illities and indwelling
emerging issues in innovation, technology and policy research

A collection of selected Research Essays
edited by Manuel Heitor and Paulo Ferrão

Center for Innovation, Technology and Policy Research, IN+
Instituto Superior Técnico, Technical University of Lisbon

June 2013
Preface

This compilation of research essays considers a selection of recent publications by researchers of the Center for Innovation, Technology and Policy Research, IN+, at Instituto Superior Tecnico, Technical University of Lisbon, which have allowed to deepen our understanding of a range of issues dealing with illsities associated with the sustainable development of our societies.

First and foremost, we thank all those at IN+ that have made possible this compilation of research essays and acknowledge all the significant progresses attained over last years. In particular, we acknowledge all those that have made significant contributions over the last 15 years, since the official installation of IN+ in 1998.

Overall, IN+ has been developed as a space for interdisciplinary research about contemporary issues in innovation, technology and policy, with main applications in engineering, science and society. The research agenda has been oriented to those who are working in engineering systems, as well as to those who have serious interest in exploring this area together with social scientists and scholars in the humanities. In addition, it aims to foster the scientific culture of all those involved in specific research areas, as well as of university students and other researchers interested in discussing challenges and new opportunities for engineering systems and related policy implications.

It is under this context that this compilation of research essays also inspires the main strategic plan for IN+ for the coming years, linking basic and applied research to technology development, and focused on the issues of sustainability and policy research. We aim at: i) consolidating successful research lines and increasing the level of interaction among them, ii) developing new lines of research that show considerable promise, and iii) fostering interaction with stakeholders and establishing common joint ventures on issues with scientific, societal and economic relevance.

A special thank also for all those individuals and institutions that have collaborated and supported research activities at IN+ over the last 15 years. The institutional support of Instituto Superior Tecnico is particularly acknowledged, as well as the collaboration with colleagues in the Associate Laboratory LARSYS. The financial support for the Portuguese Science and Technology Foundation, FCT, is recognized and acknowledged.

We hope this book helps promoting the discussion of new ideas and contributes to deepen our understanding about the social construction of technological systems.

Lisbon, June 2013

Manuel Heitor and Paulo Ferrão
List of contents

Introduction, Manuel Heitor and Paulo Ferrão

Part 1: Understanding Sustainability
Chapter 1: Manuel Heitor, “Open the box: Fostering global research networks in times of uncertainty”
Chapter 4: Abel Rodrigues, Gabriel Pita, João Mateus, Cathy Kurz-Besson, Miguel Casquilho, Sofia Cerasoli, Alberto Gomes and João Pereira, “Continuous carbon fluxes measurements in eucalypts: drought and felling”
Chapter 5: Alexandra Marques, João Rodrigues, Manfred Lenzen and Tiago Domingos, “Income-based environmental responsibility”.

Part 2: Contributing to foster Efficiency and Flexibility
Chapter 6: Jisun Jung and Hugo Horta, “Higher Education Research in Asia: A Publication and Co-Publication Analysis”.
Chapter 8: Teodoro Trindade and Edgar Fernandes, “Improving burner design through the spectrum analysis of flames chemiluminiscence”.
Chapter 9: Emanuel Teodori, Ana Moita and Antonio Luís Moreira, “Evaluation of pool boiling heat transfer over micro-structured surfaces by combining high-speed visualization and PIV measurements”.

Part 3: Learning to Build Capacity
Chapter 10: Hugo Horta, “Global and prominent higher education: internationalization, competitiveness and the role of the state”.
Chapter 11: Manuel Heitor, Hugo Horta and Joana Mendonça, “Developing human capital and research capacity: science policies promoting brain gain”.
Chapter 12: Miguel Amaral, Rui Baptista and Francisco Lima, “Serial entrepreneurship: impact of human capital on time to re-entry”.
Chapter 13: Miguel Preto and Rui Baptista, “Knowledge based start-ups, regional agglomeration and economic performance”.

Part 4: Looking at New Horizons
Chapter 14: Muriel Pádua, João Sousa, Hugo Horta and Manuel Heitor, “Managing uncertainty: on risk prevention of non-communicable diseases in the context of primary intervention among the elderly in vulnerable communities”.
Chapter 15: Carlos Silva, Gonçalo A. D. Pereira and Susana Vieira, “Diversity in energy consumer taxonomies: understanding user capabilities for sustainable cities”.
Chapter 16: Paulo Ferrão, Samuel Niza and André Pina, “Urban metabolism and sustainable cities”.
Chapter 18: Joana Mendonça and Manuel Heitor, “Industrialization, geography and policy: the changing pattern of industrial production”
Chapter 19: Manuel Heitor, “How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?”
Introduction
Manuel Heitor and Paulo Ferrão

“Illities” are non-functional technical and engineering requirements, including but not limited to sustainability, efficiency, flexibility or capability, which are increasingly associated with modern technical solutions and depend on the way people, institutions and the social environment interact with science and technology. Their understanding requires dealing with holistic perspectives on the increasing complexity of our daily life and related technical, cultural, social and economic relations. And, above all, it requires learning how to manage uncertainty.

Learning, per se, requires the personal and social appropriation of scientific and technological knowledge, including processes of inquiry, skill acquisition, and connoisseurship. In other words, “indwelling”, as first described in the 1950s by Michael Polanyi in a context where any knowledge is necessarily personal.

But how far can we foster learning and functioning effectively in a world where “reliable knowledge” is increasingly impersonal? This question has driven the rationale beyond this book. It considers a selection of recent publications by researchers of the Center for Innovation, Technology and Policy Research, IN+, which have allowed deepening our understanding of a range of issues dealing with illities associated with the sustainable development of our societies.

1. On the organization of this book
The various contributions included in this compilation of research essays are presented in four main parts, grouped as follows. The first part deals with “understanding sustainability” and the emphasis is on uncertainty and complexity, with specific works dealing with urban metabolism and related symbioses, waste management and carbon characterization, as well as modeling economic and social impacts of environmental concerns. Contributions consider:
   • Fostering global research networks in times of uncertainty;
   • methods for urban material flow accounting;
   • waste characterization and management to promote the recovery of metals from end-of-life products;
   • continuous carbon fluxes measurements in forests;
   • income-based environmental responsibility.

The second part deals with “efficiency and flexibility”, including the characterization of the way scientific communities deal with efficiency, but also the analysis of thermal systems and engine related flows. Contributions consider:
   • Scientometrics and the development of higher education in Asia;
   • Technological adjacency for emergent therapies;
   • burner design through the spectrum analysis of flames chemiluminiscence;
   • pool boiling heat transfer over micro-structured surfaces.

The third part deals with “capacity and diversity”, including the analysis of capacity building in science and technology and its potential economic relationships. Contributions consider:
   • internationalization, competitiveness and the role of the state in higher education;
   • developing human capital and research capacity through science policies;
   • the impact of human capital on entrepreneurship;
   • knowledge based start-ups, regional agglomeration and economic performance.

The book concludes with a series of research papers dealing with challenges for the future in terms of our understanding about the social construction of technological systems. Contributions consider:
   • Managing uncertainty: Behaviors, attitudes and difficult choices in the case of non-communicable diseases;
   • Diversity in energy consumer taxonomies;
   • Urban metabolism and sustainable cities;
• Dynamic energy budgeting;
• The changing pattern of industrial production;
• University global partnerships in a new era of international affairs and political and economic relations.

Overall, we may summarize that the results considered in this compilation of research essays allows to deepen our understanding about the sustainable development of our societies in the following main issues:

• **Industrial Ecology and Sustainability**: The study of the metabolism of urban areas has been developed and applied to support decision making in energy and waste management networks. Models for the energy system were developed, namely models combining renewable energies and demand side management and using weather forecasting tools for wind power forecasting. Physical and chemical beneficiation for recycling metals from end-of-life products has been carried out. Modelling and measurement of ecological metabolism in several systems was carried out. Carbon sequestration in sown biodiverse pastures was established.

• **Thermofluids, Combustion and Energy Systems**: The optimization of thermal-engines has been accomplished in terms of new research on improved efficiency and the control of pollutants, including fuel/liquid atomization, enhanced heat transfer processes, lean-combustion, burning biomass and gun-powder under unsteady combustion.

• **Technology, Policy and Management**: Technology adjacency through small and micro firms has been shown to determining the way science creates new markets in biotechnology. Entrepreneurial skills impact not only job creation, but also the quality of jobs created, with the more skilled entrepreneurs running larger firms and employing more skilled workers. The size of academic research units associated to individual skills determines the relative specialization of scientific production. Students impact faculty research productivity, while mobility patterns at doctorate and early career stage impact knowledge flows and productivity.

2. Building on disciplinary knowledge to learn understanding technological change and related societal issues

The advancement of knowledge in the themes listed above has been achieved through research work in various disciplinary fields, based on which the compilation of research essays included in this book was established. The paragraphs below describe main results of a larger sample of activities performed by researchers at the Center for Innovation, Technology and Policy Research, IN+, in recent years, from which the research essays included in the following chapters were selected.

2.1. Industrial Ecology and Sustainability

Five main research topics have been considered under the broad area of “Industrial Ecology and Sustainability”, including: i) energy planning; ii) Industrial ecology, Urban Metabolism and Sustainable Cities; iii) Ecological Economics; iv) Ecological Metabolism; and v) Waste characterization and management.

**Energy Planning**

High fuel costs, increasing energy security and concerns with reducing emissions have pushed governments to invest in the use of renewable energies for electricity generation. However, the intermittence of most renewable resources when renewable energy provides a significant share of the energy mix can create problems to electricity grids, which can be minimized by energy storage systems that are usually not available or expensive. An alternative solution consists on the use of “demand side management strategies”, which can have the double effect of reducing electricity consumption and allowing greater efficiency and flexibility in the grid management, namely by enabling a better match between supply and demand (Pina, Silva and Ferrão, 2012).

However, making use of renewable energies and “demand side management strategies” requires advanced energy systems modelling capacities, which were developed at IN+ in several studies (Abreu, Pereira and Ferrão, 2012; Suomalainen, Silva, Ferrão and Connors, 2012). They ranged from synthetic wind speed models, to global energy system models with emphasis on upgrading the “TIMES model” to a high
temporal resolution, compatible with hourly renewable energy fluctuations (Souza, Pina, Leal and Silva, 2012). We developed the capacity to model renewable energy including the operation of wind and hydro plants together with energy use scenarios, deployment of demand response technologies in the domestic sector and behavioral changes to eliminate standby power. The results obtained show that “demand side management strategies” can lead to a significant delay in the investment on new generation capacity from renewable resources and improve the operation of the existing installed capacity (Pina, Silva and Ferrão, 2012).

**Industrial ecology, urban metabolism and sustainable cities**

Major urban areas in the world are facing huge changes in land use and on their interaction with the environment, mainly due to increased levels of economic development, resulting in most cases in a huge urban sprawl and changes in their form. This clearly establishes an intertwining between economy, environment and quality of life at an urban level, whose understanding requires a new set of tools that may correlate the use of natural resources, economic activities and consumption patterns (Niza, Rosado, and Ferrão, 2009).

The *urban metabolism* concept is grounded on the analogy with the metabolism of living organisms’, as cities can transform raw materials into infrastructures, human biomass and waste. It quantifies the amount of materials that are consumed by each economic activity in urban areas. We have developed a set of new methods for quantifying urban metabolism making use of national statistical data publicly available and scaling it down to an urban level (Marteleira, Pinto and Niza, 2012).

Considerable advances were achieved aiming to develop straightforward methodologies to model the urban metabolism of world urban regions. The Lisbon Metropolitan Area (LMA) was the main case study for the validation of the methodology supported by EU and national statistical data. Aditional studies for urban sustainability include the uncovering of opportunities for industrial simbiosys in the LMA and rainwater reuse in buildindus (Patricio, Costa and Niza, 2013).

**Ecological economics**

Research in Ecological Economics had results in three main lines, as follows. First, The theoretical work on fair indicators of carbon responsibility developed in Rodrigues et al. (2006, 2008) was empirically applied. Patterns of international flows of carbon responsibility were identified, through the development and use of a “Multi-Regional Input-Output model” (Marques et al., 2012, Marques et al., 2013) and the concept of income responsibility was clarified (Marques et al., 2012). Second, life cycle assessment of bioenergy solutions were carried out, in a consequential perspective, including the effects of direct and indirect land used change (Gonçalves et al., 2013). Third, environmental impacts of the internet were estimated (Coroama et al., 2013; Müller et al., 2013).

**Ecological metabolism**

Research work on ecological metabolism was continued, both from the theoretical and modelling point of view, based on DEB theory (Saraiva et al., 2011a,b), and from the empirical point of view (Yuan et al., 2011). Carbon cycle measurements and modelling were carried out for two pools (soil, trees) in two systems: eucalyptus forest (Pita et al., 2011; Rodrigues et al., 2011); sown biodiverse permanent pastures rich in legumes (Teixeira et al., 2011). The latter work led to the establishment of “Terraprima – Environmental Services”, a spin-off at IST, which now manages carbon sequestration contracts with around 1000 farmers and c. 100 000 ha (>1% of the country).
Waste characterization and management: physical and chemical processing

Substantial research efforts have been dedicated to recycling and valorisation of residues, which required spread knowledge in several domains, including materials characterisation, physical-chemical and environmental analysis, physical separation (mineralurgical and other related techniques), chemical and metallurgical engineering, modelling, process development and design. Besides the characterization of secondary resources, natural raw materials have also been studied taking into account related interfaces with the environment (Madrid, Nogueira and Margarido, 2012).

Mais areas of specialization include recycling processes of metallic residues in the scope of mercury removal from waste sources (Margarido and Nogueira, 2011), recycling of sealed Ni-Cd and Zn-Mn type batteries (Nogueira and Margarido, 2012), valorisation of residues from military activities, and waste of electric and electronic equipment, among other waste streams (Nogueira and Margarido, 2012).

2.2. Thermofluids, Combustion and Energy Systems

Two main research topics have been considered under the broad area of “Thermofluids, Combustion and Energy Systems”, including: i) Two-phase flows and heat transfer, including surface microstructuring and flows for smart structures; and ii) Flame diagnostics, propagation and control, including flame acoustics.

Two-phase flows and heat transfer: surface microstructuring: fluids and flows for smart structures

The research program addresses the study of transport phenomena (continuum mechanics) focusing on momentum, energy, and mass transfer at fluid flow interfaces. It follows an interdisciplinary approach focusing on physical, chemical and biological properties of surfaces. New results focused on flow phenomena occurring in microfluidic devices aiming to provide insight into the background physics towards the development of new theoretical and empirical models for micro-scale flow phenomena, as well as the design of new practical systems. It is a research program combining branches of physics, mechanics, optics and fluid dynamics with material science and chemistry.

Experiments have been designed taking into consideration the specific boundary conditions of specific applications, including HCCI engines (Panão, Moreira and Durão, 2012). Experimental techniques were developed to allow characterizing the processes occurring in the region immediately adjacent to the interface between phases (Moita, Teodori and Moreira, 2012). The experiments performed included the measurement of multiphase flow quantities, such as liquid and vapor concentrations, and chemical species concentration and temperature maps with enough spatial and temporal resolution (Moita, Moreira and Roisman, 2012), using Laser Induced Fluorescence (LIF), high-speed thermography, Particle Image Velocimetry (PIV) and Phase-Doppler Interferometry (Teodori, Moita and Moreira, 2012).

Flame diagnostics, propagation and control: Flame acoustics

The research work was centered on the analysis of the competing effects between heat- mass- and momentum transfer on the dynamics of laboratory flames for different type of fuels (solid, gaseous mixtures). It considered the development of flame diagnostic sensors. The research work has addressed the following topics:

- Study the ignition process of blended fuels on lean mixtures: In the context of advanced control of NOx emissions, the ignition of lean flames poses a challenge because it requires a large amount of energy, which is also largely dependent on flame sensitivity to local stretch ratios (depending on Le number) and fuel mixtures (biofuel type). All together, they inhibit initial flame kernel development conditioning flame ignition. This work under progress aims to study the spark ignition process under reactive lean mixtures.

- ThermoAcoustic Instabilities – Flame Transfer Function of impinging flames: The challenge here was to evaluate the influence of flame deformation on flame response to upcoming oscillations (due to stretch effect). Various experimental setups were used to evaluate the burning of biomass and
burning gun-powder with emphasis on situations characterized by unsteady combustion (Mericia, Trindade and Fernandes, 2012; Fernandes and Leitão, 2012).

- **Chemiluminescence**: The main objective of this research task has been to evaluate the extent to which the chemiluminiscence diagnostic concept can be used to monitor the oxidation of fuel blends (C4H4/C3H8/H2). Although established correlations between chemiluminiscence signals and heat release and equivalence ratio exists for pure fuels, it is not certain whether a straightforward synthesis of these individual signals remain valid for practical fuel blends. The work has involved non-separable thermal and chemical effects and possible kinetic-interaction effects.

- **Development of High speed Uv-Vis spectrophotometer**: Design, assembling and testing of a High speed Uv-Vis spectrophotometer to be used in the analysis of unsteady flames and fast luminiscent processes such as flame ignition (Alekseenko, Fernandes and Shtork, 2011).

- **Development of mini sound probes - Mathematical and Experimental analysis**: The device studied is a pressure probe which configuration is modeled mathematically based on the low reduced frequency model, where the viscous acoustic shear layer is taken into account in the analysis of sound propagation inside capillary tubes, before being calibrated experimentally.

2.3. Technology, Policy and Management

Three main research topics have been considered under the broad area of “Technology, Policy and Management”, including: i) Science, higher education and policy; ii) Industrialization and innovation dynamics, including the analysis of processes of technical change and entrepreneurship; and iii) Risk Governance, science systems and the social appropriation of knowledge.

- **Science, higher education and policy**

  Research on “science, higher education and policy” has focused on broad and overlapping issues between science and technology systems and higher education systems (Heitor and Horta, 2012; Heitor, 2012). It has contributed to the knowledge pool in both literatures, providing informed evidence with a key practical relevance for emerging and developing regions worldwide (Heitor, Horta and Mendonça, 2012; Jung and Horta, 2012). We follow a systemic and thematic approach to our studies that are synergetic and feed into each other area (Heitor and Horta, 2012).

  The systemic approach tends to analyse from a historical perspective the evolution of science and higher education systems and the roles of policies, resources, investment and reforms on the development of those systems. Current challenges facing those systems in developed and developing countries have been characterized and policy implications have been discussed.

  Thematic studies performed over the last years tend to focus on relatively understudied issues such as academic inbreeding (Horta, 2013), the teaching-research nexus (Horta, Dautel and Veloso, 2012), or knowledge sources, with a mix of methodologies (Horta and Lacy, 2011). The dialectic between these approaches leads to a rich analytical set that has a level of complexity meeting the growing complexity of knowledge systems, as well as the uncertain and complex conditions that societies in general are experiencing under globalized dynamics (Heitor and Horta, 2012).

- **Industrialization and innovation dynamics; technical change and entrepreneurship**

  The rational for this research theme on industrialization and innovation dynamics is driven by the observation that industrialization has been the main driver behind rapid productivity growth achievement and social well-being improvements in different countries in the last 200 years. However, the weight of manufacturing in the economy has been decreasing substantially in many countries and regions, and production has been concentrating in certain regions, while others have been increasingly loosing their productive ability, leading to changes in employment, and raising new concerns.
The research agenda has addressed issues associated with industrialization dynamics (and related desindustrialization risk), considering development patterns through technical change, integrating emerging science and technology capacity, the role of entrepreneurial activity and the creation of new firms and industries (Mendonca and Faria, 2012). In particular, the research work has focused on economic impacts of entrepreneurship in terms of employment generation and innovation (Baptista and Preto, 2011), as well as on the value of entrepreneurial human capital, exploring its linkages with firm performance and quality of job creation (Baptista, Lima and Preto, 2012). The role of the new technology based firms is acknowledged and its relationship with FDI, internationalization and knowledge creation has also been also explored, with emphasis on Portugal and on the basis of specific employment data sets.

A particular attention was considered for the study of biotechnology and related entrepreneurial activities, with emphasis on the analysis of emerging regulatory frameworks (Couto, Perez-Breva and Heitor, 2012). A new concept of technological adjacency for emergent therapies as been discussed, in terms of knowledge creating market strategies for small biotech start-ups (Couto, Perez-Breva and Heitor, 2012).

Risk Governance, science systems and the social appropriation of knowledge

The rational for research has been driven by the need to help facilitating the social appropriation of knowledge towards designing resilient cities. The focus has been on launching a team for innovative “hands on” approach based on the design of new engineering-based products and processes to help shape perceptions and peoples’ behaviour.

Two areas of intervention were chosen in terms of risk mitigation, including energy consumption and non-communicable diseases (e.g., diabetes). The main goal is to assess risk perceptions of lay people belonging to vulnerable communities and, in the process, to examine strategies of risk communication towards these groups. While early studies on risk perception were based on “unilateral” expert views (i.e., “methods of expert elicitation”), it has become more and more clear that the involvement of lay people in the process is critical for risk governance (i.e., to ensure their participation). But, and despite these new trends, vulnerable groups remain an outlier category of this type of analysis (Pâdua and Custodio, 2012).

The research team used the knowledge representation approach developed by Morgan (2002)\(^1\) to identify misconceptions of groups from different cultural backgrounds towards diabetes. In summary, a major initiative on learning for uncertainty in urban contexts was developed, including actions to look at risk perceptions, risk communication and stakeholder engagement of lay people from vulnerable communities. The Mouraria neighborhood, in Lisbon, has been used for preliminary fieldwork, which was focused on two distinct areas for risk mitigation, namely: i) non-communicable diseases, such as diabetes (Pâdua, Santos, Horta and Heitor, 2013); and ii) patterns of consumer behaviour in energy usage.

Our research hypothesis is associated with the idea of “indwelling“, firstly introduced by Polanyi\(^2\) and recently explored by John Seely Brown\(^3\) in terms of understanding learning through processes of knowing, playing and making. We are attempting to provide new evidence on related learning processes through the distinct experiments mentioned above.


\(^3\) John Seely Brown, Douglas Thomas, A New Culture of Learning: Cultivating the Imagination for a World of Constant Change, CreateSpace 2011.
First, regarding non-communicable diseases, such as diabetes, interviews with targeted communities reveal that many misunderstand the risks they face. Through our experiment, lay people has been encouraged to ride a "stand-alone" bicycle, facilitating physical exercise and making a smoothie (healthy fruit based drink) through the preparation of crushed fruit, yoghurt or milk. This begins a learning and revelatory process concerning adequate diet, exercise and successful ways of reducing their risk by having fun together. But, above all, it facilitates making people aware of diabetes risks and to introduce the debate on the topic among specific and target communities. Experiments were successfully conducted with the elderly and specific muslin groups (Pàdua, 2012).

Second, consumer behaviours of energy usage in domestic and public spaces were assessed and monitored through the application of smart energy meters and the consequent attempt to design new products to help lay people to save energy and reduce their energy bill. Three different consumer typologies were assessed in residential properties, including two families in social housing and an apartment with Erasmus students. In addition, energy consumption in a “tasca” was monitored. In both examples, participants learn about energy flows, energy conservation, cost saving and reducing avoidable energy wastage.

Our experiments were conducted by mobilizing groups of university students in a way that also provided new insights into university learning methods. In addition to look at ways of lay people addressing risks, our results provide new insights into the modernization of university education though "hands-on" experimentation in vulnerable communities and the socialization of knowledge and knowledge networks.

3. Setting a research agenda for the next decade

The compilation of research essays included in this book describe recent research results, but they also open new horizons for future research on innovation, technology and policy. Main issues to be considered in the years to come are briefly described below.

3.1 Industrial Ecology and Sustainability

Integrated models of environment-energy-economy interaction will be developed, at multiple spatial scales (cities, regions, countries), using models such as input-output, computable general equilibrium models, and economic growth. These will be used to support policy making on sustainable use of energy and raw materials in general, and regional, urban and rural sustainability planning in particular. Based on these models, work will be carried out in the development of sustainability assessment tools and indicators, e.g. Green GDP, ecological footprint and human appropriation of net primary production. Particular focus will be given to the environmental themes of solid waste, greenhouse gases and carbon sequestration.

Urban metabolism will deserve particular attention, as described below, focusing on developing spatially comprehensive and temporally broad physical models of resource consumption of urban centres. Additionally, energy and materials consumption in buildings and new and innovative solutions to promote the concept of "Sustainable Buildings" will be studied.

In the context of the sustainable energy systems, which will constitute a major research area, the design of future intelligent energy and transportation systems which are "green", "smart", and "efficient", requires an understanding of a region's current systems, including detailed characterizations of its energy networks, supplies, and demands, and of the main factors influencing the evolution of those supplies and demands, including multiple renewable resources, and socio-economic and behavioural dynamics.

A major research topic is the development of models to facilitate the penetration and integration of forms of renewable energies. The need to foster aggressive energy efficiency end-uses introduces a new set of issues in energy systems planning that the current tools are not able to answer. This includes developing models that resolve the hourly dynamics of renewable resources in order to evaluate accurately the match between demand and supply; it is necessary to include the demand dynamics to evaluate accurately the
impact of demand side management strategies like load shedding or load shifting. In parallel, operational NWP (Numerical Weather Prediction) models for Portugal will continue to be used in forecasting availability from wind power plants. Progressively, these two lines of work will be integrated.

Environmental modelling will address multiple ecosystem services and their integration in land planning issues, at the city, municipal, regional and landscape scales, incorporating detailed climatic information from the ongoing climatic reconstruction of a normal climatological series (1979-2009) at high spatial resolution (better than 9 km) based on the most recent global reanalysis originated in the USA (~40km resolution).

Urban Metabolism and Sustainable Cities (UMSC): building a “Research Program”

The program aims at strengthening knowledge about cities metabolism drivers, to support sustainable management of urban-driven flows and defining policies for urban sustainable management. For this purpose, it is intended to create a multidisciplinary model of the urban area, including impacts of energy and materials use, energy generation within the city, urban spaces outdoors’ quality and issues of materials sustainability, namely life cycle assessment and pollution.

Additionally it will contribute on providing appropriate tools and methods for managing over space and time scales, integrate biophysical and social components, adapt and deal effectively with uncertainty, visualize and identify planning problems and work to identify trade-offs and synergies in undertaking interventions.

Three main themes will be considered: i) Urban Flows assessment; ii) Critical factors for urban sustainability; and iii) Supporting urban planning and policies for sustainability. The following paragraphs briefly describe the rational for the research planned.

Urban Flows assessment. Properly managing (optimizing and reducing) urban flows demands knowing the quantities involved. Despite broad acknowledgement of the concentration of global energy and material resource consumption in urban centres, very little is known of the actual resource consumption of contemporary cities, mainly because significant challenges remain in accounting for the resource burden imposed by cities.

Therefore, a critical task involves identifying and mapping the main flows of energy, water, materials and products in cities. Research will be performed in order to create a database format for translating, storing and crossing information from existing databases compiled by different urban agents (companies, utilities, municipal services and others).

Potential productive linkages between physical accounting tools usually used in the assessment of particular products, such as Life Cycle Assessment (LCA), assessment of economic relations between sectors (Input-Output Tables) and those tools used for spatially defined resource consumption, as Material Flow Accounting (MFA) should be evaluated. Tools integration should provide the basis for an analytical tool specifically oriented toward the assessment of urban metabolism.

Critical factors for urban sustainability. The urban subsystem is dynamic, so the status of the demand actors and the associated resource flows are constantly changing. This might be linked to the dynamics of land use, population growth or technological innovations. Some of these changes are intrinsic to the urban system under observation and other may be brought through (external) regional, national and international influences.

The research activities to be performed in the years to come will also distinguish between continuous (e.g. internal socio-economic) and discrete (deliberate intervention) influences and consider their rate of change (land use changes slowly whereas changes in tenancy or in employment status may happen relatively quickly). These changes inevitably influence urban sustainability. Therefore, critical factors should be identified and studied in order to provide guidelines towards sustainable development. Case study designs and demonstration projects will be used to understand the correlation between resource consumption and several socioeconomic, geographical, climate and governance parameters. This will lead to the articulation of criteria for the planning of communities that better serve a resource efficient urban fabric.
Supporting urban planning and policies for sustainability. While the growth of many older cities consists of a densification of urban cores and metropolitan zones, certain younger urban centres continue to expand their perimeters by extending transportation, water, waste, and power infrastructure while annexing adjacent rural and agricultural land. The nature of various types of urban growth and change and the consequences for municipal consumption requires studying cities and propose policies towards a more sustainable planning of cities.

Together with the development and refinement of accounting methods, it is important to develop a spatially comprehensive and temporally broad physical accounting of resource consumption of a variety of urban centres. Material flows dynamic stock modelling should be used to analyse the industrial metabolism dynamics, namely the complex interaction between resource use and waste production and recovery, in order to contribute for the definition of more integrated and sustainable resource management policies.

The focus of this research area will be the coupling of emerging methods for assessing and tracking the resource consumption of cities with strategies for implementing design and planning recommendations for communities. Research will examine the potential for applying to the design and planning of more resource efficient cities a variety of strategies that have emerged from the field of Industrial Ecology. Through current work in tracking the material and energy flows devoted to urban centres, this study will identify networks for symbiotic resource exchanges and productive reconfigurations of primary elements of urban form and infrastructure for sustainable city planning.

3.2 Thermofluids, Combustion and Energy Systems

The research rational in this area focuses on understanding the fundamental relationship between concurrent thermodynamic transports phenomena involved in energy conversion processes. Within this framework, research activities will be oriented towards combining efforts in developing mathematical/physical models to downsize of energy conversion systems. Flow phenomena occurring in microfluidic devices will be analysed, with and without chemical reaction, towards the development of new theoretical and empirical models for micro-scale flow phenomena, as well as on the design on new practical systems. It is a multidisciplinary research field, combining branches of physics, mechanics, optics and fluid dynamics with chemistry.

The general objective is to develop and use different tools and new approaches to support the development of energy sustainable products towards consumers, e.g. ranging from heating/cooling suppliers to medical devices.

The main topics include:

- **Study the dynamics of interfacial transport phenomena in multiphase systems with phase change.** The goal is to deal with the hierarchical complexities of pool boiling, of a velocity boundary layer flow at a flat interface and of liquid flow in micro-channels with phase change. The study must address: i) The interfacial effects and transport phenomena associated with surface tension gradients caused by temperature and concentration variations on bubble interfacial shape and dynamical behaviour including nucleation, bubble motion and coalescence; ii) The effects of surface topography and chemical properties (wettability) on interfacial transport phenomena during liquid nucleation boiling; and iii) To develop/built appropriate experimental techniques and/or methods to quantify the physics of dynamic interfacial transport phenomena or processes including jet-like flows, bubble interaction and spatial scale effects.

- **The Dynamics of Flame anchored in micro burners.** The research work is to deal with the competing effects between heat- mass- and momentum transfer on the dynamics of small flames anchored on mini burners burning blended fuels. The work will address the following topics: i) Study the ignition process of blended fuels on lean mixtures, to study the spark ignition process under reactive lean mixtures; ii) Thermo-Acoustic Instabilities – Flame Transfer Function of micro flames, to tackle the determination of the experimental acoustic flame transfer function of small conical flames followed by the flame response modelling through the G-equation; and iii) Thermo-Acoustic Instabilities – Flame Transfer Function of micro flames, to evaluate the extent to which the chemiluminescence diagnostic
concept can be used to monitor the oxidation of fuel blends (CH4/C3H8/H2). Although established
correlations between chemiluminescence signals and heat release and equivalence ratio exists for pure
fuels, it is not certain whether a straightforward synthesis of these individual signals, without
considering various non-separable thermal and chemical effects and possible kinetic-interaction
effects.

3.3. Technology, Policy and Management
The research agenda for this area will emphasize four main themes: i) knowledge for development; ii)
Systems of knowledge creation and diffusion; iii) Technological change, industrial development and
innovation dynamics; and iv) entrepreneurial behaviour and new business development. In addition, a
specific Research Program will be implemented to look at new challenges for “Science, Higher Education
and Policy” (SHEP).

Research on “knowledge for development” is aimed to foster the systematic observation and in-depth
research of issues in science and technology, higher education and public policy in developed and
developing regions worldwide. The ultimate goal is to create and promote a totally independent and
credible international observatory of science, technology and higher education policies in a way to report,
publicly and periodically, relevant information and early warnings on the state of policies and budgets at a
country and regional level. It should foster an international perspective and convey new research and
understanding of the impact of the current economic situation (mainly in Europe) on the “states of
knowledge”, including science, technology and higher education capacity.

Research on “Systems of knowledge creation and diffusion” is centred on the relation between education,
diffusion of scientific knowledge, and the risk awareness behaviours and attitudes of populations at large.
The ultimate goal is to inform policy regarding the features that education and scientific awareness
influences, and in what ways educational levels draw people to act when they are at risk. This has key
relevance for science, education and health policies in ageing, modern societies.

Research on “Technological change, industrial development and innovation dynamics” focuses on
industrialization and deindustrialization processes in both developing and developed regions of the world.
It envisions the setting-up of a large task force for the “observation” of industrialization, to cover various
aspects, including: i) The geography and dynamics of economic development and specialization – how
scientific, technological and industrial bases evolve and impact socioeconomic development; ii) The
structure, geography and dynamics of supply chains and knowledge networks in different sectors and
markets; iii) The structure and availability of human resources and competences, as the basis for industrial
activity; iv) Policy tools to foster local industrialization processes (e.g., public procurement, local production
agreements, public expenditure in R&D and training) and; v) Deindustrialization processes, characterizing
them and identifying, analyzing and governing related risks.

Research on “Entrepreneurial behaviour and new business development” will focus on entrepreneurial
human capital and skills; occupational choice and transitions into – and out of – new founded firms;
habitual entrepreneurship; founding teams; entrepreneurial activity among minority groups in society (e.g.
immigrants, disabled and older individuals); social groups that are underrepresented in entrepreneurship
(e.g. women); firm demography and industry dynamics; knowledge intensive and technology-based
businesses; high-growth firms; job creation and regional development.

Science, Higher Education and Policy (SHEP): building a Research Program
Particular emphasis will be devoted to foster the systematic observation and in-depth research of issues in
science and technology, higher education and public policy in Europe. The ultimate goal is to create and
promote a totally independent and credible international observatory of science, technology and higher
education policies and budgets across Europe in a way to report, publicly and periodically, relevant
information and early warnings on the state of policies and budgets in each country and at EU level. It should foster an international perspective and convey new research and understanding of the impact of the current economic situation in Europe on the “states of knowledge”, including science, technology and higher education capacity. It would, therefore, help to increase the public awareness of the strategic importance of science and higher education policy decisions, as well as to strengthen the motivation of scientists and the academy to engage themselves in policy action as informed and responsible citizens.

This is important because the conditions for the social construction of technological systems in both central and peripheral EU regions and societies will be addressed in terms of their impact on the emergence of new social realities in those societies, as well as their potential as factors of economic and social development on a global scale. To achieve these goals, science and technological development case studies will be developed across EU member states.

In addition to the tasks mentioned above about Europe, a research consortium is under establishment to foster the process of building human capital and scientific competencies in emerging and developing regions and countries, which do require adaptable and resilient scientific and academic institutions. It considers new research activities and fieldwork, together with the advanced training of a new generation of academic, scientific and policy leaders for emerging regions and countries, with emphasis on China, Brazil and Russia.

China, Russia, and Brazil are leading emerging regions at a world level facing the need and the opportunity for large investments in science, technology and higher education. These aim at responding to the explosive social demand for higher education and to the vast social and political transformations already induced by new waves of educated youth. These investments not only seek new skills and but also the certification of quality that may be expected from working together with well-established academic and scientific institutions from developed countries.

Also, processes of brain-circulation, associated to concurrent patterns of brain-gain, brain-return and brain-drain, urge for a need to better understand flows of people and knowledge, and how the spillovers resulting from these flows are changing the way we learn, view the world, and countries and institutions act. This relates to the need for a new approach on the design of higher education (HE) at a world level and in very different socio-economic and cultural contexts.

Overall, The Research Program on Science and Higher Education Policy (SHEP) will consider different world regions and always with an international comparative approach. This will have an integrative and multidisciplinary approach as follows:

- **The career trajectories of the highly qualified**: The analysis of the career trajectories of highly qualified people is important to understand how policies, institutional and career incentives, and other factors impact on individual career choices, networking, and productivity. In this framework, features such as understanding the impact of mobility, internationalization, interdisciplinarity, educational backgrounds and family issues along the career, are important to illuminate the knowledge on how these individuals decision-making and contribution to institutional and regional knowledge building occurs. This will be performed for those developing activities in the higher education sector but also in the business enterprise sector.

- **The role of international science base partnerships**: It is important to understand how international science base partnerships (and other science policy policies as well) contribute to reform higher education systems. This will be pursued equally regarding institutional change at universities in terms of structural change, and research and teaching activities and practices. Through an international comparative analysis of several educational hubs, it aims to understand the impact that international science partnerships with lesser or broader scopes of action can have in promoting a more active role of
universities in society, including in supporting of the knowledge base of communities and engagement in industry-university relations.

- **Employment policies for the highly qualified**: The aim is to understand on a historical perspective supply and demand policies concerning highly qualified people in emerging regions. This is to be contextualized and combined with the scientific and economic structure of the countries/regions under study, and focused on specific sectors of economic activity. The objective is to assess to what extent public policies focused on demand can be pursued in contexts that do not have developed the capacity to absorb highly qualified human resources. This has a clear linkage to public policies focused on the supply side and on the thematic of brain-gain, brain-drain and brain-circulation. Therefore, a complementary analysis is one of the fluxes of highly qualified people and Diasporas.

- **Geographies of knowledge and scientific structures**: As knowledge increasingly flows globally, it is essential to understand its geographical poles, its concentration, integration, and engagement drivers. A better understanding of the evolution of collaborations among regions, countries, and institutions, including universities and firms is of particular importance. Through analyzing evolutionary mappings of knowledge production and collaboration one can realize the patterns of change at global level, and how knowledge global flows of are interacting and being constructed by national and regional knowledge bases. In this regard, the analysis needs to take into perspective the engagement of different scientific structures; by looking at how broader or narrower they are set in terms of disciplinary fields.

References (grouped per main research topic)

**Energy Planning**


Industrial ecology, urban metabolism and sustainable cities


Marteleira, R., Pinto, G., Niza, S., (2012), Regional Water Flows – Assessing opportunities for sustainable management (Submitted)


Patricio, J., Costa, I., Niza, S. (2013), Urban material cycle closing – Assessment of industrial waste management in Lisbon Region (Submitted)

Ecological economics


Ecological metabolism


Waste characterization and management: physical and chemical processing


Two-phase flows and heat transfer: surface microstructuring; fluids and flows for smart structures


Flame diagnostics, propagation and control: Flame acoustics


Science, higher education and policy


M. Heitor (2012), ”How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?”, Technological Forecasting and Social Change, accepted for publication. A version of this paper was presented at the ASHE Annual meeting 2012, Las Vegas, November 2012.


Industrialization and innovation dynamics; technical change and entrepreneurship


Risk Governance, science systems and the social appropriation of knowledge


Chapter 1:

Manuel Heitor, “Open the box: Fostering global research networks in times of uncertainty”
Open the box: Fostering global research networks in times of uncertainty

Manuel Heitor

Abstract
The basic premise of this paper is that the central locus of innovation has increasingly become distributed and increasingly dependent upon linkages between many different institutions and sources of knowledge. First, the increasingly transnational business, technology and science require evolving from nationalistic approaches to new collaborative policy frameworks. Among these, large international collaborative arrangements play an emerging role. Second, the science and technology performance sectors, namely government, industry and academia, remain key players, but the connectivity, links and associations with other institutional players and agencies is no less important. In particular, the increasingly relevant role played by new technology-based firms is identified, which are also becoming global. This requires opening-up science policies to multiple public and private agents and promoting global research networks towards socio-economic resilience and the sustainable development of our societies.

Introduction
Why research and innovation need to go global?...and why governments need to fund global research networks beyond national borders?
These questions are gaining increasing relevance as much of the political debate in the US and EU is centered on economic competitiveness in the long term, most of the times under a rather “nationalistic” approach to innovation for growth. The question that does arises is if the acceleration of knowledge investments in China, India, Brazil and Russia (i.e., the so-called BRICS) signal the decline of the US and EU and should it be countered by aggressive “techno-nationalism” in the US and EU in any form?
This is important to assess because of the current trends in the US economy (and of other major industrialized countries) in times characterized by the increasing importance of knowledge for economic development, but mainly in terms of the recent global economic and financial crisis and the related structural transformation of the global economy (OECD, 20101). The rapid growth and development of emerging markets in the BRICS has led to a shift in economic and political powers to the extent that it is estimated that the aggregate economic weight of those economies is about to surpass that of the traditional industrialized powers of the past decades (Table 1).
Table 1: The Top 20 Economies in 2050 – GDP Forecasts

<table>
<thead>
<tr>
<th>Order in 2050 by GDP</th>
<th>Size of economy in 2050 (Bn, Constant 2000 US$)</th>
<th>Rank change between now and 2050</th>
<th>Income per capita (Constant 2000 US$) 2050</th>
<th>Income per capita (Constant 2000 US$) 2010</th>
<th>2050 Population (Mn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 China</td>
<td>24617</td>
<td>2</td>
<td>17372</td>
<td>2396</td>
<td>1417</td>
</tr>
<tr>
<td>2 US</td>
<td>22270</td>
<td>-1</td>
<td>55134</td>
<td>36354</td>
<td>404</td>
</tr>
<tr>
<td>3 India</td>
<td>8165</td>
<td>5</td>
<td>5060</td>
<td>790</td>
<td>1614</td>
</tr>
<tr>
<td>4 Japan</td>
<td>6429</td>
<td>-2</td>
<td>63244</td>
<td>39435</td>
<td>102</td>
</tr>
<tr>
<td>5 Germany</td>
<td>3714</td>
<td>-1</td>
<td>52683</td>
<td>25083</td>
<td>71</td>
</tr>
<tr>
<td>6 UK</td>
<td>3576</td>
<td>-1</td>
<td>49412</td>
<td>27646</td>
<td>72</td>
</tr>
<tr>
<td>7 Brazil</td>
<td>2960</td>
<td>2</td>
<td>13547</td>
<td>4711</td>
<td>219</td>
</tr>
<tr>
<td>8 Mexico</td>
<td>2810</td>
<td>5</td>
<td>21793</td>
<td>6217</td>
<td>129</td>
</tr>
<tr>
<td>9 France</td>
<td>2750</td>
<td>-3</td>
<td>40643</td>
<td>23881</td>
<td>68</td>
</tr>
<tr>
<td>10 Canada</td>
<td>2287</td>
<td>0</td>
<td>51485</td>
<td>26335</td>
<td>44</td>
</tr>
<tr>
<td>11 Italy</td>
<td>2194</td>
<td>-4</td>
<td>38445</td>
<td>18703</td>
<td>57</td>
</tr>
<tr>
<td>12 Turkey</td>
<td>2149</td>
<td>6</td>
<td>22063</td>
<td>5088</td>
<td>97</td>
</tr>
<tr>
<td>13 S. Korea</td>
<td>2056</td>
<td>-2</td>
<td>46657</td>
<td>16463</td>
<td>44</td>
</tr>
<tr>
<td>14 Spain</td>
<td>1954</td>
<td>-2</td>
<td>38111</td>
<td>15699</td>
<td>51</td>
</tr>
<tr>
<td>15 Russia</td>
<td>1878</td>
<td>2</td>
<td>16174</td>
<td>2934</td>
<td>116</td>
</tr>
<tr>
<td>16 Indonesia</td>
<td>1502</td>
<td>5</td>
<td>5215</td>
<td>1178</td>
<td>288</td>
</tr>
<tr>
<td>17 Australia</td>
<td>1480</td>
<td>-3</td>
<td>51523</td>
<td>26244</td>
<td>29</td>
</tr>
<tr>
<td>18 Argentina</td>
<td>1477</td>
<td>-2</td>
<td>29001</td>
<td>10517</td>
<td>51</td>
</tr>
<tr>
<td>19 Egypt</td>
<td>1165</td>
<td>16</td>
<td>8996</td>
<td>3002</td>
<td>130</td>
</tr>
<tr>
<td>20 Malaysia</td>
<td>1160</td>
<td>17</td>
<td>29247</td>
<td>5224</td>
<td>40</td>
</tr>
</tbody>
</table>

It is well known that the US economy, as well as most of the OECD economies with the notable exception of Germany, has been transformed by the change in emphasis from manufacturing to services industries and, for example, the number of U.S. manufacturing jobs has declined from nearly 20 million in 1979 to about 12 million today. This has occurred together with the rapid growth of manufacturing capacity in the developing world with much lower wage and other costs. To remain competitive, analysis has consistently shown that developed nations will have to rely more and more on knowledge intensive industries (e.g., Berger, 2005; Hidalgo and Hausmann, 2009), which are increasingly global and very much dependent on innovation trends in the US. We argue in this paper that these developments will have major consequences for the demand patterns relating to science, technology and innovation worldwide.

We argue that any new narrative on global research networks requires the analyses of, at least, the last decades and the seminal work of Sylvia Ostry and Dick Nelson (1995), among many others for the last twenty years, has called for our attention of the relationship between the globalism of firms and the nationalism of governments, as well as the related interplay of cooperation and competition that characterizes high technology and knowledge-based environments.
It should be noted that the Brookings Institution’s project of the early 90’s has promoted this debate, although in a different international context, and clearly shown that tensions about deeper integration arise from three broad sources: cross-border spillovers, diminished national autonomy, and challenges to political sovereignty. As a result, the technogobalism of the 80’s gave rise to national policies designed to help high-tech industries become more innovative and, consequently, the emergence of technonationalism.

It is under this context that the concept of “national systems of innovation” emerged in academia, mainly through economists and related schools of thought, to explain and explore how and why the systems have evolved differently in the major industrial nations, mainly US, Japan, UK, Germany and France. It was clear by then that the increasing international tensions and economic instability were largely a result of the attempt of governments to impose national technology and innovation policies on a world in which business and technology are increasingly transnational.

This brief paper draws main implications of that trend over the last decades for new science and technology policies in times of increasing uncertainty and on the rise of globalization. We start by discussing the “old” paradigm of “national systems of innovation”, which is rather limitative on addressing emerging patterns of openness and international cooperation. Then, we consider main issues on science and innovation policies, including the need to strengthen main pillars of research and education, an industrial base for socio-economic resilience and articulation with diversified stakeholders, including new technology based firms.

Beyond national systems of innovation

The concept of “national systems of innovation” has evolved during the last two decades, first in association with the need to fight against “market failures”, then against “system failures”. And it helped building new nationalistic policies all over the world, but just as business and science are becoming increasingly transnational. The end result has been a frustration of national policies, on one hand, and a further move toward the multi-nationalization of business, on the other.

This requires many observations and, certainly, deepening the debate in relation to the current economic and social situation in the US and EU, as compared to those in newly industrialized regions and the, so-called, BRICS. First, the myth of “national” high tech industries and related policies to protect them requires to be better understood, if analyzed in terms of the increasing unemployment rates. Second, the debate itself on “national innovation policies” is in any case naïve. No country, even in non-democratic regimes, ever seems to have had a broad, well coordinated innovation policy, mainly because of the complex structures associated with any “innovation ecosystem”.

Looking at the last two decades, the picture that is emerging at a global level is not very much different from that discussed by Sylvia Ostry and Dick Nelson in the early 90’s. In other words, one of increasing internationalization of private business strategies, while government innovation polices and science funding agencies remain overwhelmingly national. This is leading to new dilemmas for policymaking
and to new sources of international friction, although with new boundaries and new players. The key issues to answer include what are the implications of increasing technoglobalism for national and international innovation policies, namely US and EU innovation policies? And, also, what new approaches are required to reduce international frictions and where do public policies need wider integration?

For the case of the US, the key message that emerges from analyzes of long-term patterns of investments in S&T is that of a diversity of policies that led over time to increased opportunities for citizens, as well as to increased institutional specialization based on a clear separation of the role of private and public incentives to support S&T. In addition,

For Europe, recent analysis also argues that the debate on climate change, the recent financial crisis, and the new Chinese dominance of the world market, mean there is a need to revisit the role and design of industrial policy. This has been used to justify the need for renewed targeted sectoral intervention of governments, namely to redirect production and innovation towards clean technologies, as well as to make industrial policy more competition-friendly and more innovation-enhancing.

Analysis in the literature has also clearly shown that China’s capacity to innovate is still quite limited as compared especially to the capacity of the US. A similar comment could be raised about Brazil, India or Russia and, therefore, there is a large scope to better discuss US and EU innovation policies in a broad international context, well beyond national borders. In addition, a new paradigm of international academic, scientific and technological cooperation that seems to emerge, as discussed below.

**Strengthening the pillars**

At a time of increasing financial difficulties derived due to public budget constraints, there is the expectation that the links between research activity and its application in society will be reflected in more direct and immediate financial flows. However, this perception is leading to an institutional convergence between what universities do (and are supposed to do) and what firms and other agents do.

In fact, more than a decade after Burton Clark launched the idea of “entrepreneurial universities”, there is still much to learn about their impact and analysis has clearly considered this convergence a potential threat to the institutional integrity of the university and the future of scientific research due to the commoditization of knowledge.

The issue is not to “save the university”, but rather to understand who will play the fundamental and unique role that universities have played in the overall cumulative system of knowledge generation and diffusion. It is clear that the US and EU elites are not willing to allow this integrity to be jeopardized. By misunderstanding national policies towards university-based research, there is a grave danger that university policy worldwide will destroy these basic functions, which would be detrimental to the global production of knowledge, but would also certainly harm the development prospects of many US and EU regions.
Overall, changing the patterns of teaching and learning, making students’ work more active, and fostering student-centered education schemes, are the ultimate goals of many leading institutions, which should be better understood at an overall level. We need to allow students to determine their own learning paths and trajectories, particularly through education cycles, but also across institutions in different regions and countries.

The debate requires tertiary education institutions in general to better understand how people learn. It is clear that learning systems vary considerably across the full spectrum of disciplines, with arts and medicine using project-based approaches and, probably, engineering and the social sciences following a more intense academic drift. The recent book by Tony Wagner of Harvard, as well as that published last year by Douglas Thomas and John Seely Brown, represent insightful treatments about this theme.

Following the practices, skills, attitudes and values described above, education at all levels should take into account that learning a new practice requires moving through discovery, invention, and production not once, but many times, in different contexts and different combinations.

To achieve these objectives, we must learn from new research and, certainly, we also need to foster evidence-based project and experimental work, as well as to focus our attention on the transferable skills students should acquire. But we also need to reduce drop-out rates in tertiary education and to involve students in research activities from the early stages. In summary, we need to go beyond the structure of tertiary education and gradually concentrate our efforts on measuring and taking stock of the diversity and evolution of specific student-centered parameters.

But, in addition to that argument, we argue for a deep discussion about the complexity of stakeholder engagement and the politics of trust building in science and technology in the US and EU.

This is because, beyond any single measure, one may argue that it is the public understanding of science and the related level up to which people trust in academic and scientific institutions that determines the success of the politics of science and innovation policies. It is under this context that the systematic development and promotion of activities to foster science awareness, science education and the role of science in the daily life of citizens has been implemented in many regions and countries with a high level of priority in the innovation policy agenda.

Innovation and socio-economic resilience

It is clear that technoglobalism and the globalization of trade and supply chains led to the emergence of increasingly competitive global markets and to facilitated access to new suppliers, independently of their geographic location (Berger, 2005). This has allowed countries and regions with strong technological and industrial bases to profit on the lowering of trade barriers to access new markets, while the majority of firms located in other regions remained confined to local markets.
In addition, the analysis of the overall trend on moving towards knowledge intensive services and its relation with job creation and economic growth requires some pragmatism. This is because, in parallel to technoglobalism came post-industrialism, promoting services as the new developed countries’ economic growth overtook manufacturing industries. Captivated by the prospects of accelerated and cost-effective economic growth, many countries, the United States included, started shifting their focus from manufacturing industries to knowledge-intensive services (Hepburn, 201116; Ghani and Kharas, 201017).

The result is emerging with many regions worldwide lagging behind. In fact, evidence shows that, when compared to knowledge-intensive services, hard industries have higher labor productivities, a more balanced income distribution, higher income growth rates and the ability to generate exports, which are negligible in the case of services (Nairn, 200218; Fingleton, 199919).

Looking at the US, together with other most developed economies (including Germany), we can identify some common factors, but also opportunities that need to be understood in international comparative terms: strong industrial bases, diversified economy, and supply chain and knowledge networks’ complexity (Amsden, 200120; Hidalgo and Hausmann, 200921).

Approaching this question requires the setting-up of a large task force for the “observation” of industrialization, to cover various aspects, including:

- The geography and dynamics of economic development and specialization – how scientific, technological and industrial bases evolve and impact socioeconomic development.
- The structure, geography and dynamics of supply chains and knowledge networks in different sectors and markets.
- Policy tools to foster local industrialization processes (e.g., public procurement, local production agreements, public expenditure in R&D and training).
- Deindustrialization processes, characterizing them and identifying, analyzing and governing related risks.

It should be clear that a new generation of industries will drive the economic recovery over the next decade, fuelled by long-term changes in technology, society and geopolitics. The recession has not been only a point of change, and many argue that it has acted as a catalyst for growth. As the business landscape alters, we will see the emergence of new ways of doing business in an increasingly interconnected world.

**Engaging new technology based firms**

Technology-based entrepreneurship is increasingly seen as a key element of regional competitiveness and that has been taken as “the model” for many other regions and countries worldwide. Silicon Valley and Route 128 in the Boston-Cambridge area, the most dynamic regions in the world today in terms of growth and innovation, were propelled mainly by new technology and the creation of startups - Apple, HP, Google, and Intel, to name a few. At the same time, start-up companies are also
becoming global enterprises and engage in services, manufacturing, and research throughout the world, with strong links to universities and research groups. Others are going beyond their borders to procure products and services at lower prices, often from new companies or subsidiaries in countries like China, India and Brazil. Well-trained engineers and computer scientists from Bangalore and Shanghai are competing for jobs that traditionally went to their counterparts in Europe and the US.

At the same time, research universities worldwide are attempting to “emulate” their US counterparts and foster a range of technology transfer offices and commercialization activities, together with industrial liaison programs, mostly intended to foster entrepreneurial environments and the launching of technology-based start-ups. Bringing ideas to the market is their main goal.

Notably, beyond the concentration of people and skills in a number of regions, a key issue that has differentiated North America from many other countries and regions is the availability of a mix of public and private funding sources, in a quite diversified pattern and, most of them, of easy access to SMEs. It is in this context that a few countries have tried to emulate the SBIR program (“Small Business Innovation Research”), which remains unique in many of its characteristics. Although many difficulties have been found in the public support to continue SBIR (as well as that of TIP at NIST), its enormous success and impact should be further acknowledge. This is a program of the outmost importance and relevance that has helped American innovative firms to growth. In addition, many other schemes to fund and support new technology-based firms have been used in America in quite original ways, namely through public procurement through the Defense and Energy Departments.

**Forward looking**

The issue is certainly how far we all take advantage of opportunities that arise with the increasingly dynamic and globally distributed geography of innovation, as well as how it fosters a new global order and help others to use similar advantages at local levels.

This is because one must take up the challenge of probing deeper into the relationships between knowledge and the development of our societies at a global scale. Our inspiration comes from, among others, the seminal work of Lundvall and Johnson, who challenge the commonplace by introducing the simple, but powerful, idea of learning. Lundvall and Johnson speak of a “learning economy”, not of a “knowledge economy”. The fundamental difference is to do with a dynamic perspective. In their view, some knowledge does indeed become more important, but some also becomes less important. There is both knowledge creation and knowledge destruction. By forcing us to look at the process, rather than the mere accumulation of knowledge, they add a dimension that makes the discussion more complex and more uncertain, but also more interesting and intellectually fertile in an international context.

This closely follows the lessons Eric von Hippel, at MIT, has provided in recent years based on the American experience that user-centered innovation is a powerful and
general phenomenon. It is based on the fact that users of products and services - both firms and individual consumers - are increasingly able to innovate for themselves. It is clear that this is growing rapidly due to continuing advances in computing and communication technologies and is becoming both an important rival to and an important feedstock for manufacturer-centered innovation in many fields.

Eric von Hippel has also shown that the trend toward democratization of innovation applies to information products such as software and also to physical products, and is being driven by two related technical trends: first, the steadily improving design capabilities (i.e., innovation toolkits) that advances in computer hardware and software give to users; and second, the steadily improving ability of individual users to combine and coordinate their innovation-related efforts via new communication media such as the Internet.

In other words, beyond suitable technical infrastructure, the process of “democratization of innovation” at a global scale requires people with the ability to engage in knowledge-based networks without borders. It is about people and knowledge beyond national borders, and this constant interaction has gained particular importance in recent years.

It is clear that the emerging patterns of innovation require new perspectives for public policies, which in the US and in EU have in the past relied on supporting manufacturers and their intellectual property.

Certainly we need to move on from those days and consider better ways to integrate policies, as well as to diversify them at a global scale to better consider “win-all” approaches. A potential way to achieve this is to avoid overemphasizing current rivals sectors and competitive strategies, but rather to look at science, education and innovation policies towards new challenges that require a strong collaborative and pre-competitive approach.

Long-term challenges, namely those with current direct implications for US and EU firms (large and small), researchers and universities include the emerging opportunities associated with the democratization of human genome sequencing and the emergence of personalized medicine throughout the world, as well as the increasing convergence between heath sciences, physical sciences and engineering. But also sustainable energy systems worldwide should be a subject of priority for government intervention and innovation policies with a great potential for global impact.

In this regard, and following the emerging discussion about the future of S&T, it is clear that, by and large, the financing of higher education and of science and innovation has occurred along rather traditional lines. Yet, the history of science is rich with varied means of financing science and technological innovation. More importantly, developments in the size, integration, and technologies available in global capital markets present the opportunity to think about new financing possibilities and processes of societal engagement in S&T policies. These involve moving from traditional, “one-way” (and most of the time fragmented) government policies, to integrated multi-stakeholder policies involving a wide range of public and private agents.
Openness: research and education beyond national borders

Looking at the present and tentatively forecasting the future, we argue in this brief paper that a new paradigm of international academic, scientific and technological cooperation that seems to emerge as a major shaping factor for development at an unprecedented level.

Leading American research universities are playing a key role in this process worldwide, as a result of the accumulation of large investments in research and education over many decades now. It should be clear that this is not a new issue. For example, Morgan (1979) describes the role US universities played in helping to build and indigenous S&T base in developing countries until de 70’s and how far American Universities has engaged into that process. Some thirty years ago, Morgan recommended universities and policy makers about the future involvement on four areas: institutional building, cooperative R&D, resource base development, and education and training.

More recently this theme has been subject of various books and papers in the technical literature and, for example, the analysis of Bruce Johnstone, Altbach et al, as well as that of Knight, shows an active participation of US and EU universities in indigenous and local development practices, indicating related major advantages, as well as major challenges for them and national innovation policies in the near future. A recent report by the Royal Society further emphasizes these aspects in terms of scientific collaboration.

This emerging models of research and academic cooperation, that includes but do not seem to be a hostage of the traditional forms of services’ international commerce, may derive their uniqueness from the very nature of academic communities and from the strong meritocratic and universalistic ideals that prevail in science on an international scale. In addition, they are also driven by the flow of students and researchers, and by the citizen sense of being part of a “mission” for scientific and social development that motivates some of the best professionals in academic and research institutions worldwide.

It is under this context that innovation policies should help fostering a better understand of future international collaborative paths in education, science and innovation. Ultimately, this will become a key issue for competitiveness everywhere.

Summary

This paper attempts to enlighten new insights in science and technology policy. It ultimate goal is to promote the discussion beyond “national systems of innovation” and to help clarifying the emerging diversity of policies and increasing institutional specialization and clarification of the role of private and public incentives to support R&D.

We argue for the need to promote and integrate public and private strategies in modern societies fostering a non-hierarchical integration of formal policies and informal system linkages leading to knowledge-driven societies. This requires
opening-up science policies to multiple public and private agents and includes the continuous adaptation of systems of competence building and advanced studies, among which promoting global research networks should be highlighted.

References

1. OECD (2010) Economy: Developing countries set to account for nearly 60% of world GDP by 2030, according to new estimates. http://www.oecd.org/document/12/0,3746,en_2649_33959_45467980_1_1_1_1,00.html


7. See, for example, Galbraith, J. (2012), "Inequality and instability – a study of the world economy just before the great crisis", Oxford University Press.


15. Berger, S. (2005), "How we compete – what companies around the world are doing to make it in today’s global economy", Doubleday Press.


Chapter 2

Samuel Niza, Leonardo Rosado, André Pina and Paulo Ferrão, “Towards a standard method for urban material flow accounting”.
Towards a standard method for urban material flow accounting

Samuel Niza, Leonardo Rosado, André Pina, Paulo Ferrão

Abstract

The article presents the methodological advances towards the development of a straightforward method to assess the resource flows of urban areas. The main gaps found in urban metabolism studies are identified and the solutions developed to solve them are presented. The stages that constituted the main advances performed by the IN+ team to develop a comprehensive systemic perspective to urban metabolism studies will be described (Niza et al., 2009; Rosado et a., 2013; Pina et al., 2013). Described methods avoid the frequently adopted bulk MFA perspective, by establishing a new methodological framework intended to quantify in detail urban material flows and stocks and its correlation with economic activities.

1. Introduction

The Urban Metabolism term was first used by Wolman (1965) in a famous article in Scientific American. In his pioneer work the author quantified the flows of energy and materials into and out of a hypothetical American city enhancing the importance of this approach to study cities. Nevertheless, since then, when compared to the large number of MFA studies on the national level, studies focusing the regional or local level are still very limited. The available studies though show the importance of material flows on regional and urban metabolism but also present a large spectrum of approaches that can be defined using the Material Flow Accounting approach (Graedel 1999).

The concept of Urban Metabolism is based on a material flows model of the interrelations between the economy and the environment, where the economy functions like an environmental subsystem, depending on a permanent throughput of materials and energy (Daly 1996). Urban regions are unique in that they are typically characterized by a high metabolic rate. They induce high energy and material flows largely due to high population densities and their large and relatively dense material stocks. In order to promote a sustainable material flows management of urban regions, it is vital to understand the urban metabolism and this requires a detailed knowledge of main material flows.

Material flows accounting (MFA) allow picturing the material consumption of a system for a certain base year - corresponding to a static analysis of material flows - but also permits evaluating trends of material consumption of the economic system by developing time series. It can thus be considered a tool providing, simultaneously, data disaggregation and a timely response allowing the characterization of the dynamics of an economy metabolism (Niza and Ferrão 2005).
According to the first thermodynamics law (conservation of mass), the total of inputs must, by definition equal the total of the outputs plus the net accumulation of materials in the system:

\[
\text{Input} = \text{Output} + \text{Stock increases} - \text{Stock decreases}
\]

This material balance principle is as true for the whole economy as for any of its subsystem (economic sectors, company, household etc.). EUROSTAT (2001) defined as the main input material flows indicators the Direct Material Input (DMI) and the Domestic Material Consumption (DMC). The main output indicators are the Domestic Processed Output (DPO) and the Exports. Stock changes are accounted as Net Addition to Stock (NAS).

The DMI measures the direct input of materials for use into the economy, i.e. all materials which have an economic value and are used in production and consumption activities; DMI equals domestic extraction plus imports.

The DMC measures the total amount of material directly used in an economy for own consumption. DMC equals DMI minus exports and is defined in the same way as other key physical indicators such as gross inland energy consumption.

The DPO measures the total weight of materials, extracted from the domestic environment or imported, which have been used in the domestic economy, before flowing to the environment. These flows occur at the processing, manufacturing, use, and final disposal stages of the production-consumption chain.

DMI is not to be added across economies since it includes the fraction of materials that are going to be exported by that economy (therefore imported by an economy that could be being compared with). For example, to account the DMI of the European Union the intra-EU foreign trade flows must be netted out from the DMIs of Member States.

When accounting the metabolism of an urban area researchers are faced with certain constraints, particularly if the objective is obtaining results with a disaggregation level that allows supporting policy making.

Two main gaps are identified in the articles dedicated to review existing urban metabolism studies. First, it is clear from literature (e.g. Kennedy et al. 2011; Niza et al. 2009; Barles 2010 and Weisz and Steinberger 2010) that urban metabolism accounting lacks unified methodology. Different studies use different methodologies and this is likely to be detrimental to the consolidation and reliability of results. Without a unified methodology, urban metabolism studies will fall under a case-by-case basis, failing to draw more general conclusions that might contribute to the design of more sustainable urban systems.

The second gap, which strongly contributes to the first, concerns the lack of data. Kennedy et al (2011), for instance, mention that despite the importance of cities for the overall volume of socioeconomic material and energy use, the harmonized datasets are provided by statistical offices almost exclusively to the national level. Therefore attempts to generalize patterns and trends of urban-specific resource use from the literature have so far proven to be incomplete.
Niza et al. (2009) pioneered a method developed specifically for quantifying urban material flows based on the Eurostat (2001) methodology, downsizing it from economy-wide to urban material flows. It was tested in a case study that characterized the urban metabolism of the city of Lisbon by quantifying its material balance for 2004. This study, like the majority of the urban MFA literature (e.g., Barles 2009, BFF 2002, Hammer and Giljum 2006), adopted a categorization of materials in five classes following the aggregate material classes used in Eurostat (2001). This limitation, or gap (the third in this review), considerably reduces the practical use of its results, namely for waste management or sustainable consumption policies. For example, aluminum and steel are two material types with a significant economic value, although fundamentally different, as they require significantly different techniques to be recycled. In these studies there is no discrimination between them as they are categorized as metallic minerals. Also, the aggregate fossil fuels materials class does not identify the amount of plastics that could be recycled, nor the fossil fuels used for energy, which are responsible for CO2 emissions and cannot be retrieved for waste management purposes.

On the other hand, the decomposition of material consumption per economic activity was performed for much-aggregated sectors allowing very rough conclusions about which ones are the main drivers of material consumption. This constitutes the fourth gap of current urban flow accounting methodologies.

Another gap is associated with a limited understanding of the origin and destination of flows within the urban boundaries and the magnitude of the crossing flows (i.e. those materials that are imported and to a great extent re-exported (Hammer and Giljum, 2006). Urban MFA studies account the flows when entering or leaving the system but there is no characterization of the flows inside the system. For example, the origin of imported materials and the municipality where they are locally extracted, the materials that are transformed by the local industry and the product destination at their end of life are not usually identified. This fifth gap is detrimental to establishing policies for a more sustainable use of urban material flows.

The sixth gap relates to the lack of understanding about the dynamics of the added stocks to cities since the studies performed generally provide a static picture of material stocks. They show materials entering the economy and the accumulation in the stock but do not provide information about the dynamics of the consumption-throwaway cycle, losing potential insights that can be obtained by understanding future materials availability from products reaching their end-of-life.

The spatial scale of the studies determines additional constraints. First, the lack of commodities statistics at a low spatial scale demands estimating values. This estimation may bias material accounting since an urban area concentrates many of the regional and national administrative offices. Additionally, daily commuters, plus the large amount of flows from logistic infrastructures (e.g. big harbors, train stations or airports) to be distributed not only in the city but elsewhere, be it the rest of the country or other countries, may potentially contribute to over-estimate the material consumption. This can be identified as the seventh gap: flows that are accounted within the urban limits, but are in transit to elsewhere.

In this article the stages that constituted the main advances performed by the IN+ team to develop a comprehensive systemic perspective to urban metabolism studies will be
described (Niza et al., 2009; Rosado et al., 2013; Pina et al., 2013). Described methods avoid the frequently adopted bulk MFA perspective, by establishing a new methodological framework intended to quantify in detail urban material flows and stocks and its correlation with economic activities.

The work is presented in a chronological order from the first method developed to account the material flows of a city, through the method developed to account the flows of metropolitan areas mainly using EUROSTAT data, to the method developed to perform the assessment of the metabolism of metropolitan areas using national IO tables. The second method considerably increased the possibility of decomposing the material consumption per economic activity and municipality of the metro area while the third method established a streamlined way of accounting the material flows of any metro area in the world.

2. First stage: Uncovering urban material flows

The main purpose of the first stage of research was to establish a smart methodological framework to depict urban areas, relying on published statistical data and based on the EUROSTAT (2001) methodology. A significant effort has been carried out to extend that method, given that it was originally designed to account economy-wide material flows, and the current work focused on the city level, excluding all its surrounding municipalities of the metropolitan area or region.

It can be assumed that all materials consumed in a city come from outside its limits, including its surrounding municipalities. It may thus be considered that domestic extraction in the city is residual and may be defined as non-existent in material flow accounting terms.

Consequently, Import and Export categories assume higher relevance in the quantification of the city material flows. It is therefore crucial to quantify the amounts of goods exchanged between the city, the rest of the country and the rest of the world.

A city harbor, for instance, is an input and output gateway for a large amount of commodities to and from the country. Consequently a large amount of commodities reaching the town just cross it and should not be accounted as part of the city material flows, in order to avoid an overestimation of the city citizens and economic activities material consumption.

As domestic extraction is almost non-existent and local production is usually considerably low, exports of materials derived from the city are residual or non-existent, in terms of material flows.

1 From Niza et al., 2009.
Material inputs to the city consist of products for internal consumption and raw materials for local production. Outputs are emissions and wastes, to be processed in the surrounding municipalities. The remaining flows, that may assume considerable amounts, are products that just cross the town towards their destination.

From the previous paragraphs it may be inferred that the main catalyzer of material flows in a city is consumption and, in a much lesser degree, local production. Consequently, imports into town (when subtracted of imports that just cross it) are a proxy of the city consumption.

The developed methodology allows characterizing the following variables associated to a city material flow:

- Absolute consumption and final disposal of materials, per material category;
- Throughput of materials, per material category;
- Activity sectors material consumption;
- Waste treatment per material category and treatment type.

Each variable is processed making use of a matrix built up with the data described in the following paragraphs: the Materials Matrix provides the input and output flows of materials; the Throughput Matrix provides the materials that are added to the city materials stock; the Waste treatment matrix distributes wastes according to three treatment categories and the Activity sectors matrix intended to distribute materials consumption through different economic sectors.

The first two variables allowed for the quantification of the Material Balance, and the other two indicators are useful to characterize the city material consumption.

### 2.1 Materials matrix
The consumption/disposal accountability is formulated as a materials matrix, which was designed as a product of three matrices:

- the products composition matrix, $A_{ij}$,
- the mass flow matrix, $P_{jk}$, and
- the city quota matrix, $L_{jl}$.

**Equation 1:** \[ M = A_{ij} \times P_{jk} \times L_{jl} \]

The products material composition matrix, $A_{ij}$, where $i$ are material categories and $j$ are products, determines the composition of each product (input flows) and wastes (output flows) per material category, in mass percentage. Material categories considered are Biomass, Fossil fuels, Metals and Non-metallic minerals.

The mass flow matrix, $P_{jk}$, where $j$ are products and $k$ are mass amounts, is a diagonal matrix. Whenever $j=k$, $P_{jk}$ defines the mass of products entering the city, in a given period, and the amounts of wastes leaving it. When $j \neq k$ then $P_{jk} = 0$.

Finally, the city quota matrix, $L_{jl}$, where $j$ are products and $l$ are mass percentages, is also a diagonal matrix. When $j=l$ then $L_{jl}$ defines the fractions of each product and waste that is consumed and produced within the city. When $j \neq l$ then $L_{jl} = 0$.

### 2.1.1 Products composition matrix

The material composition of products is mainly characterized by making use of Annual Surveys on Industrial Production, as this database (for instance in Portugal) presents data in a disaggregated way allowing an assessment per product. Composition is defined according to EUROSTATs aggregated material categories: Biomass, Fossil fuels, Metals and Non-metallic minerals.

The most representative material categories in terms of weight are estimated for each set of products (grouped by the CEA – Classification of economic activities6 and designated according to the Combined Nomenclature, CN). It is assumed an average composition for products of the same CEA, avoiding a case by case analysis of about three thousand different products.

Whenever possible, manufacturers' information about products composition is consulted (e.g. chemical products). For some cases, information provided by specific waste management entities is used.

This material composition is then used for each input: products from International trade, National freight transport and Local industrial production; and output: Industrial wastes and Municipal solid wastes. Due to the aggregation degree of International trade and National Freight Transport databases, average compositions of similar products previously estimated for Industrial Production are used. For Industrial waste, the composition per material categories is based on the European Waste List;

The city electricity consumption is converted in the fossil fuels amount needed to produce it.

### 2.1.2 Mass flow matrix
This matrix characterizes the amounts (in tons) of products imported and transported in the region the city is included and wastes produced in the area, during one year.

### 2.1.3 City quota matrix

Data sources used for calculations include different geographic scales, so data has to be processed in order to characterize only urban flows. The methodology developed consisted in producing a diagonal matrix, $L_{ii}$ ($i$ are products) that quantifies the fractions of products and wastes that are consumed and produced in the municipality, in percentage. This matrix is quantified considering that the city consumption is a function of the number of workers, the number of inhabitants or their purchasing power.

Both International trade and National Freight Transport refer to the Region. In order to obtain the municipality share, a distribution of materials/products imported or transported regionally per destination activities is produced.

For wholesale and local industrial production, it was assumed that the number of workers is the limiting factor for consumption. For retail, it is considered that together with the number of workers, the limiting factor is also the inhabitants’ purchasing power. As a consequence, the value obtained by the estimation made through the ratio based on the number of workers is multiplied by the ratio of the purchasing power in the city and the purchasing power of the region.

For each Import category, destination activities are grouped. Imports to the municipality (in percentage) are obtained from the ratio between the number of workers of these activities in the municipality and in the region.

Local production in the municipality is estimated by defining for each activity a relation between the country numbers of workers and the city number of workers.

Having assumed that there are no products exported, exports equal wastes produced in the city.

### 2.2 Throughput matrix

Throughput is referred here as the product life span in the city. This results from the conjugation of:

- The product material composition, namely if it is composed of fast degrading materials (less than one year) or not, and,
- The kind of use, or function. For instance, there are products that become unusable after its first use (e.g. food packaging or cleaning products packaging) and others that last decades (e.g. buildings).
This classification is important to estimate materials added to the city material stock. The material flows’ life span is divided in four categories:

1. Flows leaving the economy before a year after their input (e.g. food, packaging, oil as fuel);
2. Flows leaving the economy after one year and before 10 years after their input (relatively durable goods as toys, computers etc.);
3. Flows leaving the economy between 11 and 30 years after their input (durable goods as machines, cars or home appliances);
4. Flows remaining in the economy for more than 30 years (long duration goods as buildings or communication infrastructures).

Distribution of products according to their throughput velocity is computed as the product of two matrices:

- The materials matrix, M;
- The products life span matrix, Rlj.

Equation 2: \[ T = M \times R_{lj} \]

In R_{lj}, \( l \) represents life span categories and \( j \) the products. Literature with information about products life span is scarce, exception made to Cooper (2005) and Hsu & Kuo (2005). Distribution of products according to the four life span categories was therefore based in these references and resulted from specific research by the authors.

### 2.3 Waste treatment matrix

Waste treatment matrix was designed as a product of three matrices:

- The waste composition matrix, \( V_{im} \) - where \( i \) are material categories and \( m \) are wastes, determines the composition of each waste per material category, in mass percentage;
- The waste flow matrix, \( X_{mn} \) - where \( m \) are wastes and \( n \) are mass amounts;
- The treatment category matrix, \( T_{ms} \) - where \( m \) refers to wastes and \( s \) the kind of treatment.

Equation 3: \[ W = V_{im} \times X_{mn} \times T_{ms} \]

In \( T_{ms} \), industrial waste and municipal solid waste data is distributed according to the following treatment categories: Recycling, Incineration (waste-to-energy recovery) and Controlled landfill.

### 2.4 Activity sectors matrix

Distribution of products per sector is computed as the product of two matrices:

- The materials matrix, M;
The products per sector matrix, $S_{ih}$.

Equation 4: $S = M \times S_{ih}$

In matrix $S_{ih}$, $i$ represents products and $h$ sectors. Materials were distributed according to three activity sector categories: a) Restaurants, Hotels and Services, b) Housing and c) Industry and Construction. Again, distribution of imports to Lisbon per sector (in %) was based in the number of workers in wholesale, retail and specific industrial activities in Lisbon.

The described method allowed namely quantifying the Lisbon material flow balance for 2004.

Figure 2 – Lisbon Material Balance, 2004 (Niza et al., 2009)
3. Second stage: UMAN (Urban Metabolism Analyst) – opening the black box

The UMAN method is based in four main components in which a set of nested calculation routines provide an adequate processing of statistically available data to elaborate a detailed map of material resources and throughput in different economic activities. The sub models developed are:

- Platform – theoretical foundations of the model specifying the adjustments made to the economy-wide MFA, addressing specific issues of urban areas, particularly the relevance of imports, national and international, and also the potential existence of significant cross flows.
- Statistics – data sources needed to operate mathematical calculations in the model.
- Plugins – databases describing the flows of products and goods (according to the standard nomenclatures used by OECD and Eurostat) in terms of their material composition, average lifespan, and life-cycle phase.
- Calculator – the operational core of the model, involves four sequentially nested steps that use the MFA method, the statistics and the built plug-in databases, allowing accounting several variables, namely: products and materials and related MFA indicators; throughput projections based on products and materials added to the stock; assignment of products and materials to the economic activities of the metropolitan area and, the spatial disaggregation of the products and materials.

The proposed UMAN model is intended to overcome the methodological gaps identified as described in Table 1.

<table>
<thead>
<tr>
<th>Gaps</th>
<th>UMAN developments to bridge the gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Lack of a unified methodology</td>
<td>Using statistical data provided by OECD and Eurostat to perform urban MFA, allows for the establishment of a model that has the potential to be universally applied.</td>
</tr>
<tr>
<td>2. Lack of material flows data at urban level</td>
<td>The use of generalized national statistics coupled with regional statistics allows for the extrapolation of consumption within the metropolitan areas, hence providing a way of generalizing patterns of consumption for urban areas.</td>
</tr>
<tr>
<td>3. Limited discrimination of material types</td>
<td>The disaggregation and harmonization of material types using the material composition plug-in database and the accounting of products allows for measuring flows beyond 5 highly aggregate material types to 28 material types.</td>
</tr>
<tr>
<td>4. Limited resolution of economic activities consumption</td>
<td>National transportation and international trade statistics are used to provide for attributing an economic sector to each product transaction.</td>
</tr>
<tr>
<td>5. Limited understanding of the origin and destination of flows</td>
<td>The characterization of the supply chain is made by combining the products and materials decomposition and the phase of the product plug-in database. This allows identifying, within the metropolitan area, the economic activities involved to manufacture final products for consumption.</td>
</tr>
<tr>
<td>6. Lack of understanding about the dynamics of added stock</td>
<td>The measurement of the dynamics of stocks is performed by combining the lifespan plug-in database and the throughput description parameter, allowing measuring the potential amount of materials becoming obsolete each year.</td>
</tr>
</tbody>
</table>

---

2 From Rosado et al., 2013
In the next paragraphs the components and intermediate phases of the model are presented.

### 3.1 Platform

MFA method plays an important role in the proposed Urban Metabolism model but it needs to be adapted from the economy wide to urban areas. Urban areas are regions where the amount of materials extracted and the production sector are of a lesser importance than that of entire countries (e.g., Barles 2009). Potentially a large part of the material flows in an urban area originate from outside its limits (Bai 2007): Import and export categories assume a higher relevance in the quantification of material flows at this scale of analysis. Furthermore, these areas can have flows that cross the defined boundary toward their destination. The crossing flows can assume a great importance in metropolitan areas where the “Rotterdam effect” is felt (as is the case of Lisbon (Niza et al. 2009) or Hamburg (Hammer and Giljum 2006)), referring to the role of big harbors (e.g., Rotterdam or Antwerp) serving as gateways for international trade (Weisz et al. 2005).

Additionally, since metropolitan areas are part of a country, a split between the imports and exports from foreign countries and other regions of the country should be performed.

One of the major issues here relates to the openness of cities, i.e., the lack of a clear boundary in terms of existing statistical information. Since data for imports and exports are not explicit about their final destination or origin, it is difficult to understand if imports are staying within the economy or just passing through, likewise for exports.

### 3.2 Statistics

Data sources used by the Eurostat and national statistical offices to compile economy-wide MFAs can be applied to the specificities of metropolitan areas but an additional major data source must be used: the Standard Goods Classification for Transport Statistics which is available at NUTS 2 level. The use of this source ensures a proper accounting of flows in and out of metropolitan areas that match a specific NUTS 2 level, since it tracks all movements of products and goods within the system. In the case where there is no overlap between NUTS 2 and the defined metropolitan area the MFA will have to be made to the NUTS 2 and when the disaggregation by municipality (NUTS III) is made, further extrapolation is possible to attain and account for the Metropolitan area, using the spatial decomposition part of the UMan model.

Some of the data sources, to estimate the domestic extraction in the urban area, are directly used while others involve extrapolation from the national level to the metropolitan area. Again, to extrapolate the data from the national level it was, assumed a linear relation with the number of workers for the economic activity. In this case, the weight will be a percentage of the total amount equal to the percentage of workers of that economic activity in the metropolitan area. For instance the absence of workers in the economic activity Mining of iron ores (13.10 NACE code, rev 1.1) suggest that, there is no iron extracted within the metropolitan boundaries.
3.3 Plugins

In order to provide the material disaggregation, the lifetime and the life-cycle dimensions of the model, three supporting databases were built: one characterizes the material composition of products; another identifies the average lifespan of products; and another identifies the life-cycle phase of each product. These databases describe 13,135 types of products, corresponding to the fifth level of disaggregation of the combined nomenclature used in the International Trade Statistics (European Commission 2011).

3.3.1 Material Composition

A new nomenclature with 28 material categories was developed to decompose products into materials, the MatCat (Table 2). This nomenclature represents:

1. a balance between the aggregation of the nomenclature defined by the Eurostat (2001) for imports and exports and the disaggregation of the whole spectrum of material categories defined by the same agency for domestic extraction (e.g., for minerals it includes 153 categories);
2. a resource-management-based classification, a proxy to the material recovery potential of the materials that compose products.

<table>
<thead>
<tr>
<th>Fossil Fuels (FF)</th>
<th>FF1</th>
<th>Low ash Fuels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF2</td>
<td>High ash Fuels</td>
</tr>
<tr>
<td>FF3</td>
<td></td>
<td>Lubricants and Oils and Solvents</td>
</tr>
<tr>
<td>FF4</td>
<td></td>
<td>Plastics and Rubbers</td>
</tr>
<tr>
<td>Metals (MM)</td>
<td>MM1</td>
<td>Iron, Steel Alloying Metals and Ferrous Metals</td>
</tr>
<tr>
<td>MM2</td>
<td></td>
<td>Light Metals</td>
</tr>
<tr>
<td>MM3</td>
<td></td>
<td>Non-Ferrous Heavy Metals</td>
</tr>
<tr>
<td>MM4</td>
<td></td>
<td>Special Metals</td>
</tr>
<tr>
<td>MM5</td>
<td></td>
<td>Nuclear Fuels</td>
</tr>
<tr>
<td>MM6</td>
<td></td>
<td>Precious Metals</td>
</tr>
<tr>
<td>Non-metallic minerals (NM)</td>
<td>NM1</td>
<td>Sand</td>
</tr>
<tr>
<td></td>
<td>NM2</td>
<td>Cement</td>
</tr>
<tr>
<td></td>
<td>NM3</td>
<td>Clay</td>
</tr>
<tr>
<td></td>
<td>NM4</td>
<td>Stone</td>
</tr>
<tr>
<td></td>
<td>NM5</td>
<td>Other (Fibers, Salt, inorganic parts of animals)</td>
</tr>
<tr>
<td>Biomass (forestry, crops and animal products) (BM)</td>
<td>BM1</td>
<td>Agricultural Biomass</td>
</tr>
<tr>
<td></td>
<td>BM2</td>
<td>Animal Biomass</td>
</tr>
<tr>
<td></td>
<td>BM3</td>
<td>Textile Biomass</td>
</tr>
<tr>
<td></td>
<td>BM4</td>
<td>Oils and Fats</td>
</tr>
<tr>
<td></td>
<td>BM5</td>
<td>Sugars</td>
</tr>
<tr>
<td></td>
<td>BM6</td>
<td>Wood</td>
</tr>
</tbody>
</table>
MatCat can be formulated as a material composition matrix (\(M\)) where \(n\) represents each product of the combined nomenclature (comprehending 13,135 products) and \(m\) represents its material composition in 28 categories (Table 2) as a percentage of the total weight (Equation 5).

Equation 5: 
\[
M_{n,m} = \begin{bmatrix}
m_{1,1} & m_{1,2} & \cdots & m_{1,m} \\
m_{2,1} & \vdots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
m_{n,1} & m_{nm} & \cdots & m_{nm}
\end{bmatrix}
\]

with \(n = 1, \ldots, 13,135\) products and \(m = 1, \ldots, 28\) material categories

In order to create this matrix extensive literature research was conducted. They include UNU (2008), European Union (2011), and Defra (2009), among others. As an example, according to Defra (2009), the combined nomenclature category 87032190 “Motor cars and other motor vehicles principally designed for the transport of persons” has the following weight ratios of materials: 4% of Lubricants and oils and solvents; 22% of Plastics and rubber; 58% of Iron, Steel Alloying Metals and Ferrous metals; 2% of Light metals; 4% of Non-ferrous heavy metals; 1% of Sand (in the form of glass); 4% of Textile biomass and 5% non-specified.

### 3.3.2 Products Lifespan

To understand the dynamics of resource flows and improve the knowledge of future availability of resources (be they for reuse, recycling, or recovery) a lifespan database was developed. This database includes information about the average lifespan of products (in years) and a distribution function for each. The chosen distribution function was the Weibull (as proposed by some authors, e.g., Elshkaki et al. 2005; Melo 1999) since it provides a good fit for the lifespan of most of the products. The distribution is described by Equation 6:

Equation 6: 
\[
f(x; \lambda, k) = \begin{cases} 
k \frac{x^{k-1}}{\lambda^k} e^{-\left(\frac{x}{\lambda}\right)^k}, & x < 0 \\0, & x \geq 0
\end{cases}
\]

Value \(x\) represents the average lifespan of a product and the shape parameter \(k\) value is chosen arbitrarily based on values used in the scarcely available literature (e.g., Melo 1999; Davis et al. 2007), to adjust the \(k\) value to empirical information for each product (Melo 1999). The value 10 was chosen for most of the products to measure the potential amounts flowing out of the system on an annual basis. A throughput matrix (Equation 7)
was built (T) where \( n \) is product types and \( y \) the percentage of products becoming obsolete each year.

**Equation 7:**

\[
T_{n,y} = \begin{bmatrix}
t_{1,1} & t_{1,2} & \cdots & t_{1,y} \\
t_{2,1} & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
t_{n,1} & \cdots & t_{(n-1),y} & t_{n,y}
\end{bmatrix}
\]

with \( n = 1, \cdots, 13,135 \) products and \( y = 1, \cdots, x \) years

Taking the mentioned example (Defra 2009), the combined nomenclature category 87032190 has an average lifespan of 12.5 years. Its obsolescence is distributed annually according to the Weibull distribution.

### 3.3.3 Life Cycle Phase

Materials and products from abroad (be they from national or international origin) enter the metropolitan economy in different life cycle phases (e.g., raw materials or final goods). As final goods they will eventually become wastes (namely municipal solid wastes) and emissions. As raw materials or preprocessed goods, they will be transformed within the metropolitan area contributing to industrial wastes and emissions, becoming part of the final goods that are consumed within the metropolitan area or exported. In order to assure a correct balance of materials it is important to identify the final goods going directly to consumption and the amount of preprocessed goods, so that they can be balanced with sectorial wastes and emissions and exports. The state of products characterization matrix \( S \) was built to perform the assignment of phases to products (Equation 8) with \( n \) being the product and \( g \) the percentage in the following product state (life cycle phase) categories: livestock \((g=1)\), raw materials \((g=2)\), intermediate products \((g=3)\), final goods \((g=4)\), and waste \((g=5)\).

**Equation 8:**

\[
S_{n,g} = \begin{bmatrix}
s_{1,1} & s_{1,2} & \cdots & s_{1,g} \\
s_{2,1} & \ddots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
s_{n,1} & \cdots & s_{(n-1),g} & s_{n,g}
\end{bmatrix}
\]

with \( n = 1, \cdots, 13,135 \) products and \( g = 1, \cdots, 5 \) state classes of goods

Again, using the same example for the combined nomenclature category 87032190 it can be identified that this is a final good.

### 3.4 Calculator

#### 3.4.1 Materials Accounting

Domestic extraction is calculated using harvesting and production data of the urban area (e.g., agricultural production and industrial extraction) whenever available. If not available, estimation may be based on MFA national values (Eurostat 2001) weighted by
the number of workers per economic activity and additional information, like the purchasing power, as in Niza et al., (2009).

Transport statistics are used to measure imports and exports. These are Eurostat statistics available in 20 categories of the Standard Goods Classification for Transport Statistics nomenclature (NST).

The calculation of the Domestic Material Consumption, disaggregated in the NST nomenclature \( (DMC_{nst}) \), is performed considering the amounts of national \( (nat) \) and international \( (int) \) imported products \( (IMP) \), domestically extracted materials \( (DE) \), and national and international exported products \( (EXP) \). The DMC is formulated mathematically in Equation 9, where \( nst \) refers to each of the 20 product categories defined in the NST nomenclature.

\[
DMC_{nst} = DE_{nst} + IMP_{nst, int} + IMP_{nst, nat} - EXP_{nst, int} - EXP_{nst, nat}
\]

Equation 9:
\[
\begin{bmatrix}
d_1 \\
d_2 \\
\vdots \\
d_{nst}
\end{bmatrix}
+ \begin{bmatrix}
imp_{1, int} + imp_{1, nat}

imp_{2, int} + imp_{2, nat}

\vdots 

imp_{nst, int} + imp_{nst, nat}
\end{bmatrix}
- \begin{bmatrix}
exp_{1, int} + exp_{1, nat}

exp_{2, int} + exp_{2, nat}

\vdots 

exp_{nst, int} + exp_{nst, nat}
\end{bmatrix}
\]

with \( nst = 1, \ldots, 20 \) classes of products

After, the DMC in NST is converted to the combined nomenclature (of the International Trade Statistics, describing all traded products in an economy). This step allows obtaining the DMC for 13,135 products. The distribution is done assuming a mass percentage related to the sum of all CN product categories that belong to the respective NST category:

- for international trade \( (it) \) – ratios are defined by the international trade \( (int) \), for imports \( (its) \) and exports;
- for national trade \( (nt) \) – ratios are defined by the domestic extraction \( (de) \), industrial production \( (np) \) and international trade \( (its) \). National import ratios are calculated taking into account the mixture of products that are extracted, produced and imported by the country \( (nat) \). Similarly, national exports ratios are calculated taking into account the mixture of products extracted, locally produced and exported by the country.

In mathematical terms, Equation 10 and Equation 11 present the distribution for international trade imports and national trade imports.

\[
imp_{it, nst} = \frac{its_n}{\sum its_n \cap nst} \times IMP_{nst, int}
\]

with \( n = 1, 2, \ldots, 13,135 \) products \( (CN) \); \( nst = 1, \ldots, 20 \) products \( (NST) \); \( int = \) international imports
Equation 11:

\[ \text{impnt}_{n,\text{nst}} = \frac{\text{its}_n + \text{n}\text{p}_n + \text{de}_n}{\sum (\text{its}_n \cap \text{n}\text{st} + \text{n}\text{p}_n \cap \text{n}\text{st} + \text{de}_n \cap \text{n}\text{st})} \times \text{IMP}_{n,\text{nst, nat}} \]

with \( n = 1, 2, \ldots, 13,135 \) products (CN); \( \text{nst} = 1, \ldots, 20 \) products (NST); \( \text{nat} = \text{national imports} \)

Equation 12 provides the domestic material consumption of 13,135 products in CN (DMCn).

Equation 12:

\[ \text{DMC}_n = \text{DE}_n + \text{IMP}_n - \text{EXP}_n = \begin{bmatrix} \text{de}_1 \\ \text{de}_2 \\ \vdots \\ \text{de}_n \end{bmatrix} + \begin{bmatrix} \text{imp}_{1,1} + \text{impnt}_1 \\ \text{imp}_{1,2} + \text{impnt}_2 \\ \vdots \\ \text{imp}_{n,1} + \text{impnt}_n \end{bmatrix} - \begin{bmatrix} \text{exp}_{1,1} + \text{expnt}_1 \\ \text{exp}_{1,2} + \text{expnt}_2 \\ \vdots \\ \text{exp}_{n,1} + \text{expnt}_n \end{bmatrix} \]

with \( n = 1, \ldots, 13,135 \) products

Since the combined nomenclature refers to goods in different life cycle stages (e.g. raw materials, intermediate goods, final goods), a fraction of the DMC are final goods and the other is raw materials or intermediate products, that are likely to be transformed by the industry into final goods. In order to understand which are going to be part of durable goods and which are going to be part of non-durable goods, local production data is used to simulate the processing of these raw and intermediate goods into final goods. The result is the local production of final goods. This step is critical 1) to account the materials that will constitute the material stock of the urban area (materials that will be used by the urban economy for more than a year) and 2) to help closing the balance with wastes and emissions.

Using the state of products characterization matrix (\( S_{n,g} \)), as depicted in Equation 8, and multiplying it by a diagonal matrix with the domestic material consumption (Equation 12) it is developed a matrix splitting the DMC products into the different states (DMCsp) (Equation 13).

Equation 13:

\[ \text{DMCsp}_{n,g} = \text{DMC}_{n,n} \times S_{n,g} = \begin{bmatrix} \text{dmc}_{1,1} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \text{dmc}_{n,n} \end{bmatrix} \times \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,g} \\ s_{2,1} & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,g} \end{bmatrix} \]

with \( n = 1, \ldots, 13,135 \) products \( g = 1, \ldots, 5 \) state of products

In order to describe the amount of non-final goods that are processed by the local industry it is used the Industrial Production Statistics (IPS) dataset. This set is available at the NUTS I (National) level and uses the PRODCOM nomenclature. From this, the ratio of the number of workers at the NACE 4 digit level is used to estimate the local production (LP) for the NUTS II.

In Equation 14, final goods locally produced and allocated to the domestic material consumption (DMCp) are estimated. These represent the fraction of the total final goods
locally produced (LP) that matches the amounts accounted for non-final goods (including waste) in the split matrix (DMC_{op}) and the emissions for the local industrial sector (E).

Equation 14:

\[ \text{DMC}_{lp,n,g} = \text{LP}_{n,g} \times \left( \frac{\sum \text{SP}_{n,g}, g \neq 4}{\sum \text{LP}_{n,g} + \sum E} \right) \]

with \( n = 1, \ldots, 13,135 \) products \( g = 1, \ldots, 5 \) state of products

The remaining local production of final goods is exported (EXP_{lp}). The mathematical operations presented so far allow the making of the material balance for a metropolitan area. The Domestic extraction (DE), national and international imports (IMP) and exports (EXP) (including waste) and emissions (E) are computed. A material balance with crossing flows can be represented by Equation 15.

Equation 15:

\[ \text{DE}_{n} + \text{IMP}_{n} + \text{IMP}_{n} - \text{EXP}_{lt,n} - \text{EXP}_{n,t} - E = 0 \]

with \( n = 1, \ldots, 13,135 \) products

Including the local production it is possible to isolate a partial fraction of the crossing flows (EXP_{cf}) and a more detailed material balance can be formulated (Equation 16).

Equation 16:

\[ \text{DE}_{n} + \text{IMP}_{n} + \text{IMP}_{n} - \text{EXP}_{lp,n} - \text{EXP}_{cf,n} - E = 0 \]

with \( n = 1, \ldots, 13,135 \) products

The material balance provided by this method allows the accounting of several variables: 1) domestic extraction, 2) imports and exports disaggregated by origin (national or international) and by product, 3) emissions to nature, and 4) locally produced final goods disaggregated by destination (domestic consumption and export). A measure of the crossing flows is possible to obtain and partially avoid the seventh identified gap. For the crossing flows amount to be fully described further data sources need to be used to identify, from the local production, the exact amount of non-final goods exported.

The identification and accounting of 1) the direct final goods, 2) the non-final goods that are transformed in the urban area, as well as 3) the identification of the amount of crossing flows provides information about the supply chain in the urban area, hence covering aspects identified as the fifth gap in this article. Finally through Equation 17, the final goods are decomposed in 28 material categories (MC). This is performed by multiplying the diagonal matrix of domestic material consumption of final goods (DMC_{fg}) with the material composition matrix (M).
Equation 17:

\[
\text{MC}_{z,m} = \text{DMCF}_{g_{z,z}} \times \text{M}_{z,m}
\]

\[
\begin{bmatrix}
dmcf_{g_{1,1}} & 0 & \cdots & 0 \\
0 & \ddots & \vdots & \vdots \\
0 & 0 & \ddots & dmcf_{g_{z,z}} \\
0 & 0 & \cdots & dmcf_{g_{z,z}}
\end{bmatrix}
\begin{bmatrix}
m_{1,1} & m_{1,2} & \cdots & m_{1,m} \\
m_{2,1} & \ddots & \vdots & \vdots \\
m_{z,1} & \ddots & m_{(z-1),m} \\
0 & 0 & \cdots & dmcf_{g_{z,z}}
\end{bmatrix}
\]

with \(z = 1, \ldots, 9,785\) final goods and \(m = 1, \ldots, 28\) material categories

From equation 17 it is possible to provide answers to the third gap by providing detailed material consumption in 28 material categories at the urban area.

3.4.2 Throughput

The materials throughput (MT) over time (Equation 18) is obtained by multiplying a diagonal matrix of each material of the final goods decomposed matrix (MC), with the throughput matrix (T) (Equation 7). A timeframe \(y\) of 50 years is chosen to show namely the throughput dynamics of stocks that includes infrastructures and buildings (on average, since in some cases they might have more than 50 years lifetime). Though, the model is flexible enough to provide the throughput dynamics for more or less years depending on the user needs or the case study.

Equation 18:

\[
\text{MT}_{z,m} = \overline{\text{MC}}_{z,z} \times \text{T}_{n,y} =
\begin{bmatrix}
m_{c_{1,1}} & 0 & \cdots & 0 \\
0 & \ddots & \vdots & \vdots \\
0 & 0 & \ddots & m_{c_{z,z}} \\
0 & 0 & \cdots & m_{c_{z,z}}
\end{bmatrix}
\begin{bmatrix}
t_{1,1} & t_{1,2} & \cdots & t_{1,y} \\
t_{z,1} & \ddots & \vdots & \vdots \\
t_{z,1} & \ddots & t_{(z-1),y} \\
t_{z,1} & \ddots & t_{(z-1),y} & t_{z,y}
\end{bmatrix}
\]

with \(z = 1, \ldots, 9,785\) final goods; \(y = 1, \ldots, x\) years; \(m = 1, \ldots, 28\) material categories

The computation of the materials throughput matrix described by Equation 18 provides information that covers the sixth gap.

3.4.3 Decomposition per Economic Activity

The International Trade Statistics (ITS) provides the destination economic activities (AD) for each imported product. The distribution of products per sector is then computed using this data source, assuming that imports from national trade have the same destination as international imports of the same type of goods. The distribution matrix (in percent) is described in Equation 19 where \(c\) refers to the economic activities as described in the level 2 of NACE nomenclature.

Equation 19:

\[
\text{AD}_{n,c} =
\begin{bmatrix}
ad_{1,1} & ad_{1,2} & \cdots & ad_{1,c} \\
ad_{2,1} & \ddots & \vdots & \vdots \\
\vdots & \ddots & ad_{(z-1),c} \\
ad_{z,1} & ad_{z,2} & \cdots & ad_{z,c}
\end{bmatrix}
\]

with \(z = 1, \ldots, 9,785\) products and \(c = 01, \ldots, 99\) level 2 digit NACE code
The bulk DMC fiz vector combined with the distribution into the different economic activities (EA) is described in mathematical terms in Equation 20. This operation allows describing economic activities in terms of their material consumption (tons).

Equation 20:

$$ \mathbf{E_A}_{z,c} = DMC_f g_{z,x} \times \mathbf{A_D}_{z,c} = \begin{bmatrix} dmc_{1,1} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots \\ \vdots & \ddots & 0 & \vdots \\ 0 & \cdots & dmc_{z,z} & \vdots \end{bmatrix} \times \begin{bmatrix} ad_{1,1} & ad_{1,2} & \cdots & ad_{1,c} \\ ad_{2,1} & \ddots & \vdots & \vdots \\ \vdots & \ddots & ad_{(z-1),c} & \vdots \\ ad_{z,1} & ad_{z,m} & \cdots & ad_{z,c} \end{bmatrix} $$

with $z = 1, \ldots, 9,785$ final goods and $c = 01, \ldots, 99$ level 2 digit NACE code

The allocation of the domestic material consumption to the different economic activities coupled with the information of the amounts of products and respective original economic activity allows an increase of information, helping solving the fourth and fifth gaps.

3.4.4 Spatial Decomposition

The Spatial distribution matrix (SD) is a function of the mix of economic activities in the sub-areas (NUTS 3) within the urban area. The decomposition of the material consumption per sub-area is based on the assumption that there is a linear relation between the number of workers per economic activity in each area and the total number of workers per economic activity at the NUTS 2 level (Equation 21) with sa referring to the NUTS 3 level sub-areas.

Equation 21:

$$ sd_{c,sa} = \frac{w_c}{\sum w_c} $$

with $sa = 1, 2, \ldots, n$ NUTS 3 level areas; $c = 01, \ldots, 99$ level 2 digit NACE and $w = \text{number of workers}$

The spatial distribution matrix (SD) is depicted in Equation 22.

Equation 22:

$$ \mathbf{S_D}_{c,sa} = \begin{bmatrix} sd_{1,1} & sd_{1,2} & \cdots & sd_{1,sa} \\ sd_{2,1} & \ddots & \vdots & \vdots \\ \vdots & \ddots & sd_{(n-1),sa} & \vdots \\ sd_{c,1} & sd_{c,2} & \cdots & sd_{c,sa} \end{bmatrix} $$

with $sa = 1, 2, \ldots, n$ NUTS 3 level areas; $c = 1, \ldots, 99$ level 2 digit NACE code

Using Equation 20 and Equation 22 the distribution of materials per sub area (MSD) is obtained through Equation 23.

Equation 23:
The spatial decomposition will complement the previously described variables and help solve the second, fourth and fifth gaps.

The global urban metabolism model described is summarized in Figure 1.
Table 3 - Domestic Material Consumption per Economic Activity and Municipality of the LMA, 2005 (tons) (Rosado et al., 2013)

| Economic Activity | A+B | C | D | E | F | G | H | I | J | K | L+M | N | O+P+Q | Municipal DMC (per capita) | Average DMC for the LMA (purchase power per capita) |
|-------------------|-----|---|---|---|---|---|---|---|---|---|---|----|---|--------|--------------------------|----------------------------------|
| Agriculture       | 868 | 3 | 109 | 421 | 28 | 19 | 92 | 47 | 47 | 57 | 126 | 57 | 66 | 268 | 35.4 |
| Fishing           | 97  | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0.7  |
| Mining and quarrying | 28  | 0 | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38   |
| Manufacturing     | 126 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0    |
| Electricity       | 0   | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0    |
| Agriculture       | 868 | 3 | 109 | 421 | 28 | 19 | 92 | 47 | 47 | 57 | 126 | 57 | 66 | 268 | 35.4 |
| Fishing           | 97  | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0.7  |
| Mining and quarrying | 28  | 0 | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38 | 0  | 91 | 38   |
| Manufacturing     | 126 | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0    |
| Electricity       | 0   | 0 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0    |

**LEGEND:** A- Agriculture, hunting, and forestry B- Fishing C- Mining and quarrying D- Manufacturing E- Electricity, gas and water supply F- Construction G- Wholesale and retail trade; repair of motor vehicles, motorcycles, and personal and household goods H- Hotels and restaurants I- Transport, storage, and communication J- Financial intermediation K- Real estate, renting and business activities L- Public administration and defense; compulsory social security M- Education N- Health and social work O- Other community, social, and personal service activities P- Activities of households Q- Extraterritorial organizations and bodies.

4. Third stage: Streamlined Urban Metabolism – establishing a common framework

The economic structure of an economy may be described by means of an Input-Output table (IO table, monetary units), a usually available tool in international organizations’ websites. This tool may be used to estimate the use of materials by economic activities in a country using a conversion mass/price. Usually, the description of data in International Trade Statistics provides a way of allocating material categories of imports of raw materials and intermediate goods to economic activities. Assuming that the domestic extraction categories are distributed among the same activities as imports, it is possible, through the IO table, to estimate the use of goods per economic activity.

The **Streamlined Urban Metabolism** method currently being developed and validated consists on estimating the metabolism of an urban area by making use of Input-Output tables (IO tables) published by OECD. These tables characterize the sales from each economic sector to the others, the consumption of households, the consumption of the government, the acquisition of buildings and machinery by companies, households and the government (gross fixed capital formation, GFCF) and the exports. These are the demand

---

3 From Pina et al., 2013
parameters of IO tables. The calculation of the input of materials to the country was based on domestic extraction and trade statistics (FAO, UN Comtrade among others).

In order to distribute the input of materials by the multiple economic sectors, the final consumption and the exports, each material and product entering the economy is first allocated to the economic sectors that process them. This is performed using correspondence tables linking commodities (expressed in SITC, EW-MFA, HS or CN nomenclatures) to economic activities (expressed in nomenclatures such as ISIC and NACE), and conversion tables for nomenclatures of materials and of economic activities. The economic activities considered in this work are consistent with the ISIC nomenclature, with the grouping of economic activities used being shown in Table 4.

---

5 International Standard Industrial Classification
6 Statistical Classification of Economic Activities in the European Community
The distribution of materials through the economy is then made based on the sales registered by each economic sector, with the materials originated from each sector being split by all the economic sectors, final consumption and exports according to the purchases made from that sector. These calculations enable the estimation of how materials are distributed within a country economy.

The analysis of the large variety of materials and products that enter an economy has been facilitated by the structured nomenclature of categories of materials named MATCAT (Rosado et al., 2013).

<table>
<thead>
<tr>
<th>Table 4 – Economic sectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic sectors</td>
</tr>
<tr>
<td>Agriculture and mining</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Biomass related products</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Chemicals and fuel products</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Construction products</td>
</tr>
<tr>
<td>Metallic products</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Machinery and equipment</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Utilities</td>
</tr>
<tr>
<td>Construction</td>
</tr>
<tr>
<td>Commerce and services</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
The MATCAT nomenclature establishes a correspondence between products listed in the CN and the materials that constitute them (Table 2). It considers 6 categories of materials (fossil fuels, metallic minerals, non-metallic minerals, biomass, chemical and others) and 28 subcategories enabling identifying the types of materials of which an economy is more dependent.

The key methodological step to scale down consists again on using the share of the country’s workers that exist in the urban area, per economic activity, as well as the share of population. The share of workers is used to estimate the amount of materials consumed by each economic sector, as well as the amount of materials and products produced by each economy sector that are for international export. The final consumption by households and by the government was estimated using the share of population.

This method determines that the material input to an urban economy is composed of the materials that enter the urban area to be transformed by its productive sectors and the materials that are locally consumed by the economic sectors (households, government and companies): material input = products for transformation + local consumption. For each sector, the products for transformation are described as the products that enter an economic sector and leave it to be consumed elsewhere. The local consumption includes the consumption by households, the consumption by the Government and the consumption of materials by companies (materials that enter the economic sectors and only leave it as waste). The international exports are estimated based on the international exports estimated at the national level. To ensure a proper balance, the domestic exports are calculated as the material input minus the local consumption and the international exports: domestic exports = material input – local consumption – international exports.

Therefore, when comparing to the EUROSTAT material flows indicators (EUROSTAT, 2001) the material input to an urban area, as calculated by this method, is equal to the Direct Material Input (Domestic Extraction + Imports), with the local consumption being equal to the Domestic Material Consumption (DMI – Exports).

This methodology does not calculate the crossing flows within an urban area. However, this is not an issue, since the total material input to an urban area is determined by the total final consumption and the exports, and therefore the materials that cross the urban area are also not accounted in the inputs, thus leading to a correct material balance.

A preliminary diagram of the urban metabolism of Lisbon Metropolitan Area using the streamlined method for the year 2000 is presented in Figure 4, encoding the mapping of resource consumption of each of the urban economic activities. The diagram quantifies the consumption of materials by the different economic sectors. The materials are represented on the left side of the diagram and are organized in 6 categories: biomass, chemicals, fossil fuels, metallic minerals and non-metallic minerals. The economic sectors (15) are represented on the right side of the diagram. The areas that bridge them are proportional to the magnitude of the material flows that annually are required to sustain the economic activities.
Figure 4 - Direct Material Input per Economic Activity, Lisbon Metropolitan Area, 2000.
5. Conclusions

This article provides a review of the developed and on-going stages to establish at IN+ a straightforward method to perform urban material flow accounting with a disaggregation of results that allows supporting decision-making.

Vital to these methodological advances is the set of databases developed at the centre characterizing more than thirteen thousand commodities in terms of their material composition, their average lifespan and the life-cycle phase of each product.

The main achievement of the results of this work consists in providing an exclusive understanding of how natural resource uses correlate with urban economic activities. It does also demonstrate the potential of urban metabolism to characterize urban typologies if the work is replicated in a sound set of metropolitan areas and, for each of them, to benchmark the relative resource intensity of the economic sectors, as a valuable tool to learn how different development options can be taken as a reference to be adopted in the urbanizing context.

References

The authors would like to acknowledge the Portuguese Science and Technology Foundation (FCT) for its financial support in the aim of projects ResiSt (Reference: PTDC/SEN-ENR/103044/2008) and MeSur (Reference: PTDC/SEN-ENR/111710/2009) (PTDC/SEN-ENR/121747/2010) as well as the Ph.D. scholarship of Dr. Leonardo Rosado (Reference: SFRH / BD / 30457 / 2006).

Asian Development Bank is also acknowledged for its support in the aim of the project Key Indicators for Asia and the Pacific 2011 (TA-7792 REG).
References


Chapter 3:
Fernanda Margarido, Carlos A. Nogueira, Fátima Pedrosa and Paula Oliveira, “Waste characterization and management: promoting the recovery of metals from end-of-life products”
Chapter 3

Waste Characterization and Management:
Promoting the Recovery of Metals from End-of-Life Products

Fernanda Margarido\textsuperscript{(1a)}, Carlos A. Nogueira\textsuperscript{(2)}, Fátima Pedrosa\textsuperscript{(2)}, Paula C. Oliveira\textsuperscript{(2)}

1 - Center for Innovation, Technology and Policy Research – IN+, Instituto Superior Técnico, Technical University of Lisbon (TULisbon); Av. Rovisco Pais; 1049-001 Lisboa, Portugal;

2 - LNEG – Laboratório Nacional de Energia e Geologia, I.P., Estrada do Paço do Lumiar; 1649-038 Lisboa, Portugal

\textsuperscript{(a) corresponding author: fernanda.margarido@ist.utl.pt}

Abstract

Today lifestyle demands an increased consumption of products with specific requirements. Technological innovation and market expansion have been contributing to a quick replacement of these products, increasing the amount of waste. The well-known examples of this type of products are electric and electronic equipment and portable batteries.

Electric and electronic equipment and also batteries are multicomponent products and with a very heterogeneous composition, containing several types of materials, being metals one of the most important constituents. Recycling is an important issue concerning the correct management policy of these residues. Its importance is due not only to the environmental problems related with the contained hazardous materials, but also to the economic reasons promoted by the recovery of the valuable metals.

In electronic appliances, printed circuit boards (PCB) represent the most important component, containing high grade of valuable metals, besides organic resins and some ceramic materials. Copper is the major metal in PBC’s composition (normally higher than 20\% \textit{w/w}) but many other secondary and minor metal elements, including precious metals can be found.

In the case of portable batteries, the most relevant electrochemical systems are Zn-MnO\textsubscript{2} (alkaline and saline), nickel-cadmium (Ni-Cd), nickel-metal hydride (Ni-MH) and lithium-ion (Li-
ion). Several important metals, usually with high grades, are contained in most of these battery types, making their recovery an economic and sometimes a strategic issue.

A precise and accurate physical/chemical characterization is the first step aiming at the development of adequate treatment processes for end-of-life products. The characterization of these two important waste streams, printed circuit boards and spent batteries, is here presented and discussed. The valorization processes are subsequently analyzed.

1. Introduction

Metals have been and are still crucial materials for the Human development. The rapid consumption of the last decades was responsible for the huge increase in metals demand. Mineral resources, namely the metal-containing minerals, were strongly explored and nowadays we are faced by an increasing and worrying depletion of resources. Europe has particular problems in this domain due to the shortage of metal primary resources, mainly concerning some rare metals. These metals, used in high-tech applications and in emerging technologies of important fields such as energy, environment, security and health, have been identified as critical, due to its scarcity and supply risk associated. Examples are cobalt, gallium, germanium, indium, platinum-group metals and rare earths. But, in fact, even the resources of base metals such as copper, zinc and nickel are being rapidly depleted and therefore require the same concern. Besides recycling allowing an effective materials management, also promotes energy efficiency from a life-cycle point of view, since it is well known that recovery of metals from wastes is normally less energy consuming than their recovery from primary mineral resources.

A correct management of metal-containing wastes and an efficient recycling can strongly contribute to the solution of this difficult equation of metals supply. Two important waste streams, for environmental and economic reasons, that contain some of the previous identified metals, are considered in this chapter: batteries and printed circuit boards of waste of electric and electronic equipment. Their characterization and processing are herein reviewed and discussed. A summary of the research work carried out recently by the authors in this field is also presented.

2.1 Battery Flows

Batteries are complex products with a large variety of sizes, shapes and electrochemical systems which causes serious difficulties for predicting and carrying out its valorization process.

Domestic-type batteries are essentially of the Zn-MnO$_2$ system and with cylindrical shape, varying from the large size Mono type to the small size Micro type [1]. Zinc-manganese batteries (saline and alkaline types, both primary cells) are the most frequently found in domestic flows. The other main systems are the secondary (rechargeable) batteries of the systems nickel-cadmium (Ni-Cd), nickel-metal hydride (Ni-MH) and lithium-ion (Li-ion).

Fig. 1 shows the weight distribution of spent batteries in Portugal. In the other European countries, figures are essentially similar. Primary batteries of Zn-MnO$_2$ system (alkaline and saline) are still the most relevant accounting to more than 70% of the market share, the alkaline being clearly predominant. The saline type is now only about 12% and is continuously decreasing. The actual trend is to increase the usage of secondary batteries, mainly the Li-ion systems, and to decrease the primary battery consumption. The Ni-Cd system is also declining.

![Fig. 1 - Market weight distribution of spent portable batteries by electrochemical system in Portugal (data from 2008 to 2012).](image-url)
Table I shows the typical elemental composition of spent batteries considering only the main metals for each system. All batteries are high metal concentrates, containing some valuable metals (e.g. Ni, Co, rare earths). Therefore, besides the environmental concern, economic benefits are expecting from treatment of these residues. The main solid phases constituting the electrode materials are also presented in the same Table.

Table I - Elemental composition ranges and electrode solid phases of spent batteries.

<table>
<thead>
<tr>
<th>Element</th>
<th>Elemental composition (wt%) of batteries</th>
<th>Main solid phases in electrodes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Zn-MnO₂</td>
<td>Ni-Cd</td>
</tr>
<tr>
<td>Zn</td>
<td>18-25</td>
<td>-</td>
</tr>
<tr>
<td>Mn</td>
<td>25-30</td>
<td>-</td>
</tr>
<tr>
<td>Fe</td>
<td>20-25</td>
<td>30-45</td>
</tr>
<tr>
<td>Ni</td>
<td>-</td>
<td>20-30</td>
</tr>
<tr>
<td>Cd</td>
<td>-</td>
<td>10-15</td>
</tr>
<tr>
<td>Co</td>
<td>-</td>
<td>0.3-1.0</td>
</tr>
<tr>
<td>RE</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cu</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Al</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Li</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

RE = rare earths, manly La, Ce, but also traces of Pr and Nd.

2.2 Characterization of Spent Batteries

Batteries are composed of electrodes, normally as pressed powders in the form of plates, stacks or rolls (Fig. 2). Nowadays these materials are manufactured with a highly porous structure and a large surface area to provide a low internal resistance and a capability for high performance in properties such as energy/power density. Some examples of morphology of battery powders are shown in Fig. 3. Nanoparticles are common to found in recent battery electrodes. Positive and negative electrodes are connected by an ion conductive media (the electrolyte). The active materials are supported in plates (made of steel or Ni plated steel) or simply compressed with a powder conductor (like graphite or nickel powder) to form a plaque. Besides electrodes and electrolytes, the external cans (casings) are a common component of all batteries. Iron in different steel alloys or in Ni/Sn plated steel is therefore always present.
Fig. 2 - Several arrangements used in electrodes manufacturing for portable cells.

a) compressed masses in cylindrical cells; (b) jellyroll electrodes in cylindrical cells;
(c) stack tablets in prismatic cells.

Fig. 3 - SEM micrographs showing electrode materials morphology of some samples of batteries: (a and b) Ni-MH electrode; (c) Alkaline electrode; (d) Ni-Cd electrode.
The distribution of the main components of batteries is presented in Fig. 4, being the specificity of composition of each battery type described as followed.

![Graph of battery component distribution](image)

**Fig. 4 - Distribution of main components of portable batteries (weight %).**

**Zn-MnO\(_2\) batteries**

Alkaline and saline (also called zinc-carbon) batteries are the most common portable cells of the Zn-MnO\(_2\) electrochemical system. They are typically the same type of cells, varying essentially in the electrolyte used (potassium hydroxide in the first one and zinc/ammonium chloride in the second one). Although having similar chemistry, the usual assembling of these two batteries is quite different. The anode is metallic zinc, transformed to zinc oxide during its lifetime, and the cathode is MnO\(_2\), partially transformed to manganese oxides with lower valence during discharge.

Concerning the mass balance of materials [2], the cathode represents about 50% while the anode accounts to near 22%. As can be seen, the electrodes are quite predominant in these cells. Iron components from casings contribute to about 22% and the remaining materials (insulators, electrode separator) are about 6%. Elemental composition shows high grade of Zn and Mn, typically from 20 to 30% each. In spent batteries, zincite (ZnO) is the most probable zinc species, which is readily soluble in acid media. Several manganese oxides are probably found and their reactivity is quite different depending on the oxidation state of manganese. When considering the aqueous dissolution of Mn(IV) oxide, besides the use of an acid the addition of a reductive agent is mandatory to achieve high Mn yields.
**Ni-Cd batteries**

The use of Ni-Cd batteries in domestic appliances have been drastically reduced due to environmental concerns regarding the cadmium metal. However, all the predictions of the overall reduction of cadmium usage in batteries were failed, and Ni-Cd cells have still usage in some applications, namely in large scale (called “industrial batteries”) and also in some cordless tools.

The chemistry of these batteries is based on a Ni(OH)$_2$ cathode and a Cd(OH)$_2$ anode, which are the species present in discharge conditions. When charged, Ni is oxidized to NiOOH and Cd reduced to metallic cadmium. The electrodes represent about 49% of mass balance, equally distributed by cathode and anode, steel is about 40% and the remaining 11% are plastics and other substances. Elemental chemical composition shows high metal grades, typically 20-30% Ni and 10-15% Cd [3]. Recycling of Ni-Cd batteries is driven by the environmental control of Cd flows and by the interest on Ni recovery.

**Ni-MH batteries**

This type of battery has been gradually substituting Ni-Cd cells in portable appliances. They are based on a similar cathode (NiOOH/Ni(OH)$_2$) but the anode is substantially different, constituted by metal alloys which form an hydride when charged. These metal alloys are based on AB$_5$ and AB$_2$ compositions. For the first type, A represents a rare earth and B a transition metal, usually based on the composition LaNi$_5$. For the second type, less usual, A is a transition metal with low atomic number and B transition metals with higher atomic number.

The electrode materials are the main portion of Ni-MH batteries, being about 42% of materials, where the cathodic material is predominant. The steel components constituting the metallic casings and supporting grids contribute to about 23%, essentially made by Ni-plated steel. Plastics, insulators and other materials represent 8% [4].

Regarding chemical composition, a portable Ni-MH battery can contain 22-32% of Ni, which gives economic interest to these waste materials. The cobalt content cannot be ignored due to its commercial value. The total rare earths grade (until 5-7%) is also an interesting value that should be taken into account.

Concerning phase composition, there are essentially two types of solid phases in electrode materials: hydroxides, mainly Ni(OH)$_2$, and the metal alloys, e.g La/CeNi$_5$. The reactivity of these two main forms is different and shall be taken into account in any process development.
Li-ion batteries

The Li-ion system is the most recent and promising of all battery types. These cells take advantage owing to the high electrochemical potential of lithium, together with its low atomic weight, to allow achieving a substantially higher discharge potential when compared to the other known systems. The most usual cathode of portable cells is composed by a lithium-cobalt oxide, LiCoO$_2$. In large scale application, other cathodes are more common for economic and safety reasons, namely the lithium-iron phosphate. The anodes are essentially made of an intercalated structure of graphite-lithium. During the charge-discharge cycles, the Li ions migrate between the two electrodes. Due to the high reactivity of these materials, the electrolytes are non-aqueous, normally a lithium salt in an organic solvent. For supporting the electrodes, two metal sheets are normally used, an Al foil in the cathode and a Cu foil in the anode.

As observed for the other batteries, the electrodes represent the major proportion of the weight of the Li-ion cells, being about 50%. The steel cases contribute to near 29% while plastics are about 8%. The remaining weight is essentially distributed by the electrode supporting foils (5% for Al and 8% for Cu).

The main element of a Li-ion portable cell is cobalt (about 15-20%) which gives a special economic and strategic interest to recycle these wastes. Lithium content is substantially lower (typically 2-5%) and economically much less attractive. However, when considering the expected boost in Li demand for electric vehicle batteries, lithium will be rapidly considered a strategic metal. Then its recovery from spent batteries shall be at once a major concern. Regarding metals recovery from spent batteries, the reactivity of electrodes is relatively favorable. For example, using acid media in relatively low concentrations will allow substantial dissolution of Li and Co from electrode materials.

3. Processing of Spent Batteries

The chemical treatment of batteries can be performed by pyrometallurgy or hydrometallurgy [5,6].

Pyrometallurgical processes applied to batteries are carried out in smelters which process other materials (primary/secondary resources), the batteries being normally a smaller portion of the furnace charge. So, for some types of batteries, the use of facilities already existing can
be an advantage. However, sometimes this is not possible (such as when there is mercury in the battery composition) but in some cases a previous pre-treatment can be eventually necessary. The processes involve usually a smelting operation producing a metal stream for valorization and a slag to discharge.

The hydrometallurgical processes for spent batteries are not so common. They involve a mechanical/physical treatment (shredding, separation of ferrous materials and non-metallic fractions), followed by leaching, usually in acid media, and a series of hydrometallurgical separation operations (precipitation, solvent extraction, electrowinning) to produce the valorized metals or metal compounds. The iron from case components is always a problem in aqueous processing since the Fe dissolved will require the introduction of purification steps.

A more detailed discussion of processes for some battery types will be further presented.

**Processes for Zn-MnO₂ batteries**

Several high temperature furnace processes were developed and applied to spent Zn-MnO₂ batteries. These processes operate in dedicated plants [7,8]. An initial step of thermal treatment is normally used where the pyrolysis of plastics occur, as well as the distillation of mercury, if it is present. The pyrometallurgical process proceeds with zinc volatilization and condensation as metal or oxide, to be reused. The iron and manganese can be recovered as an alloy or discharged as slag.

Industrial applications of hydrometallurgical processes are also known [9,10]. The batteries are mechanically treated and fragments of different materials are separated by means of physical operations. The electrodes constituting the “black mass” are therefore leached in sulfuric acid media, and zinc and manganese are recovered applying sequential and selective precipitation steps. Solvent extraction for Zn/Mn separation was also proposed and applied for hydrometallurgical treatment of these batteries.

**Processes for Ni-Cd batteries**

The Ni-Cd batteries were the first cell types to be recycled in dedicated plants, some decades ago. The processes were operated by some battery manufacturers allowing recovery and reuse of recycled cadmium for producing new batteries [11,12]. The process scheme is more or less similar to that of Zn-MnO₂ batteries, involving a pyrolysis step followed by cadmium distillation (volatilization and condensation) and its recovery on a pure form. Nickel can be slagged or else alloyed with iron to be potentially utilized. No evidence of cobalt recovery is addressed by
these processes, in spite of the content of this metal in these cells. Hydrometallurgical processes were also developed but not applied to this type of batteries.

**Processes for Ni-MH batteries**

Pyrometallurgical processing of Ni-MH batteries involves smelting for nickel recovery as a metal phase. Other elements such as the rare earths are lost in the slag. So, these processes are relatively simple and objective, but the loose of important elements cannot be disregarded. The alternative hydrometallurgical approach can give a more efficient recovery of metals, but requires more steps. After mechanical processing, the metals are leached and the rare earths can be precipitated as sulfate double salt or extracted by adequate organic extractants and subsequently stripped and recovered. Nickel remaining in solution can also be further recovered by another precipitation operation.

**Processes for Li-ion batteries**

Processing of Li-ion batteries can be done by a cryogenic process (together with primary Li cells) aiming at Li recovery, by pyrometallurgy or hydrometallurgy. The pyrometallurgical processes are usually related with Co recovery and can be carried out in integrated smelters where other secondary resources are treated. Cobalt is melted and purified as pure metal while lithium is lost in the slag. The more recent hydrometallurgical approaches allow recovery of cobalt and lithium by means of mechanical processing/deactivation, leaching and purification/precipitation operations [13].

4. Development of New Processes

Several processes for recycling batteries were developed by the present authors in the last years. They are usually of hydrometallurgical type, involving physical processing steps, leaching and operations for separation, purification and metals recovery. The target was always to reach the maximum metals recovery, the minimum reagents consumption, and the maximum energy efficiency. Special attention was made for minimizing losses, mainly concerning the metals that are commonly lost in other process approaches, like lithium and rare earths, as described in the previous sections. Three recycling examples developed by this group for Ni-Cd, Zn-MnO₂ and a mixture of batteries are here presented.
Recycling of Ni-Cd batteries by hydrometallurgy can be carried out according with the diagram of Fig. 5 [14]. The process is based in physical processing, acid leaching with sulphuric acid solutions and aqueous treatment of leachates by two successive solvent extraction circuits were cadmium, cobalt and nickel are separated. The physical and chemical (leaching) operations were integrated to allow the maximum recovery of the electrode materials and the maximum metal leaching yields. The batteries are shredded and the supra fraction containing the ferrous scrap is separated and washed. Since some of the electrodes are stacked in the supporting grids, the washing is made with the pregnant leach liquor in a mechanically stirred vessel, in order to assure simultaneously the leaching of some electrode material and disaggregation of remaining particles from the grids. This arrangement also allows to attain a partial consumption of the excess of acid in the leachate. The particles in the pulp are reclaimed to the main leaching reactor while the rich leach liquor is neutralised to remove the iron dissolved and then sent to the separation steps. In the leaching reactor, the dissolution of the main metals, cadmium, cobalt and nickel occur. Special attention is required to the dissolution of the metallic Ni mesh contained in the particles, which requires higher acid concentration and temperature [15,16]. By the contrary the metal hydroxides are quite soluble even at moderate conditions.

The separation of cadmium, cobalt and nickel is performed by solvent extraction using organophosphorous extractants [17]. In the first circuit, cadmium is selectively extracted and the second circuit cobalt is extracted against nickel. In both circuits, final concentrated strip liquors are produced, close to saturation, allowing easily the crystallization of metal salts with minimum energy consumption. Nickel can be recovered as sulphate or carbonate. Depending on the Ni recovery form chosen, the solution can be recycled back to the leaching step to promote water and reagent savings.
For the Zn-MnO\textsubscript{2} system (alkaline and saline batteries) the process developed [18] is quite similar regarding the unit operations involved (Fig. 6). After shredding the batteries, the coarse fraction containing the steel and plastic scrap can be sieved and removed. The fines are sent to the leaching operation, in sulphuric acid media [19]. An alternative study involved the leaching of the electrodes together with the iron scrap, since Fe in the metal form can act as reductive agent promoting the reaction and dissolution of the Mn(IV) species, otherwise very insoluble. The zinc species were always very reactive and the reaction can proceed with a minimum acid consumption. After purification, the leach solution is treated by solvent extraction, where zinc is selectively extracted against manganese. The solvent extraction circuit can be design to be connected with an electrowinning recovery step, or alternatively to produce a zinc salt. The raffinate containing the manganese allows the production of any Mn salt. This feature provides versatility to the process, which can be easily adapted according with the market needs.
From the previous work described, it seems that the several types of battery flows can be treated by a similar way (physical processing, leaching and separation), providing that the separation steps, using hydrometallurgical operations, are adequately designed. Therefore, a recent development involves the proposal of a combined process valid for the treatment of mixed battery flows (BATMIX). The diagram developed, summarized in Fig. 7, describes this innovative proposal.

Zn-MnO₂ batteries represent about 70-80% of the market share and therefore any process dealing with simultaneous treatment of all the portable cells will become technically unacceptable since valuable metals like cobalt, nickel, rare earths and lithium, will be strongly diluted, affecting negatively its recovery. Hence two main separated treatment flows are proposed, one for Zn-MnO₂ batteries, using the same technology previous described for these cells, and the other for secondary cells (Ni-Cd, Ni-MH and Li-ion), rich in the metals Ni and Co [20]. In this later flow, after the physical processing, the rich electrode fraction is leached with H₂SO₄ solution and, after purification, metals are separated in three solvent extraction circuits (the first for rare earths, the second for Cd-Mn and the third for cobalt) leaving nickel and
lithium in the final raffinate for further processing. Expecting recoveries are higher than 90%. The process still requires improvements concerning impurity control and demonstration/validation in higher scale.

Fig. 7 - Flowsheet for recycling mixtures of portable batteries.


The industry of electric and electronic equipment (EEE) is one of the sectors of manufacturing that has the highest growth in the Western World. Rapid technological innovation has allowed the access to better and cheaper electronic products, but the other hand leads to the products
become obsolete more quickly shortening its life cycle. As result, the generation of waste of
electric and electronic equipment (WEEE) has increased sharply in recent years.

EEE are distributed into ten classes, depending on their size and functionality, including: (1) large household appliances (being the cooling line a special group of these, often considered separately); (2) small household appliances; (3) computer equipment and telecommunications (including computers, monitors, printers/copiers and telephones); (4) consumer equipment, including audio, television and video, as well as photovoltaic equipment; (5) lighting equipment; (6) power tools; (7) toys, electric/electronic games; (8) medical equipment; (9) monitoring instruments, and (10) automatic dispensers. Since a large number of these devices are for domestic use, it has been increasing the proportion of this flow of waste in municipal solid waste (MSW), constituting recently in Europe, about 4% by weight of the total MSW with tendency to increase.

EEE is constituted by a set of materials with complex and varied physico-chemical composition, being some of its constituents recognized as hazardous [21]. Each equipment consists of several parts or modules, highlighting the wrappers or outer boxes, usually plastic or metal, and inner constituents. In most equipment, the internal parts are a set of electric or electronic systems containing printed circuit boards, cables or wires, capacitors, resistors, relays, switches and circuit breakers, sensors, connectors, batteries and display screens (cathode ray tube or liquid crystal display). The chemical composition of these constituents is very diverse, but some substances must be standing out for their potential hazardous after their end-of-life, namely:

- the heavy metals, very toxic, such as mercury (present in switches and circuit breakers), lead (present in the glass of cathode ray tubes and some welds), cadmium (present in some types of batteries and in the composition of some plastics), chromium and arsenic;

- the other heavy metals such as nickel and copper, constituting the electrical cables and circuit boards;

- the halogenated substances, including flame retardants such as polybrominated diphenyl ethers and polybrominated biphenyls - PBDE and PBB (used in some flame retardant plastics), PVC, chlorofluorocarbons and hydro-chlorofluorocarbons - CFCs and HCFCs (used in cooling systems and insulating materials) and polychlorinated biphenyls - PCBs (used in capacitors). These compounds may produce dioxins and furans when incinerated;

- the fibrous materials.
Although the use of many of these toxic substances has already been prohibited by the European legislation, there are still huge amount of equipments in use and WEEE containing these potential hazardous substances.

The estimated weight distribution of the main materials constituents of WEEE is represented on Fig. 8, where metals represent almost 50%, followed by plastics and glasses (particularly those of the cathode ray tube containing high levels of lead).

![Fig. 8 - Average composition of main type of materials (weight %) on WEEE](image)

In terms of metals (Fig. 9a), stands out copper, which is the main component of devices containing electrical circuits, but there are also other important metals such as iron, zinc, nickel, tin, aluminum, and lead. However, many other metals are part of the constitution of WEEE, some of them with toxic characteristics and others with high economic value.

Regarding the polymeric materials [22], there is a wide range in electronic equipment (Fig. 9b). It can be mentioned the polystyrene (PS), acrylonitrile-butadiene-styrene (ABS), polypropylene (PP), polyurethane resins (PU), and polycarbonate (PC), all of them from the class of thermoplastics. But other plastics belonging to the thermoset class like resins such as phenol formaldehyde (PF) and the epoxy, can also be found.
In Europe it is estimated that, on average, are generated about 20 kg of WEEE per inhabitant per year. This quantitative depends on the countries concerned, considering that in Western Europe varies from 14 to 24 kg, being in Eastern European countries 6-12 kg. The total WEEE generated annually in Europe (EU27) was then estimated at about 9 Mt, being on the EU15 about 7-8 Mt. In Portugal, it is estimated that the quantitative is about 130-140 kt annually.

5.2 Characterization of Printed Circuit Boards

Printed circuit boards (PCB) are present practically in all electric and electronic devices, especially computers and other related equipment. Bearing in mind that computers are a very important item of WEEE, it can be concluded the relevance of PCB within the WEEE.

The PCB (also called PWB - "printed wiring board") consist of two main parts:

- the plate itself, constituted by the conductor circuit of copper printed on (or in) a matrix (substrate) of insulating plastic material (resin);

- the components mounted on the plate, including chips, resistors, capacitors, switches/protectors (relays, fuses), magnetic devices (inductors, transformers, amplifiers, solenoids), transistors, transducers, diodes (bridges, LED's), terminals and connectors.

The plate (not assembled) represents about 74% by weight of the PCB while the remaining 26% is attributed to the components [23].
Sometimes the designation of PCB refers only to the circuit without the components. After assembling the components, the final board may be called by printed circuit assembly (PCA) or mounted printed circuit board assembly (PCBA). In this chapter, we will use the simplified term PCB to denote the already assembled plate, except where expressly provided otherwise.

According to the European legislation, the PCB is one of the components that must be selectively removed from WEEE by operators during the dismantling phase, due to the economic value of some materials present but also to the potentially toxic nature of some of its constituents.

The quantitative composition of PCB is reported in the literature, showing the existence of a very wide range of elements and substances [24].

The key elements are the metals copper, iron and aluminum, plastic materials (epoxy resin included) and ceramics. Plastics represent about 40% by weight of the plates, followed by ceramics (about 20%), copper (also about 20%), iron and aluminum (both slightly less than 10%). The dispersion of the composition is very high depending on the type of equipment, demonstrating the high heterogeneity in this type of components.

As secondary elements stand out lead, tin and zinc (on average 1-2 %), the halogens Cl and Br (about 3 %) and nickel (about 1 %). The dispersion of these elements in various equipments analyzed is also very high, being of the order of magnitude of their average values.

The analysis of some data collected in the literature should be done with caution. Regarding the composition of most metals, this does not seem problematic, but some mistakes can occur with other materials such as ceramics and plastics. Concerning to ceramics, these may include silica, alumina, titanates (barium, for example), among others. Thus, it is sometimes difficult distinguish, if aluminum, silicon, barium, titanium, presented in the elemental analysis is in the form of oxides (ceramics). With regard to plastics, its weight composition is often determined by the physical separation and weighing. Thus, certain components that may be included in the plastic compounds and that are not organic, like fiberglass, are also recorded on this weight. Other examples, although less important are antimony which is used as a flame retardant or sometimes as a silica filler composite.

Based on the information gathered in the literature, a compiled average composition of PCB is shown in Fig. 10. This does not intend to represent the composition of all plates, due to its high variability (shown previously) but can provide a good baseline for these devices. The metals account for about 40% in this distribution, and the nonmetallic fraction of approximately 60%.
The metal fraction is the relevant part of the PCB from an economic point of view, given the high metal content, some of them of great value. Thus, these residues may be regarded as metal concentrates from which recovery is strongly recommended.

![Pie chart showing the composition of PCB](image)

**Fig. 10** – General average composition of printed circuit boards, considering the main constituents (individually or in groups).

### 6. Processing of Printed Circuit Boards

The most effective way of management of the PCB wastes is recycling, which allows the recovery of contained materials and avoids the problems and costs associated with landfilling.

The recycling of PCB involves the following steps: 1) Disassembly and Sorting; 2) Physical Processing and (3) Metallurgical Processing.

The diagram in Fig. 11 illustrates the sequence of main operations and steps of recycling of PCB.
6.1 Dismantling and Physical Processing

After collection and dismantling of equipment containing PCB, which is usually performed manually, these must be removed for individual processing. The verification of the state of PCB, if they are still in operating conditions, is a priority option, which enabling reuse in secondary markets and thus increases the lifetime of these devices.
Due to technic and economic problems of removal of components of PCB the most common procedures involve the direct processing without the dismantling phase. The first step is the physical processing, which involves the fragmentation and subsequent separation of the constituent materials based on the difference in physic-chemical properties, in particular the magnetic or electrostatic properties, particle size and density [25].

The aim of fragmentation is normally the individualization ("liberation") of different constituent materials in particles of single materials for subsequent separation by physical processes. To achieve efficient release of materials, the dimension of particles after fragmentation would be less than 1 mm (typically 0.5 mm or less, depending on the subsequent separation technology).

Considering that PCB are constituted by two major groups of materials, plastics and metals, the characteristics used for separation are magnetic properties, giving the separation of the ferrous metals (ferromagnetic) of the remaining material, and subsequently the separation of nonferrous metals (copper, aluminum) diamagnetic based on eddy current separators. For separating the polymeric materials it is used the electrostatic separation, which are subject to electrification having metals a passive role. Other alternative physical processes are based on density differences between plastics and metals, including the flotation (with foam, air induced or dissolved air), the elutriation, gravity separation with dense media and hydrocyclone [25].

The physical process is often considered crucial in the overall recycling process because it can facilitate the subsequent chemical treatment. An efficient separation of materials makes the resulting fractions as "pure raw material", suitable for further processing, being technically and economically more favorable.

6.2 Metallurgical Processing

The metallurgical processing can be pyrometallurgical or hydrometallurgical. The first is based on the reactions of metals at high temperatures and the impurities slagged, obtaining a melt which is purified to obtain the pure metal. The second is based on the solubilization of metals in suitable aqueous medium, proceeding then to the treatment of the resulting solutions to obtain the metals in recoverable forms.

The material fed to a metallurgical process can be processed directly without any prior physical separation. In some pyrometallurgical processes, the plastic fraction of PCB can be used as a
power source for furnaces, thus carrying a waste energy recovery. The organic matter also plays the role of reducer agent, which is critical in melting furnaces, thereby decreasing the use of other carbon materials. However, there is the risk of hazardous gaseous emissions, but usually the operators ensure the control of emissions.

There are several "smelters" exclusively dedicated to the processing of scrap (copper metal and its alloys) and other residues (dust, ash, sludge, slag). Many "smelters" accept electronic scrap, but usually do not pay the value inherent to this raw material based on content of all contained metals.

The pyrometallurgical processing is based on the fusion of copper in water jacket furnaces [26,27]. May also be involved steps relating to the metallurgy of lead [27], as found in the UMICORE integrated smelter. Valuable metals, can be considered in two groups of elements with different behaviors for the molten copper: those that are soluble in the melted copper such as Au and platinum group elements - PGM (Pt, Pd, Rh, Ir, Ru) - and others forming oxides which go to the slag (such as Fe, Pb, As, Sb, Sn, Bi, In). Some elements such as silver will be distributed by both metal phases, but with greater affinity for the slag.

A typical pyrometallurgical copper process allows obtaining secondary recovery yields of 99% Cu, 90% Ag, 98% Au, 90% Pt, 90% Pd, 85% Ni, 85% Pb, 80% Sn and 85% Zn [26].

Despite the advantages of pyrometallurgical processing, particularly in terms of energy recovery (not yet fully optimized) contained in the plastics of PCB, there are also disadvantages in terms of the possible generation of hazardous gaseous emissions (dioxins, etc.), excess of production of slag due to some metals (Fe, Al) and ceramics, and some loss of valuable metals in outputs of the processes which are not optimized.

The alternative route for recovery the base metals of PCB is the hydrometallurgical processing [28,29], but its industrial application does not exist or, if any, will be scarce and for small processes. The main steps of hydrometallurgical processing are metal leaching, purification of the obtained liquor, separation technologies such as solvent extraction, selective precipitation and ion exchange, and finally the separated elements are recovered by electrolysis (as metal), precipitation or crystallization (as compound).
7. Development of New Processes

A recycling process for PCB’s was developed by the authors, based on physical processing and hydrometallurgical routes, aiming at recovering the contained copper, which is the higher grade metal present in this residue.

The treatment process consisted of applying physical and chemical technologies, including several steps: shredding, copper leaching in nitric media, copper solvent extraction, separation and purification and copper recovering by crystallization.

PCB’s waste was shredded in two steps (primary and secondary, respectively in a grab shredder and in a cutting mill) resulting a grinded material (Fig. 12) with 90% (weight) with particle size less than 2.1 mm and an average particle diameter of 1.2 mm. Chemical analysis of granulometric fractions showed that the base metals like Cu, Zn, Pb and Sn concentrated mainly in intermediate size fractions (0.4-1.7 mm) being the fines very rich in epoxy resin composite. About 80-90% of the principal metals were recovered in that size range [30].

Fig. 12 – Images of shredded PCB: (a) fragments after primary shredding; (b) detail of particles after secondary shredding.

The first step of chemical treatment was the acid leaching of metals. Due to the high complexity of PCB’s, which contain numerous chemical elements in their composition, selective leaching using different agents is a possible approach to be considered, since can allow to extract selectively some groups of metals in successive leaching stages, improving and
simplifying the following separation steps. In this sense, the solubility of metal ions in the various media (e.g. sulfate, nitrate, chloride) is the starting point for the development of the selective leaching process. As example, silver is practically insoluble in chloride media while lead is very insoluble in sulfate media. Noble metals such as Pd, Pt and Au would be only soluble under extreme conditions such as in aqua regia.

Several leachants were compared (sulfuric, hydrochloric and nitric acids) aiming at maximizing copper recovering. Sulfuric acid leaching was not very promising concerning metals dissolution being only effective for iron. Hydrochloric acid allowed the leaching of tin up to 60% and about 50% of lead, as well as the iron. Nitric acid was the most efficient leachant due to its oxidizing properties, allowing efficient copper dissolution. The use of the other acids is only optional providing that an additional oxidant agent is used.

Nitric acid was chosen to proceed with the process development. At appropriate conditions, the efficient solubilization of base metals like Cu, Ni, Zn, Pb and Ag was allowed [31]. More than 90% recovery of Cu, Zn and Ni were achieved using relatively low HNO₃ concentration. The obtaining leach liquors contained between 8.0 to 10 g/L of copper. Precious metals did not react under these conditions and their recovery would be attained in a subsequent leaching step using highly concentrated acidic/oxidizing solutions.

The pregnant copper liquor was processed by solvent extraction using Acorga M5640, a hydroxoxime extractant, allowing to selectively extract the copper and to recover it in a crystallization circuit. Extraction efficiency was 98.9% and a final copper sulphate monohydrated was produced.

The developed process, exemplified in Fig. 13, allowed recovering copper from PCB by using relatively simple and versatile technology with high efficiency (above 90%) and without the environmental disadvantages and the higher energetic consumption which are associated with the classic pyrometallurgical copper processes.
Metals recovery from end-of-life products through recycling is nowadays mandatory, due to environment, economic, as well as strategic reasons related with scarcity and depletion of mineral raw materials. Recycling contributes to decreasing the supply risk of metals, provides environment protection and energy efficiency when compared with primary extraction.

A complete and accurate characterization of the secondary sources is the first step to allow the development of recycling schemes and to define suitable technological approaches. Physical processing plays an important role in materials separation, improving the efficiency of forthcoming metallurgical operations. Innovative hydro and pyrometallurgical technologies, designed to achieve the maximum recovery yields of metals, minimizing materials and energy consumption, are the tools needed to allow a more sustainable metals production industry, so important for the supplying of materials for today and tomorrow needs.


17. C.A. Nogueira, F. Delmas, New flowsheet for the recovery of cadmium, cobalt and nickel from spent Ni-Cd batteries by solvent extraction, *Hydrometallurgy*, 52, 1999, 267-287


Chapter 4:
Abel Rodrigues, Gabriel Pita, João Mateus, Cathy Kurz-Besson, Miguel Casquilho, Sofia Cerasoli, Alberto Gomes and João Pereira, “Continuous carbon fluxes measurements in eucalypts: drought and felling”
Eight years of continuous carbon fluxes measurements in a Portuguese eucalypt stand under two main events: Drought and felling

Abel Rodrigues\textsuperscript{a}, Gabriel Pita\textsuperscript{b}, João Mateus\textsuperscript{b}, Cathy Kurz-Besson\textsuperscript{c}, Miguel Casquilho\textsuperscript{d,\,*}, Sofia Cerasoli\textsuperscript{e}, Alberto Gomes\textsuperscript{a}, João Pereira\textsuperscript{e}

\textsuperscript{a} Unidade de Silvicultura e Produtos Florestais (Unit of Silviculture and Forest Products), INRB (IP), Ave. da República, Quinta do Marquês, 2780-159 Oeiras, Portugal
\textsuperscript{b} Instituto Superior Técnico (Technical University of Lisbon), Ave. Rovisco Pais, IST, 1049-001 Lisboa, Portugal
\textsuperscript{c} Instituto Dom Luís (CGUL), Faculdade de Ciências (University of Lisbon), Edifício C8, Campo Grande, 1749-016 Lisboa, Portugal
\textsuperscript{d} Centro de Processos Químicos, Instituto Superior Técnico (Technical University of Lisbon), Ave. Rovisco Pais, IST, 1049-001 Lisboa, Portugal
\textsuperscript{e} Instituto Superior de Agronomia (Technical University of Lisbon), Tapada da Ajuda, 1349-017 Lisboa, Portugal

\textbf{A R T I C L E   I N F O}

Article history:
Received 27 July 2010
Received in revised form 15 December 2010
Accepted 16 December 2010

Keywords:
Eddy covariance
Carbon
Tree felling
Water stress
Drought
GEEs

\textbf{A B S T R A C T}

This paper reports on results from eddy covariance measurements of carbon uptake and evapotranspiration in the eucalypt site of Espirra in Southern Portugal (38°38′N, 8°36′W). This site was included in the “Carboeurope” European network and is part of a 300 ha eucalypt forest, with about 1100 trees ha\textsuperscript{−1}, intensively managed as a coppice for pulp production and characterized by a 12-month annual growing period. The climate is of Mediterranean type with a long term (1961–1990) annual average precipitation of 705 mm and an annual average air temperature of 15.90 °C. During the measurement period (2002–2009) two main events took place, which changed the annual sink pattern of the forest: a drought period of two years (2004–2005) and a tree felling (October and November 2006). We analyzed the daily, seasonal and inter-annual variation of carbon uptake and evapotranspiration, and their relationships with the events and the variability of the main meteorological variables. Before the felling, annual net ecosystem exchange (NEE) increased from −865.56 g C m\textsuperscript{−2} in 2002 to −356.64 g C m\textsuperscript{−2} in 2005 together with a deep decrease in rainfall from 748 mm in 2002 to 378.58 mm and 396.64 mm in 2004 and 2005, respectively. For the same period, seasonal patterns of carbon uptake showed maximum values in April and decreased in July–August. The eucalypt stand recovered its carbon sink ability since June 2007 and had a NEE of −209.01 g C m\textsuperscript{−2} in 2009. After the felling, the carbon uptake occurred from mid-February to mid-October, following an almost opposite pattern than that of the trees in the term of their productive cycle. A quantitative approach using generalized estimating equations (GEEs) was made for the period before the felling to relate monthly NEE and GPP with accumulated photosynthetic active radiation, water vapour pressure and precipitation. In conclusion, our study showed the relevant effects of water stress and anthropogenic interventions in the daily, seasonal and annual patterns of carbon uptake, under a context of good environmental conditions for carbon sequestration.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

In the present context of climate change, studies on the role of forest stands in carbon sequestration have been reported since the 1990s. These studies, relying on direct atmospheric carbon flux measurements by the eddy covariance method (Aubinet et al., 2000; Baldocchi, 2003), were supported by research programs such as Carboeurope, or global science networks such as Fluxnet. The results provided by these researches allowed an improvement of knowledge of factors explaining the seasonal and inter-annual variability of the carbon balance components. Relevant knowledge was acquired about the variability of carbon uptake with latitude and season of the year (Falge et al., 2002). Carbon uptake (NEE) and assimilation (GPP) in forest ecosystems is associated to factors related to plant biology and physical environment such as: temporal variation of meteorological conditions, leaf area index (LAI), physiological activity, length of growing season, and soil temperature and moisture content. These factors affect the carbon balance components differently (Schmid et al., 2000). While gross primary productivity (GPP) is mainly dependent on intercepted solar radiation (a function of the photon flux of photosynthetic active radiation, PAR, and LAI), total ecosystem respiration (TER) responds mostly to air and soil temperature (Carrara et al., 2004; Baldocchi, 1997; Reichstein et al., 2002). The contribution of total ecosystem respiration in European forest stands to annual NEE differences increases with latitude (Valentini et al., 2000).
Table 1: Annual and three-month sums of NEE, GPP and TER. Also shown are mean air temperature $T_{a}$, cumulative global radiation, $R_g$, precipitation, Prec (mm), and evapotranspiration ($E$ (mm)).

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean $T_{a}$ (°C)</th>
<th>$R_g$ (MJ m$^{-2}$)</th>
<th>Prec (mm)</th>
<th>NEE (gC m$^{-2}$)</th>
<th>GPP (gC m$^{-2}$)</th>
<th>TER (gC m$^{-2}$)</th>
<th>$E$ (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>15.30</td>
<td>6007.81</td>
<td>748.29</td>
<td>-865.56</td>
<td>2206.04</td>
<td>1340.47</td>
<td>474.87</td>
</tr>
<tr>
<td>2003</td>
<td>16.06</td>
<td>6021.75</td>
<td>706.58</td>
<td>-791.33</td>
<td>1995.35</td>
<td>1204.02</td>
<td>590.20</td>
</tr>
<tr>
<td>2004</td>
<td>16.15</td>
<td>6225.97</td>
<td>378.58</td>
<td>-724.24</td>
<td>1834.88</td>
<td>1110.64</td>
<td>722.55</td>
</tr>
<tr>
<td>2005</td>
<td>16.01</td>
<td>6377.06</td>
<td>396.64</td>
<td>-356.64</td>
<td>1253.11</td>
<td>899.41</td>
<td>391.64</td>
</tr>
<tr>
<td>2006</td>
<td>16.51</td>
<td>6093.91</td>
<td>805.92</td>
<td>-619.07</td>
<td>1816.56</td>
<td>1197.91</td>
<td>756.10</td>
</tr>
<tr>
<td>2007</td>
<td>15.95</td>
<td>6572.38</td>
<td>443.05</td>
<td>-11.06</td>
<td>939.44</td>
<td>928.07</td>
<td>654.06</td>
</tr>
<tr>
<td>2008</td>
<td>15.86</td>
<td>6064.92</td>
<td>508.81</td>
<td>-200.73</td>
<td>1226.00</td>
<td>1024.97</td>
<td>726.73</td>
</tr>
<tr>
<td>2009</td>
<td>15.78</td>
<td>6278.04</td>
<td>569.53</td>
<td>-209.01</td>
<td>1294.05</td>
<td>1086.43</td>
<td>533.37</td>
</tr>
<tr>
<td>Mean</td>
<td>16.17</td>
<td>6175.23</td>
<td>569.68</td>
<td>-472.205</td>
<td>1570.94</td>
<td>1098.99</td>
<td>606.23</td>
</tr>
</tbody>
</table>

One of the features of the present climate change is an increase in weather variability. Since the 1970s, the frequency and severity of droughts increased in the Western Mediterranean region due to higher air temperatures and diminished winter–spring precipitation (Miranda et al., 2002). As water shortages generally decrease both GPP and carbon uptake in forests (Ciais et al., 2005; Granier et al., 2007; Pereira et al., 2010), drought has a strong relevance to determine the inter-annual and seasonal variation in ecosystem carbon exchange with the atmosphere. Indeed, water stress in these regions is a major factor controlling plant carbon uptake, due to stomatal limitation of photosynthesis (Farquhar and Sharkey, 1982) and atmospheric evaporative demand. Discussion about water stress influence on leaf and canopy gas exchange is provided in multiple references, e.g. by Tenhunen et al. (1983) and Pereira et al. (1986), in Portuguese Quercus coccifera and Eucalyptus globulus stands by Lebaube et al. (2000) and Granier et al. (2008), in a French beech forest or McCaughey et al. (2006), in mixed wood forest in Canada. Other natural or anthropogenic disturbances, such as defoliation or tree felling, also contribute to determine the amounts and the time patterns of carbon fluxes by forests (Jiang et al., 2002; Xiao et al., 2003).

This paper describes the evolution of NEE, GPP, TER and evapotranspiration ($E$) regime in an eucalypt stand at Herdade da Espirra, at Pegões, southern Portugal, included in the Carboeurope consortium, during the period 2002–2009. Eucalypt (Eucalyptus globulus Labill.) forest plantations intensively managed for pulp production are highly productive with 16 m$^3$ ha$^{-1}$ year$^{-1}$ of roundwood in a 12-month growing period and cover about 19% of Portuguese forest area (647 ha). Worldwide, intensively managed plantations, providing biomass for energy and industry, correspond to about 17 (25 million ha) of plantation forests (FAO, 2010), and their expansion launched a debate about their environmental impact (Canadell and Raupach, 2008; Markewitz, 2006; Paquette and Messier, 2010; Rotenberg and Yakir, 2010).

The consideration of forest management used and the weather pattern allowed us to consider two interesting events which affected the experimental site: a two-year drought in 2004 and 2005 and a tree felling in October 2006. After the felling young
sprouts underwent a thinning of three shoots out of four per stump in October and November 2008 and suffered as well leaf frost damage in winter 2007, 2008 and. In the aforementioned context, the main objectives were: (i) analysis of the impact of drought and felling in temporal patterns of carbon uptake, energy partition in the ecosystem and decoupling coefficient, (ii) the derivation of relationships between NEE and GPP and meteorological variables, and (iii) the analysis of the evapotranspiration regime and coupling coefficient by the Penman–Monteith big leaf analysis. The importance of the present work is corroborated by the fact that there are not many studies on atmospheric carbon exchanges, under severe drought conditions and anthropogenic disturbances, in this type of ecosystems.

2. Materials and methods

2.1. Site description

This study was part of the Carboeuroflux (2000–2003) and Carboeurope (2004–2008) projects. The experimental site is located in a 300 hectare eucalypt (Eucalyptus globulus Labill.) plantation (38°38′N, 8°36′W), extending from 700 m to 1800 m asl, part of Espirra Estate, and managed as a coppice. The site is located on a flat terrain, and the soil is a Dystric Cambisol with a mean depth of 1.3 m. Climate is of Mediterranean type with a long term average (1961–1990) precipitation of 709 mm and a mean annual temperature of 15.9°C. Trees were planted in 1986, with a distance of 3 m, following a twelve year rotation plan after a first nine year productive cycle. In October 2006, in the end of the second rotation, a felling was made to the 12 year trees of a 20 m height average. After felling, coppice sprouting regenerated the canopy. The new stems reached 7 m height in October 2009. In winter periods of and after 2007 air temperature fell below 0°C, and the young juvenile leaves were severely damaged by frost. A thinning of sprouts was made in October and November 2008 to remove 3 shoots out of 4 in each stump.

2.2. Instrumentation and calculations

The eddy covariance unit was installed at the top of a 33 m tower (13 m above canopy), and is comprised of an ultrasonic Gill R2 anemometer and an open path IRGA LI-7500 analyzer with a 21 Hz acquisition rate. Subsequently, after the felling, the eddy covariance unit was moved to a height of 12 m above the ground. The distance from the tower to the edge of the stand varied between 700 m and 1800 m. Calibration of the gas analyzer with a reference gas was carried out annually. Measurements of eddy fluxes in the constant flux layer were made since January 2002. Data for fluxes consisted in averages over 30 min periods. Half hour fluxes calculation involved
two axis coordinate rotation, linear detrending by least squares regression (Gash and Culf, 1996), Webb-Leuning (Webb et al., 1980) correction for density fluctuations and Schotanus correction for sonic temperature (Schotanus et al., 1983).

Meteorological data were sampled every 30 s with an automatic weather station (Campbell Scientific CR10 data logger) and averaged over 30 min periods. Precipitation was calculated using the integral of half hour periods data. Mean air temperature was measured at 25.2 m, 26.7 m, 29.2 m and 31.6 m with self produced Cu-Cons thermocouples of 0.15 mm diameter. The wind velocity was measured at the same heights as air temperature with cup anemometers (Vector Instruments, A110R), and wind direction was measured at the top of the tower with wind vane of the same brand, model W200P. Air humidity, incident solar radiation (Kipp &
Zonen, model CM 6B), photosynthetic active radiation (PAR) (SKYE Instruments, model SKE510), and net radiation (Campbell Scientific, model Q6) were also measured at the top of the tower. Soil moisture data were continuously recorded with a probe Delta-T, Model PR2 every 2 h since January 2007 at depths 10 cm, 20 cm, 30 cm, 40 cm, 60 cm and 1 m. For additional equipment descriptions refer to Rodrigues et al. (2005).

The flux related to storage change of carbon dioxide in the layer below the location of the eddy covariance system was calculated by the concentration measurement of CO₂ at 33 m following the approach of Greco and Baldocchi (1996), and added to the measured covariant flux. The extension of the homogeneous cover over the flat terrain is a guarantee for good fetch.

An integrated analysis for the period from 6 July to 29 November 2004 based on climatologic footprint analysis showed that a major contribution of 87% to the site’s atmospheric fluxes was due to the eucalypt forest (Göckede et al., 2005) with 97.6% of all fluxes exceeding the threshold of 80% contribution from the target land cover (Göckede et al., 2008). Possible underestimation of carbon fluxes motivated by the Licor 7500 open path analyzer heating effect (Burba et al., 2008) were considered negligible because air temperature rarely dropped below freezing and carbon fluxes were mostly large throughout the years.

The reported carbon fluxes were submitted to quality control procedures based on the three-flag scheme presented by Mauder and Foken (2004). After the calculation of the mean half hour fluctuations covariances, a filtering removed data fluxes corresponding to (i) deviations of mean vertical velocity from zero greater than 0.35 m s⁻¹, (ii) high frequency spikes affecting single instantaneous
measurements in a percentage above 1%, (iii) the existence of occasional spikes in the half hourly flux data, using the median of the absolute deviation about the median described by Papale et al. (2006) and (iv) a friction velocity below the threshold of 0.2 m s$^{-1}$ (Mateus et al., 2006). Flux data remaining after this filtering process were submitted to stationarity and integral turbulence characteristics. The average percentages of half hour data accepted for gap filling were 55% for carbon flux (77% at day and 33% at night) and 69% for latent heat flux (83% at day and 55% at night).

Gap filling and NEE partitioning in GPP and TER were made using the online software Eddyproc (2010) (http://gaia.agraria.unitus.it/database/eddyproc/EddyInputForm.html) according to the methodology proposed by Reichstein et al. (2005).

The analysis of the evapotranspiration regime was made by the evaluation of the decoupling coefficient $\Omega$ calculated as (Monteith and Unsworth, 1990):

$$\Omega = \frac{(\Delta/\gamma) + 1}{(\Delta/\gamma) + 1 + (r_c/r_a)}$$

with $r_a$ being the aerodynamic resistance (s m$^{-1}$), $r_c$ the canopy resistance (s m$^{-1}$), and $\Delta$ the rate of change of saturation vapour pressure with air temperature (Pa K$^{-1}$). The $\Omega$ coefficient is thus associated to canopy resistance and stomatal dynamics controlling water vapour and carbon dioxide fluxes. Typical values of decou-
Fig. 8. Averages of GPP for typical day: (a) 2002–03; (b) 2005; (c) 2008–09 (+, January–March; – – –, April–June, ♦, July–September; —, October–December).

Aerodynamic resistance \( r_a \) is given by (Monteith and Unsworth, 1990):

\[
ra = \frac{u}{u^*} \tag{2}
\]

where \( u \) is the mean horizontal wind velocity (m s\(^{-1}\)) and \( u^* \) is the friction velocity (m s\(^{-1}\)).

Canopy resistance, \( r_c \), was obtained inverting the Penman–Monteith equation (Monteith and Unsworth, 1990) using latent heat flux, \( LE \), obtained by eddy covariance,

\[
r_c = r_a \left( \frac{\Delta (R_n - G) + \rho c_p [e_S(T_a) - e] / r_a - \Delta}{LE} \right) / \gamma - 1 \tag{3}
\]

with \( R_n \) being the net radiation, \( G \) the soil heat flux (W m\(^{-2}\)), \( \gamma \) the psychrometric constant (Pa K\(^{-1}\)), \( e_S(T_a) \) the saturation vapour pressure, \( e \) the vapour pressure, and \( \rho \) and \( c_p \) the air density and specific heat at constant pressure, respectively.

Soil heat flux was assumed as the difference between net radiation and the sum of latent and sensible heat fluxes. Saturation vapour pressure \( e_S(T) \) was calculated by the following equation (Campbell and Norman, 1998),

\[
e_S(T_a) = 611 \exp \left( \frac{17.5027}{\frac{T_a}{234.57}} \right) \tag{4}
\]

with \( T_a \) being air temperature (\(^\circ\)C).
The long-term success of afforestation and the associated carbon sequestration potential in semi-arid climates must also be linked to the consequences in surface energy balance (Rotenberg and Yakir, 2010). For a preliminary analysis of the impact of drought in the radiative energy partition between sensible (H) and latent heat (LE) fluxes, we used the Bowen ratio ($\beta = H/LE$). The ratio $\beta$ is typically 0.4–0.8 for temperate forests and 2–6 for semi-arid areas (Oke, 1992).

Relationships between NEE and PAR were analyzed using half-hour data with higher quality (flags 0 and 1) in the months January, April and August for every year of the period considered. The analysis was restricted only to diurnal data ($R_g > 10$ W m$^{-2}$). In 2007, 2008 and 2009, November was additionally analyzed. The fitted expressions using the Marquardt method (Seber and Wild, 1989; SAS software, ver. 9.3.1, procedure NLIN, 2003) were based on the Michaelis–Menten equation,

$$\text{NEE} = \frac{\beta - \gamma \text{PAR}}{\alpha + \text{PAR}}$$ (5)

where $\beta$ is the respiration parameter derived by extrapolating the light response curve to zero irradiance, $\gamma$ the maximum rate of photosynthetic assimilation and $\alpha$ corresponds to the PAR radiation at which photosynthesis is one half of $\gamma$. All variables are expressed in $\mu$mol m$^{-2}$ s$^{-1}$. From these parameters, quantum yield can be obtained as the ratio $\gamma/\alpha$.

In order to establish possible useful practical equations relating the main meteorological variables with monthly NEE and GPP for the period preceding tree felling, a modelling approach was done based on the application of general estimating equations (GEEs) methodology. GEEs were developed by Liang and Zeger (1986) in the context of extending generalized linear models to Gaussian and non-Gaussian longitudinal clustered response data (Schaubener and Pierce, 2002). In GEEs, correlated data are modelled using the same link function and linear predictor as in the general independent case, with the difference that the covariance structure of the correlated measurements must also be modelled.

In this work GEE data analysis was done with the SAS software (ver. 9.3.1) procedure Genmod. Basically GEEs permit a consistent iterative, quasi-likelihood estimation of the vector of regression parameters $\theta$ as

$$\theta_{i+1} = \theta_i + \left( \sum_{n=1}^{N} \frac{\partial \mu'_i}{\partial \theta} \frac{1}{\sigma'_i} \frac{\partial \mu'_i}{\partial \theta} \right)^{-1} \times \frac{\partial \mu'_i}{\partial \theta} \frac{1}{\sigma'_i} (Y_i - \mu_i)$$ (6)

with $Y_i$ and $\mu_i$ corresponding, respectively, to the vectors of measurements and means in the $i$th subject, $\hat{V}_i$ an estimate of $V_i$, the covariance matrix of $Y_i$, and $N$, the total number of measurements. The term corresponding to the inverse of the summation in Eq. (6) is the model-based estimate of $V_i$, which would be used if $\hat{V}_i$ were the correct variance–covariance matrix. The GEE estimation uses a so-called “sandwich” or empirical estimator of the variance matrix of clustered quantitative variables by the various levels of the classification variables. This estimator includes a working correlation matrix (banded $m$-dependent, exchangeable or autoregressive), and successive estimates of covariance matrices allow to obtain iterative estimates of regression parameters, till convergence. An adequate choice for the working correlation structure is indicated by a reasonable similarity between matrices of model based and empirical covariance estimators (Hedeker and Gibbons, 2006).

The regression coefficients obtained by GEEs are consistent estimators of the population regression parameters (Fitzmaurice et al., 2004).

### 2.3. Biomass measurements

Measurements of total tree height, crown length and diameter at breast height were made in January 2002, 2003, 2005 and 2006 in five plots of 225 m$^2$ adjacent to the tower, for allometric estimation of annual carbon biomass. The equations used were those reported by António et al. (2007) for individual trees. Each fraction of biomass was individually estimated as a function of tree diameter, height, crown length and the dominant height of the stand. The total above ground biomass was calculated by the sum of all fractions.

### 3. Results and discussion

#### 3.1. Meteorological conditions and GEE equations

Table 1 shows, for the whole period, data of NEE and GPP and the concomitant variation of key meteorological variables such as average air temperature ($T_a$), incoming global radiation ($R_g$), precipitation ($P$), and evapotranspiration ($E$). These data are aggregated in annual and quarterly periods, aiming to investigate the seasonal and annual variations. Fig. 1 shows a steady seasonal pattern for $R_g$, phased with $T_a$ in the whole eight year period.

Averaged annual temperatures in 2003, 2004, 2005, 2006 and 2009 exceeded the long term mean of 15.90 °C (Table 1). Annual precipitation was the variable with the most significant variation relatively to the long term average (1961–1990) of 709 mm, with reductions of 47%, 44%, 37%, 29% and 20% comparatively to this average in 2004, 2005, 2007, 2008 and 2009, respectively. The prolonged drought of 2004 and 2005 was the most severe in 140 years (Garcia-Herrera et al., 2007). In 2004, $E$ was 722.55 mm, higher than in 2002 and 2003 and almost twice the precipitation (Table 1) due to soil moisture depletion. In 2005 $E$ decreased to 391.64 mm, of the same order of magnitude of the precipitation.

Monthly patterns of rain events were typically Mediterranean with almost no rainfall on summer months and more precipitation in winter and spring. The years of 2004 and 2005, besides the lower precipitation showed uneven monthly pattern of rain distribution along the year. Indeed, rainfall in the first quarter of 2004 (210.03 mm) corresponded to 55.5% of the precipitation in the whole year (Table 1) and in 2005 about 38% of the scarce rainfall occurred in March and about 40% in the last quarter (Table 1). The bulk evapotranspiration in 2004, 610.86 mm corresponding to 68% of the total, occurred in the period April–December characterized by the lowest rainfall.

Before the felling, the monthly averaged vapour pressure deficit (VPD) followed the trend of drought’s seasonal and annual conditions, with averages of 6.18 hPa, 7.56 hPa, 8.67 hPa and 7.31 hPa, in 2002–2003, 2004, 2005 and 2006, respectively. After the felling, the average monthly VPD for 2007–2009 lowered to 5.34 hPa. In the period 2007–2009, soil moisture increased with depth from 3.47% to 11.49% at 1 m, outreaching the wilting point for sandy soils (5%) below 60 cm.

GEE modelling was used for a quantification of the influence of the main meteorological variables in monthly carbon fluxes in the period from January 2002 to October 2006. An extensive analysis of distinct combinations was done regarding: classification (month, year and month nested in year); quantitative variables; and working matrices. The fitted equations, considering identity link function and the attested normal distribution of data, were:

$$\text{NEE} = -42.97 - 0.0903\text{PAR} + 0.0062\text{VPDm}$$

$$\text{GPP} = 98.67 + 0.16\text{PARm} + 0.0085\text{VPDm}$$ (7)

where the independent variables are the accumulated monthly data: VPDm (hPa) of VPD, and PARm (MJ m$^{-2}$) of PAR radiation. These equations are plotted in Fig. 2. The classification variable considered in the selected models was the month, and the pro-
posed working matrices types for these equations were banded 1-dependent (GPP), and autoregressive (NEE).

Table 2 shows some measures of these model equations. Statistics used to select the models were the similarity of empirical and model-based covariance matrices, z scores and p-values for regression parameters. The coefficients $R^2$, evaluated after the GEE model selection, improved comparatively to the ones of usual regression models.

Eqs. (7) show the influence of meteorological parameters associated with atmospheric humidity and radiation in NEE and GPP in the drought period, and are interesting under a practical point of view. The inclusion of vapour pressure deficit reflects the fact that VPD in forests exerts strong control in photosynthetic uptake (Baldocchi, 1997; McCaughey et al., 2006).

3.2. Impact of the two events in temporal variation of carbon fluxes

3.2.1. Daily patterns

A simple analysis of the graphs and results of carbon fluxes and meteorological variables at daily, monthly and annual timescales clearly demonstrates the link between variables associated to water stress, e.g., precipitation and water vapour deficit, and NEE and GPP. From Table 1 we can establish a criterion to divide the drought period of 2004 and 2005 in two stages. A first stage corresponds to 2004 with reduction in precipitation and annual carbon fluxes of the same order of magnitude of 2002 and 2003, and a second stage corresponds to 2005 when, under low rainfall, a drastic reduction in carbon uptake occurred.

During the eight year period, the daily uptake of carbon followed distinct patterns reflecting the distinct environmental conditions and disturbances. Inter-annual evolution of daily GPP, NEE, and TER is shown in Fig. 3. In 2002, 2003 and 2004 the percentages of days with carbon uptake were 92%, 90% and 89%, respectively. In 2005, the year when the drought effects in carbon fluxes were greatest, the percentage of days with carbon uptake lowered to 62%. In 2006 the percentage of days with carbon uptake recovered to about 78%, despite the tree felling in October and November. In 2007, 2008 and 2009, carbon uptake occurred in about 50%, 61% and 62% of the days, respectively.

Seasonal patterns of the hourly averaged typical day carbon and latent heat (LE) fluxes and meteorological data were analyzed on a quarterly basis in the periods 2002–2003, 2005 and 2008–2009 concerning, respectively, normal productive years, drought and sprouting after the felling (Figs. 4–9). As a rule $R_g$ phased and peaked with NEE and GPP at about noon, whereas TER phased with vapour pressure deficit and air temperature at about 15 h. The phasing of peaks of TER, GPP and NEE with air temperature, VPD, and $R_g$ is indicative of the driving role of these meteorological variables in the distinct carbon fluxes (Falge et al., 2002; Carrara et al., 2004).

In all the periods, an approximate synchrony between typical day curves of GPP and LE (not shown) was indicative of the fundamental role of stomatal closure in controlling atmospheric carbon exchanges and evapotranspiration. Typical day curves representative of NEE (Fig. 7), and GPP (Fig. 8) showed asymmetry, with maxima before noon, reflecting the effects of water stress in evapotranspiration and GPP.

In 2004 the typical day patterns (not shown) were similar to 2002 and 2003, revealing approximate phasing between LE and GPP curves and asymmetry in the July–September period. In 2005 three of the four NEE and LE curves analyzed followed an asymmetric pattern with maximum NEE and GPP occurring in the period January–March (Figs. 7 and 8).

In the period 2008–09 the eucalypt coppice behaved again as a carbon sink, with lower NEE and GPP in comparison to the period corresponding to the end of rotation cycle. The asymmetric patterns of typical day in summer were maintained (Figs. 7 and 8).

3.2.2. Annual and seasonal patterns

3.2.2.1. Drought effects. Annual and seasonal patterns of NEE, GPP and TER in the period 2002–2006 are shown in Table 1 and Fig. 11. Before the felling, monthly averaged NEE had averaged maxima of $-102.98$ g C m$^{-2}$ in mid-spring, and minima in late summer of $-10.53$ g C m$^{-2}$. This monthly maximum in mid-spring agrees with the discussion by Rotenberg and Yakir (2010) about a tendency of GPP time peaks in European pine forests shifting from July–August to mid-March, with decreasing latitude.

Under environmental conditions appropriate to eucalypt growing, global values of GPP and NEE in Espirra prior to the felling (Table 1) are high, comparatively to data reported in studies for other sites in Europe (e.g. Falge et al., 2002). A prevalence of GPP over total ecosystem respiration at Espirra is evidenced by the higher annual ratios GPP/TER, varying between 1.43 (2005) and 2.13 (2003). These values are higher than those reported by other studies, e.g., 1.35 from a fast growing beech forest of Hesse (France) in an eight year period (Granier et al., 2008), and 1.25 from 18 European forest ecosystems (Janssens et al., 2001).

Annual averaged remotely sensed MODIS LAI (Fig. 10) was 4.72 in 2002, 5.26 in 2003 and 5.33 in 2004, decreasing to 3.58 in 2005.

![Fig. 10](image)

Monthly averaged MODIS LAI in the whole period: measured values (○) and moving averages (−).
due to leaf yellowing under intense water stress. In 2006, before the felling, annual LAI recovered to 4.37 with a maximum of 5.95 in April.

In the period 2002–2006, monthly and quarterly GPP phased approximately with evapotranspiration (Table 1) as attested by the good linear relationship \( R^2 = 0.71 \), Fig. 12) between the monthly ratios GPP/LAI and \( E/LAI \).

Between 2002 and 2006, annual monthly averaged NEE and GPP varied inversely with vapour pressure deficit, with the September fall in carbon uptake coinciding with maxima averaged monthly VPD (Fig. 13). This is due to the fact that forest stomatal conductance tends to be higher at low VPDs, as shown, e.g., by Granier et al. (2000) for a set of 21 broadleaved and coniferous forest stands or by David et al. (1997) for Eucalyptus globulus in Portugal.

The effects of drought on NEE and GPP were felt mainly between September 2004 and December 2005 (Fig. 11). This period followed a six month period with a total rainfall of only 39.15 mm and an evaporation of 475.03 mm, when values of GPP, NEE and TER were of similar magnitude as those averaged from the same period in 2002 and 2003 (Fig. 11). In 2005 the totals of NEE (−356.64 g C m\(^{-2}\)) GPP (1255.11 g C m\(^{-2}\)) and TER (898.48 g C m\(^{-2}\)) (Table 1) were substantially lower than in the previous years, and the eucalypt stand behaved as a carbon source from July till November. The prolonged lack of rainfall inducing water stress and lower LAI was determinant for this restriction of carbon atmospheric exchange. Steady patterns of solar radiation and air temperature before the felling were never limiting factors to the development of continuous carbon uptake and photosynthesis along the annual periods.

The influence of drought in restraining carbon fluxes due to water limitations had been shown, e.g., by Migliavacca et al. (2009) in an intensively managed poplar forest in Zerbolò (Italy) and by Reichstein et al. (2002), in two Mediterranean holm oak (Quercus ilex) forests (Puéchabon, France). In the holm oak forests both GPP and respiration are constrained under drought and contingent on soil water contents at different depths. This evidence allows to hypothesize that, under the first drought stage in 2004 and given the low soil water content at surface, roots got access to lower soil depths. Thus, the decrease of GPP was minimized and the drought’s most severe consequences were delayed. Indeed, Moroni et al. (2003) indicated that 6 year old Eucalyptus globulus trees in Tasmania under drought stressed conditions were encouraged to develop a higher root frequency and length than under irrigated conditions, and were thereby able to penetrate the dry soil. The worst effects of drought were shown in 2005 when, despite the possibility of the trees to develop a root system able to tap water at deeper soil depth, water replenishment by the scarce rainfall was not enough to sustain the same levels of GPP as before.

NEE and \( \gamma \) in January–March 2005 (Figs. 10 and 17) were −190.29 g C m\(^{-2}\) (Table 1), and 25.37 \( \mu \)mol m\(^{-2}\) s\(^{-1}\), high when compared to equivalent periods in other years (e.g., 2002, or 2006, Table 1). This fact was mainly determined by the lower values of mean air temperature in January (8.94 °C) and February (8.71 °C) 2005, which restricted TER in the first quarter to 182.15 g C m\(^{-2}\) (Table 1). This TER value, simultaneous with a drought-induced decrease in GPP to 372.44 g C m\(^{-2}\), was the second lowest in the same periods of the years included in this study. In conjunction with air cooling, a precipitation of only 7.65 mm in January and February 2005 probably also contributed to a strong restraining of microbial soil respiration, enhancing thereby net carbon uptake.

The last month when the ecosystem behaved as carbon source (Fig. 11) was October 2005, when an increase of precipitation to 154.07 mm, corresponding to 39% of the annual total, replenished available water in the ecosystem resulting in peaks both in TER (128.34 g C m\(^{-2}\)), due mainly to soil respiration, and GPP (101.68 g C m\(^{-2}\)).

The decrease in carbon uptake in 2005 was due mostly to a decrease in GPP. Indeed, on an average monthly basis, GPP in 2005
decreased by about 35% relatively to the remaining years before the felling. The corresponding decrease in TER was 18%. The averaged monthly ratio between GPP and TER decreased from 2.13 in 2003 to 1.69 in 2004 and 1.43 in 2005, increasing to 1.58 in 2006. In the Mediterranean holm oak forest under drought conditions (Reichstein et al., 2002) the decline in this ratio is a symptom indicative of low soil water availability. Soil dryness induces a hormonal signal sent from the roots, causing stomatal limitation of gas exchanges (Baldocchi, 1997). The preponderance of GPP decrease agrees with the studies of Valentini et al. (2000) and Falge et al. (2002) concluding that the importance of ecosystem respiration in carbon exchange in European ecosystems increases with latitude.

In 2006 with the increase of precipitation, the eucalypt stand recovered its carbon sink capacity (Fig. 11) with a NEE of $-544.38 \text{ g} \text{C m}^{-2}$ in the months till the felling, equivalent to the carbon uptake in the same period of a normal year (e.g., 2002). The fast recovery of GPP and NEE after the drought, reflects the canopy plasticity and the reversibility of the mechanisms responsible for the drop of GPP and NEE in 2005.

3.2.2. Felling effects. The eucalypt stand recovered its sink capacity after June 2007, to totals of $-200.73 \text{ g} \text{C m}^{-2}$ and $-209.01 \text{ g} \text{C m}^{-2}$ in 2008 and 2009 (Fig. 14). In these years, carbon uptake lasted 233 days, from mid February till mid October, following thereby a pattern almost opposite to the trees in the term of their productive cycle. Monthly $\gamma$ and quantum yield variation (Fig. 17) agreed with the reduction of carbon uptake and with the change of carbon uptake’s seasonal pattern.

As happened with drought, the reduction in carbon uptake in 2008 and 2009 was due mostly to a decrease in GPP. Indeed, the monthly averages of GPP and TER decreased, respectively, 38% and 15% in 2008–2009 as compared to 2002–2003.

An estimated 13.6 ton biomass ha$^{-1}$ of litter (leaves, branches and twigs) left in the soil, corresponding to about 10% of the total biomass, remained after the felling in 2006. The heterotrophic decomposition of this biomass contributed to the high values, above $920 \text{ g} \text{C m}^{-2}$, of annual TER in 2007–2009 (Table 1).

The main explanation to the change of NEE seasonal pattern of young trees in 2008 and 2009, under similar weather conditions,
should be related to an improvement of summer water stress conditions, attested by a smaller asymmetry of NEE (Fig. 7) and LE typical day curves. This improvement was mostly due to the fact that the young coppice inherited from the felled trees a deep root system, which remained in the soil. The resulting imbalance between the root system and aerial plant in the coppice reduced the shoot/root biomass ratio of about 4.29 in mature eucalypt trees (Madeira et al., 2002) by a factor of about 3.20. The deep roots were therefore possibly able to extract enough water in summer below the 60 cm level, where soil moisture was higher than the wilting point of 5%, and water stress was minimized sustaining stomatal carbon uptake in a new canopy with smaller aerial plant size and lower LAI. Summer water stress after felling was also minimized both by a more scattered distribution of annual precipitation along the years and by reduced atmospheric VPD (Fig. 13).

Monthly averaged remotely sensed MODIS LAI also decreased to 2.41, 3.46 and 2.80 in 2007, 2008 and 2009, respectively (Fig. 10). During the three years after the felling, monthly LAI, GPP and NEE decreased in winter, due to the thinning in October to November 2008 and to the effects of frost in young leafs in winter in 2007, 2008 and 2009. Height growing of young trees, reaching about 7 m in October 2009, agreed with data from growing tables available for the eucalypt site (Goes, 1977).

3.2.3. WUE, energy partition and evapotranspiration regime

On a monthly basis, annual averaged water use efficiency, WUE, defined as the ratio of GPP to $E$, decreased from 2002 (4.84 g/L) and 2003 (3.36 g/L) to 2004 (2.75 g/L) and 2005 (3.32 g/L) (Fig. 15) with the onset of drought and increase in soil evaporation and VPD.

In 2004 with the beginning of the increase of evapotranspiration a steep decrease in monthly WUE occurred in April (Fig. 15) to 2.32 g/L, followed by a stationary pattern in the rest of the year. Monthly WUE showed higher seasonal variation in 2005 with a two-month peak in January and February 2005 caused by a drastic decrease in $E$ (5.07 g/L and 5.48 g/L), heightening the annual average to 3.32 g/L. A decrease in WUE followed with minima below 2 in July–September, due to a steeper decrease of GPP. Thus, drought under its first stage in 2004 depressed monthly WUE due mostly to an increase in evapotranspiration. In 2005, under the second stage, the decrease in WUE was caused by the high decrease in GPP.

To our knowledge, most studies with analysis of WUE patterns in forest ecosystems concern short drought summer periods. We think that, under a two year drought, the non-steady monthly WUE evolution reflects a more complex interaction between the fundamental underlying stomatal control, evident from phased curves of daily and seasonal GPP and $E$, and other non-stomatal factors (Baldocchi, 1997; Reichstein et al., 2002), e.g., decrease in mesophyll conductance, stomatal patchiness, and dynamics of soil moisture. In 2002 and 2003 the lesser monthly averaged VPD and decreased water stress were factors determinant to higher WUE, allowing also for a steadier monthly pattern. A steady pattern of monthly WUE, prevailed after January 2006 as well.

After the felling, annual averaged monthly WUE were 1.62 g/L, 1.86 g/L and 2.35 g/L, in 2007, 2008 and 2009 respectively, showing an increasing tendency, still lower than this in mature trees, motivated mainly by the lower GPP.

In the context of a prolonged drought, annual average monthly Bowen ratio was 0.55 in 2004 peaking to 1.70 in 2005, a value typical of transition from temperate to semi-arid regions (Oke, 1992). Bowen ratio lowered in 2004, under the drought’s first stage, when LE increased to almost twice the precipitation, due to the depletion of available soil water. The higher Bowen ratio in the drought’s second stage reflected a shift in radiant energy dissipation from evapotranspiration to convective heating and certainly contributed to an increase in leaf temperatures, promoting foliar photorepiration and reduction of carbon gain (Migliavacca et al., 2009; Reichstein et al., 2002).

Monthly evolution of decoupling coefficient, $\Omega$, obtained by inverting the Penman–Monteith equation is shown in Fig. 16. Summer decreases in all the years before felling reflected fluctuating atmospheric vapour pressure deficit and water stress conditions. Annual averaged values of $\Omega$ in 2004 and 2005 were 0.26 and 0.11, respectively. This increase in coupling to weather conditions also shows what happened in the drought’s two stages. Under the first stage in 2004 in a context of low rainfall, the high annual evaporation (722.55 mm, Table 1) was more dependent on the ecosystem’s available radiant energy, and in the second stage in 2005 stomatal control reinforced its role in restraining total evaporation (391.64 mm), contributing to the increase in Bowen ratio.

In the period from 2007 to 2009, the mean $\Omega$ value was 0.30 (Fig. 16), with a monthly flatter pattern along these years, with no decrease in summer. Thus, as expected, the sparser canopy of younger plants showed a lesser coupling to weather conditions than the denser canopy of trees corresponding to the end of their productive cycle.
events occurred: a long drought period between 2004 and 2005, and a tree felling in October and November 2006.

The impact in NEE and GPP of the strong reduction of annual precipitation in 2004 and 2005 to about 47% and 44% of the long term mean was felt mostly in 2005. Indeed annual NEE increased from $-856.56 \text{ g C m}^{-2}$ in 2002, $-791.33 \text{ g C m}^{-2}$ in 2003 and $-724.24 \text{ g C m}^{-2}$ in 2004, respectively, to $-356.64 \text{ g C m}^{-2}$ in 2005. This impact, beginning in April 2004, was twofold. In a first stage, in 2004, evapotranspiration almost doubled precipitation due to soil depletion, and NEE ($-791 \text{ g C m}^{-2}$) and GPP ($1834 \text{ g C m}^{-2}$) were not affected by the reduction of 47% of annual precipitation relatively to the long-term mean. The effects of drought were felt mainly in a second stage in 2005 when: evapotranspiration fell to 391.64 mm, of the same order of magnitude of precipitation; decoupling coefficient decreased to 0.11; Bowen ratio increased to 1.70; and NEE and GPP were reduced to $-357 \text{ g C m}^{-2}$ and 1255 g C m$^{-2}$, respectively. The seasonal pattern of carbon uptake in the period preceding the felling, characterized by a peak in late spring and a decrease in summer due to water stress, was not changed in 2005. Average typical day curves of LE and GPP phased, and their asymmetry increased in drought periods due to water stress and stomata control of tree transpiration. The reported decrease of NEE as a consequence of the two events was due mostly to a decrease to GPP. Average monthly WUE under the two events also diminished, except for January and February 2005. A GEE modelling approach to the carbon fluxes before the felling allowed for a quantification of the influence of VPD, PAR, radiation and precipitation on NEE and GPP on a monthly basis.

As expected, the felling induced a drastic reduction of sink capacity, with the young eucalypt coppice behaving as a carbon source in the first seven months of the new rotation. The seasonal pattern of GPP in 2008 and 2009, with a higher level in summer and a decrease in winter, was distinct from the one before the felling. In summer, this was due mainly to an enhanced capacity of the deep root system of young plants, inherited from the canopy, to extract water from deeper soil horizons, thereby minimizing summer water stress. The decrease of GPP in winter was mainly related to the diminishing of LAI motivated by an enhanced sensitivity of young leaves and shoots to frost and harsher weather conditions. The Michaelis–Menten parameters (maximum rate of photosynthetic assimilation and quantum yield) followed the seasonal tendencies of NEE before and after the felling.

All the results demonstrated the ability of the eucalypt forest as an annual carbon sink, the interplay between atmospheric carbon and water fluxes, and the clear restricting role of this drought in the carbon sinking at daily, seasonal and annual timescales. Long-term Mediterranean forest climate projections should thereby address the impact of prolonged droughts in carbon sequestration, under distinct scenarios.

### Acknowledgements


The authors thank the field facilities made available by Portucel (Herdade da Espirra: “Grupo Portucel Soporcel”, Complexo Industrial de Setúbal, PO Box 55, 2901-861 Setúbal, Portugal) and their respective institutions, namely: 1st author (A. Rodrigues), Instituto Nacional de Recursos Biológicos (National Institute for Biological...
Resources); and corresponding author (M. Casquilho), Centro de Processos Químicos (Centro for Chemical Processes, Technical University of Lisbon).

References


Chapter 5:
Alexandra Marques, João Rodrigues, Manfred Lenzen and Tiago Domingos, "Income-based environmental responsibility".
Income-based environmental responsibility

Alexandra Marques a,⁎, João Rodrigues a, Manfred Lenzen b, Tiago Domingos a

⁎ IN+, Center for Innovation, Technology and Policy Research, Energy and Environment Scientific Area, Mechanical Engineering Department, Instituto Superior Técnico, Technical University of Lisbon, Av. Rovisco Pais n°1, 1049–001 Lisbon, Portugal

a Integrated Sustainability Analysis, School of Physics A28, University of Sydney, NSW 2006, Australia

1. Introduction

In the economic process primary factors of production, such as natural resources, labor or capital are applied in production processes to deliver consumption goods and services. Typically these processes generate adverse environmental pressures. There is a widespread perception that the current levels of consumption and the consequent environmental pressures are unsustainable, leading to environmental problems, such as changes in climate patterns, scarcity of drinking water or degradation of arable land. In order to control and minimize the negative externalities that result from the economic process it is necessary to quantify environmental pressures and to allocate responsibility for them to the economic agents involved.

Climate change is currently a priority area of environmental policy (EEA, 2010; OECD, 2008; UNEP, 2007), with a lot of attention focused on anthropogenic greenhouse gas (GHG) emissions. Many scientists believe that in order to reduce these risks a substantial reduction of GHG emissions is needed (mitigation).

The global nature of climate change requires global action; for example, if one large emitter does not commit to its responsibility and mitigation target, it is unlikely that the rest of the world can compensate for it. Moreover any emitter that stays out of any agreement will benefit from the action taken by others; mitigation efforts to cope with climate change can be considered public goods which allow for free-riding phenomena (Stern, 2007), potentially impairing or delaying climate policy. The extent of global participation is critical because the higher the participation rate in any action taken, the lower the cost (OECD, 2003). One of the goals of the current climate debate is to design fair and equitable climate policies that are globally accepted.

In this work, we explore a principle of responsibility seldom addressed in the literature, downstream responsibility. First, Section 2 describes production-based responsibility, the metric adopted by current climate policy. Section 3 describes consumption-based responsibility, a frequently proposed alternative to production-based responsibility. Section 4 describes downstream responsibility and presents the rationale for a novel nomenclature. Section 5 compares the three approaches. In Section 6 a case study is presented. Finally, Section 7 concludes.

2. Production-based Responsibility

Since 1992, in the Earth Summit, international negotiations are in place seeking a global agreement for the attribution of GHG emissions responsibility and reduction. But only in 1997 significant results were achieved with the Kyoto Protocol. Under the Protocol each country should report, through a national GHG inventory, the ‘emissions and removals taking place within national (including administered) territories and offshore areas over which the country has jurisdiction.’ For the Annex I countries who ratified it, a binding target of GHG emissions’ reduction was also established: 5% (on average), during the 2008–2012 period against 1990 levels (UNFCCC, 1998).

Under the Kyoto Protocol, a country holds responsibility for all the emissions that are directly generated by the production processes that take place within its borders. This type of responsibility is often called producer-based responsibility, and accounts for the direct emissions of a country (Fig. 1).

The geographic boundary established for GHG inventories does not consider the transfer of emissions through international trade, enabling
carbon leakage (Pedersen and de Haan, 2006; Peters et al., 2011b), and creates difficulties for the assignment of emissions generated through international activities, like aviation and maritime transportation (currently not included in national inventories) (IPCC, 1997).

The limitations of this approach became evident when big emitters, like the USA and China, refused to ratify the protocol or to commit to binding targets. Underlying these decisions were issues of fairness and fear of impaired competitiveness among countries (Whalley and Walsh, 2009). Developing economies that are highly dependent on exports, like China and India, claim that they should not bear the responsibility for production from which they do not benefit in terms of consumption (BBC, 2009). On the other hand, the USA and other developed economies, fear that their economies’ competitiveness will be impaired if they have to cope with any binding target that is not also applied to developing economies. Environmental regulations can draw away investors, promoting the relocation of industries to environmentally unregulated economies (the pollution haven hypothesis) and enhancing any potential carbon leakage (Stern, 2007). These positions have not changed, as could be seen in the COP17 meeting, held in Durban in 2011. In these negotiations, USA, China and India acknowledged and committed to the need of curbing carbon emissions, however they did not establish any target. On the other hand Canada withdrew from the Kyoto Protocol while Japan and Russia refused to sign for a second commitment period (UNFCCC, 2012).

To address the issues of competitiveness and carbon leakage, there are two often suggested mechanisms: they are, an integrated emissions trading system and carbon-motivated border tax adjustments (Alexeeva-Talebi et al., 2008; Ismer and Neuhoff, 2004). In an integrated emissions trading system firms exporting to participating countries are included, and requested to obtain emissions certificates for the CO₂ content of their exports (Alexeeva-Talebi et al., 2008). Carbon-motivated border tax adjustments consist in tariffs imposed on imports from foreign producers that should reflect the cost of carbon as if the goods imported were to be produced in the participating country (Ismer and Neuhoff, 2004; Whalley, 2009). This mechanism also allows rebates for the domestic carbon tax on exported goods (Alexeeva-Talebi et al., 2008; Ismer and Neuhoff, 2004). Thus, participating countries would be partially refunded from their carbon abatement costs, whereas non-participating countries exporting to participating countries would face a penalty (Ismer and Neuhoff, 2004) and thus have an incentive to join the scheme. The effectiveness of these measures as well as their compatibility with the World Trade Organization are issues still under debate (Alexeeva-Talebi et al., 2008; de Cendra, 2006; Ismer and Neuhoff, 2004; Whalley, 2009).

Another option to overcome production-based responsibility limitations is to consider a measure that includes indirectly generated emissions (Eder and Narodoslawsky, 1999; Peters and Hertwich, 2008b). Indirect emissions can be generated due to two distinct and opposite driving forces: supply and demand (Eder and Narodoslawsky, 1999). The responsibility principles that result from this are describe in the next Sections.

3. Consumption-based Responsibility

 Consumption-based responsibility has been discussed by several authors, as an alternative to production-based responsibility that could address its main limitations, namely carbon leakage (for example, Ahmad and Wyckoff, 2003; Andrew and Forgie, 2008; Bastianoni et al., 2004; Davis and Caldeira, 2010; Eder and Narodoslawsky, 1999; Munksgaard and Pedersen, 2001; Peters and Hertwich, 2008a; Peters et al., 2011b).

This metric measures the emissions generated to produce a country’s final demand for goods and services. These equal the emissions stemming from within the national territory minus the domestic emissions required to generate exports plus the foreign emissions required to generate imports (Fig. 1). So in fact, consumer-based responsibility is a carbon trade balance, as was pointed out by Rodrigues et al. (2010), Serrano and Dietzenbacher (2010) and Kanamoto et al. (2012).

For a certain product, this metric takes into account all the emissions generated along its supply chain prior to the delivery to final demand; for that reason these are often called upstream embodied emissions.

The adoption of this type of responsibility is supported by China. Around 20% of China’s emissions are generated in the production of exports, therefore China has claimed that a fair agreement should take into consideration that those emissions take place to produce goods that are not consumed by the Chinese people: ‘We produce products and these products are consumed by other countries, especially the developed countries. This share of emissions should be taken by the consumers but not the producers’ (BBC, 2009).

Unlike production-based inventories, consumption-based responsibility accounts for the emissions generated through international trade, minimizing the effects of carbon leakage by holding countries responsible for the emissions embodied in their trade balance (Andrew and Forgie, 2008; Bruckner et al., 2009; Marques et al.; Peters, 2008).

A measure of responsibility based on consumption takes into consideration international trade. In the context of climate policy the application of this responsibility principle would work as a type of “border adjustment,” but not interfering with WTO regulations, since it is not formally a trade-measure (Peters, 2008). However, consumption-based GHG inventories can be used to measure the carbon content of imported goods and services and aid in the implementation of border tax adjustments that would reflect the costs of a certain product in terms of carbon (Peters, 2008).

Fig. 1. Schematic representation of a country’s (A) production-based responsibility, (B) consumption-based responsibility and (C) income-based responsibility. Arrows represent the flow of emissions.
4. Income-based Responsibility

Another metric of environmental pressure is downstream responsibility (Gallego and Lenzen, 2005; Lenzen and Murray, 2010; Lenzen et al., 2007; Rodrigues and Domingos, 2008; Rodrigues et al., 2006, 2010). This metric measures the emissions enabled by primary suppliers and that are required to generate a country’s income through wages, profits and rents (payments to primary factors of production). These equal the emissions generated within national boundaries minus the domestic emissions generated downstream of imported products plus the foreign emissions generated downstream of exported products (Fig. 1).

From a technical point of view, the accounting of downstream responsibility has the same characteristics as consumption-based responsibility; it also accounts for the emissions generated through international trade and thus can be used to minimize the effects of carbon leakage.

Downstream responsibility also works as a carbon trade balance (Rodrigues et al., 2010). For a certain product, this measure accounts for all the emissions generated downstream in its supply chain until delivery to final demand; for that reason (Lenzen and Murray, 2010) named these types of emissions downstream enabled emissions.

There are applications of the concept at the country level (Lenzen and Murray, 2010; Marques et al.; Rodrigues et al., 2010), and at the corporate level (Lenzen and Murray, 2010; WBCSD and WRI, 2010). For example, at the country level Rodrigues et al. (2010) analyzed the downstream responsibility for six world regions, and showed that developed economies and fossil fuel exporters are the regions who enabled more emissions to generate their income, which indicates that these regions are receiving an economic benefit from emissions that did not occur within their borders. Lenzen and Murray (2010) studied the downstream responsibility associated with the provision of inputs to a certain number of Australian industries: this study was the first quantification of downstream responsibility at the industry level. Moreover Lenzen and Murray (2010) also provided a substantial clarification of the term, by providing a match between the less known downstream-based vocabulary and the well-known consumption-based vocabulary. At a corporate level a recent study (Schücking et al., 2011) showed how the supply of capital by banks, through investment decisions, enabled the emissions of GHG. The results showed that the self-proclaimed greener banks are actually those who enable more GHG emissions through investments in the coal industry (Petherick, 2012; Schücking et al., 2011).

Downstream responsibility has never received the same discussion as its consumption-based cognate. The current market-based economy puts emphasis on the consumption process since its main driving forces are consumers’ needs (Lenzen et al., 2007). Therefore, accepting that who benefits from emissions are the consumers is a natural choice. Consumers can opt for products whose production causes less environmental impacts, pushing companies to adopt cleaner production processes. But the price of such products is likely to differ from the price of the dirtier alternative (Laroche et al., 2001). Thus, the expression of environmental preferences through purchases cannot be dissociated from the budget constraint allocated to consumption and, ultimately, to the consumer’s source of income.

What if more environmentally efficient products are the most expensive ones? The truth is that most of the time the cost of a product influences the consumer choice because it will reduce its real income. People want to consume more and for that need to generate more income. This is done through the supply of primary factors of production.

In a downstream approach the emphasis is put on how this money is earned. The benefit from the generation of the emissions is passed to suppliers in the form of income. Actual examples of applications of this type of responsibility are easy to find. For example, if we accept that cigarette companies should compensate their customers for any health problem they might incur due to cigarette consumption, we are following a downstream responsibility principle. Or, for example, if we believe that banks and other financial institutions should follow the Equator Principles (EP, 2006), then we are also following a downstream responsibility principle.

However, the size of the literature and research on this topic is modest. We believe that this situation results from lack of clear intuition on the concept due to a naturally oriented market-based thinking. Here we aim to clarify and illustrate the potential applications of the concept of downstream responsibility, continuing the work of Lenzen and Murray (2010).

Will the general public recognize the term ‘upstream responsibility’ as meaning ‘consumption-based responsibility’? The term upstream, per se, indicates the direction from where carbon responsibility arrives to an agent, but it says nothing about the economic process that caused the emissions. Thus this term was replaced in the literature by a more intuitive one: consumer-based responsibility. Downstream responsibility also describes from where carbon responsibility arrives to an agent, but it also says nothing about the economic process that caused the emissions. In this case the process in question is the supply of primary inputs to production processes that enable emissions to occur. Therefore the term downstream responsibility could be replaced by supply-based responsibility. However if we agree that responsibility should be placed on those that take some benefit from carbon emissions, supply-based responsibility is not informative regarding that benefit, whereas consumption-based responsibility is. The supply of primary inputs, to a production process involves the provision of labor (or any other primary factor of production) in return of a salary, or, in a broader view, an income. Therefore we propose the term income-based responsibility to address downstream responsibility.

Whereas the use of consumption-based responsibility as a metric to levy a carbon tax would be a tax on consumption, the equivalent use of income-based responsibility would function as an income tax. For example, such a tax would reduce more strongly the income earned by a shareholder of a coal-fired power plant than the income earned by an investor in renewable energy. This assuming that producers are not facing perfectly inelastic demand and that therefore the amount of the tax is not completely passed on to consumers. The same logic applies to countries. The way the income is generated, in a country whose main activity is oil extraction, would enable more emissions, thus resulting in a higher responsibility (and thus a higher tax rate), than if that country’s main activity was fruit production.

A recent report by IDE-JETRO and WTO (2011) draws attention to the drastic changes that occurred in the structure of international trade in the last decades. Many products are no longer made in a single country, but instead production chains have become fragmented, with different countries specializing in specific stages of the supply chain, leading to a move from trade in goods to a trade in tasks (Hummels et al., 2001; IDE-JETRO and WTO, 2011). This shift increases the trade volume of intermediate goods, which are re-exported several times during the processing stage, before reaching the country of assembly into a final good, which can itself be exported.

The emergence of global production chains changed the paradigm of international trade, from a situation in which the last step in a supply chain accounted for the most added value to a situation in which it only represents a small fraction (IDE-JETRO and WTO, 2011). Under this current paradigm new statistical metrics, complementary to the traditional ones, need to be implemented in order to provide a clear view of international trade. IDE-JETRO and WTO (2011) propose the use of international trade of value added. This measure enables the correct determination of the relative importance of each region that takes part in a global supply chain. This report and the new framework it presents opens the door for wider applications of income-based responsibility.

5. Comparison of Responsibility Principles

The extent to which an agent is responsible for a certain environmental pressure depends on the responsibility principle chosen. Different
principles will yield different results, mainly due to differences in the boundaries adopted (Eder and Narodoslawsky, 1999).

In production-based responsibility the agent is only responsible for the direct pressures it generates. In consumption and income-based responsibility the boundary is expanded to include all the pressures, direct and indirect, that occur respectively upstream of the activity under consideration (consumption-based), or downstream of the activity under consideration (income-based) (Eder and Narodoslawsky, 1999; Lenzen and Murray, 2010; Rodrigues et al., 2006; Zhou et al., 2010).

Production-based responsibility accounts for emissions generated through production processes. The agent who directly emits should hold responsibility for its emissions. In the consumption-based responsibility all the emissions that were generated upstream along the supply chain to deliver goods and services to final demand are accounted. The agent holding the responsibility is the consumer, who is viewed as the agent who benefits from emissions; had consumers not demanded goods and services the emissions would not have occurred. The income-based perspective accounts for all the emissions that were generated downstream along the supply chain due to the supply of primary factors of production. The agent holding the responsibility is the supplier of primary factors of production, who benefits from emissions because it receives a payment from the inputs supplied. Both producer and income-based approaches see income as the benefit taken from emissions. However the boundaries differ; for the former only the production process is analyzed, accounting for direct effects; whereas in the latter the whole supply chain is taken into consideration and both direct and indirect effects are accounted for. Consumption and income-based accounting work as a carbon trade balance. For an agent consumption-based responsibility considers the emissions embodied in the goods it imported, and leaves out the emission embodied in the goods exported (Kanemoto et al., 2012; Rodrigues et al., 2010; Serrano and Dietzenbacher, 2010). For income-based responsibility, the emissions enabled by the goods it exported are considered whereas the emissions enabled by the goods imported are left out (Rodrigues et al., 2010). As a result, agents who embark on international exchanges typically exhibit more different results depending on the responsibility principle analyzed. In Table 1 a succinct comparison between these responsibility principles is presented.

As mentioned, the adoption of a responsibility principle will yield different outcomes for different agents, making them better or worse off depending on the principle chosen. If the production-based responsibility principle is applied exporting countries will display higher responsibility than developing countries (see, for example, Davis and Caldeira, 2010; Peters and Hertwich, 2008a; Rodrigues et al., 2010). This metric is seen as fairer by developing economies than production-based responsibility, as it would force developed economies to change their lifestyle patterns in order to meet their GHG goals (Zhou et al., 2010). To date and up to our knowledge, there is no concrete policy recommendation based on consumption-based responsibility. However, we believe that if such a policy would be implemented it would raise issues of fairness as production-based responsibility raised. For example, consuming countries provide an economic benefit to producing countries that is not accounted for if a strictly consumption approach is followed.

Through an income-based approach, it would be primary suppliers, instead of producers or consumers who would assume total responsibility for the emissions generated. Countries exporting goods who will enable emissions in other regions will be worse off than countries whose exports do not enable emissions. This metric would push countries to analyze how they generate their income. It is contradictory that a country that domestically promotes clean technologies generates the majority of its income through the exports of fossil fuels. It is reasonable to accept that it is fair that those benefiting more, in terms of income, from emissions should bear the responsibility. However, in terms of climate policy, any measure placing full responsibility on one actor is unlikely to be accepted, since it will never be perceived as fair for all the agents involved in negotiations.

One option to overcome this issue is the adoption of a shared responsibility approach. With this type of approach responsibility is distributed, according to a pre-defined criteria, amongst the different agents. Several combinations of producer, consumer and income-based responsibility are possible. Rodrigues et al. (2010) provide an extensive review of this topic. For example, Bastianoni et al. (2004) developed a shared responsibility indicator, the Carbon Emissions Added (CEA). For a country this would equal their producer responsibility, plus total upstream embodied emission, normalized by the total of emissions worldwide. Ferng (2003) proposed a shared responsibility indicator, according to which a country would be responsible for a share of its production-based responsibility plus a share of its consumption-based responsibility. Ferng (2003) states that for each country a fair share should be established but no allocation rule was determined. Gallego and Lenzen (2005) and Lenzen et al. (2007) developed a shared responsibility framework that assigns responsibility either to producers and consumers (production-based and consumption-based responsibilities) or producers and workers-investors (production-based and income-based responsibility). The share of responsibility allocated to each agent should be proportional to its value added (Lenzen et al., 2007). Rodrigues et al. (2006) axiomatically derived a shared responsibility indicator that is the average between an agent’s consumption-based responsibility and income-based responsibility.

There are some empirical results on shared responsibility (for example, Andrew and Forgie, 2008; Cadarso et al., 2012; Ferng, 2003; Lenzen, 2007), focusing on producers and consumers. We believe that the lack of results on shared responsibility including suppliers is due to the lack of understanding of this approach, as referred in Section 4.

The construction of a GHG inventory based on the production-based responsibility principle, or on the consumption or income-based principles, requires different types of data. Production-based inventories use data that are already available through statistical offices (IPCC, 1997), and therefore present less uncertainty (Peters, 2008). Consumption-based and income-based inventories need more data, that is not usually available through statistical offices, namely bilateral trade data, and thus have more uncertainty associated (Peters, 2008; Peters et al., 2011b). As a consequence, a production-based responsibility measure is probably seen as more transparent and straightforward than a measure based on consumption or income-based responsibility. To overcome this, the scientific community has been committed to provide reliable data that can be used to compute the latter types of responsibility. Examples of this are the recently launched global databases: EORA from the University of Sydney (Lenzen et al., 2012), EXIOPOL and

<table>
<thead>
<tr>
<th>Responsibility</th>
<th>Production</th>
<th>Consumption</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effects</td>
<td>Direct</td>
<td>Indirect</td>
<td>Indirect</td>
</tr>
<tr>
<td>Emissions</td>
<td>Direct</td>
<td>Embodied</td>
<td>Enabled</td>
</tr>
<tr>
<td>Direction</td>
<td>-</td>
<td>Upstream</td>
<td>Downstream</td>
</tr>
<tr>
<td>Scope</td>
<td>Producer</td>
<td>Final consumer</td>
<td>Primary suppliers</td>
</tr>
<tr>
<td>Agent</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
A. Marques et al. / Ecological Economics 84 (2012) 57–65
Table 2
Producer, consumer, and income-based responsibilities, population and GDP of GTAP
regions (year 2004) (Mt CO2, million, thousand million USD2004).
GTAP Region's name
code
AUS
NZL
XOC
CHN
HKG
JPN
KOR
TWN
XEA
KHM
IDN
LAO
MYS
PHL
SGP
THA
VNM
XSE
BGD
IND
PAK
LKA
XSA
CAN
USA
MEX
XNA
ARG
BOL
BRA
CHL
COL
ECU
PRY
PER
URY
VEN
XSM
CRI
GTM
NIC
PAN
XCA
XCB
AUT
BEL
CYP
CZE
DNK
EST
FIN
FRA
DEU
GRC
HUN
IRL
ITA
LVA
LTU
LUX
MLT
NLD
POL
PRT
SVK
SVN
ESP
SWE
GBR

Australia
New Zealand
Rest of Oceania
China
Hong Kong
Japan
South Korea
Taiwan
Rest of East Asia
Cambodia
Indonesia
Lao PDR
Malaysia
Philippines
Singapore
Thainland
Vietname
Rest of Southeast
Asia
Bangladesh
India
Pakistan
Sri Lanka
Rest of South Asia
Canada
United States of
America
Mexico
Rest of North
America
Argentina
Bolivia
Brazil
Chile
Colombia
Ecuador
Paraguay
Peru
Uruguay
Venezuela
Rest of South
America
Costa Rica
Guatemala
Nicaragua
Panama
Rest of Central
America
Caribbean
Austria
Belgium
Cyprus
Czech Republic
Denmark
Estonia
Finland
France
Germany
Greece
Hungary
Ireland
Italy
Latvia
Lithuania
Luxembourg
Malta
Netherlands
Poland
Portugal
Slovakia
Slovenia
Spain
Sweden
United Kingdom

Producer

Consumer Income

Pop.

GDP

315.27
28.34
17.22
4071.13
54.70
924.98
344.35
220.70
75.80
2.81
295.57
1.40
125.32
67.38
38.20
192.72
72.93
7.44

305.47
34.30
17.75
3147.11
97.23
1214.09
335.40
167.41
57.27
3.71
261.55
2.00
68.91
72.68
58.33
144.05
67.97
9.08

376.61
19.94
29.45
3.99
15.56
8.71
3450.95 1307.99
92.20
6.96
1058.83
127.92
316.03
47.64
199.11
22.76
65.68
25.36
2.19
13.80
307.17
220.08
1.11
5.79
145.18
24.89
52.22
81.62
46.76
4.27
148.13
63.69
55.31
83.12
18.02
51.30

637.79
96.44
21.28
1674.13
163.01
4658.74
676.50
305.29
25.59
4.88
254.70
2.45
114.90
84.48
106.81
161.70
43.03
5.59

28.78
919.76
111.19
10.86
8.35
460.01
4879.14

41.11
860.79
126.67
15.63
13.97
424.99
5511.71

25.00
139.21
55.91
760.23 1087.12
641.26
88.15
154.79
94.73
9.25
20.57
20.08
7.77
56.32
13.90
512.81
31.96
979.13
4650.48
295.41 11673.38

327.08
3.15

353.65
4.95

189.41
1.49

105.70
0.13

683.24
5.89

118.20
8.96
234.81
54.98
45.19
17.31
2.87
25.09
4.02
123.52
1.86

88.41
8.52
215.53
44.07
48.14
21.09
4.55
30.06
6.30
88.30
2.37

124.44
9.39
241.66
52.67
62.14
26.82
4.79
27.98
3.39
179.07
1.23

38.37
9.01
183.91
16.12
44.92
13.04
6.02
27.56
3.44
26.28
1.39

150.40
8.78
616.54
89.64
97.46
29.97
8.42
68.63
13.69
108.23
3.52

4.14
8.47
3.51
4.87
11.00

6.39
13.49
4.56
7.86
16.51

5.11
6.47
1.85
5.24
8.51

4.25
12.29
5.38
3.18
14.07

19.47
27.45
4.39
12.60
24.15

142.85
52.27
72.39
7.05
99.41
44.27
15.03
57.67
255.58
599.25
74.78
42.71
33.97
332.60
6.45
9.42
9.73
2.73
165.81
240.70
50.19
24.67
12.62
266.76
37.41
438.29

139.65
82.86
124.15
9.24
81.15
62.00
13.55
69.19
410.46
804.45
94.89
52.20
46.59
476.05
11.84
14.51
11.25
3.43
172.01
212.64
64.34
25.91
13.93
324.69
69.97
657.35

89.92
68.14
85.20
5.45
81.68
65.20
10.68
64.69
302.32
733.92
77.69
37.33
52.95
340.80
5.85
9.15
9.83
1.56
155.14
202.00
45.92
20.51
12.97
258.40
64.65
541.88

38.45
8.17
10.42
0.83
10.23
5.41
1.34
5.24
60.26
82.65
11.10
10.12
4.08
58.03
2.32
3.44
0.45
0.40
16.23
38.56
10.44
5.40
1.97
42.65
9.01
59.48

193.12
292.31
352.31
15.42
108.03
243.73
10.22
185.92
2046.47
2740.50
205.20
99.65
182.24
1677.82
13.47
21.20
31.86
5.32
578.98
233.62
167.72
41.55
32.52
1039.90
346.41
2123.60

61

Table 2 (continued)
GTAP Region's name
code
CHE
NOR
XEF
ALB
BGR
BLR
HRV
ROU
RUS
UKR
XEE
XER
KAZ
KGZ
XSU
ARM
AZE
GEO
IRN
TUR
XWS
EGY
MAR
TUN
XNF
NGA
SEN
XWF
XCF
XAC
ETH
MDG
MWI
MUS
MOZ
TZA
UGA
ZMB
ZWE
XEC
BWA
ZAF
XSC

Switzerland
Norway
Rest of EFTA
Albania
Bulgaria
Belarus
Croatia
Romania
Russian
Federation
Ukraine
Rest of Eastern
Europe
Rest of Europe
Kazakhstan
Kyrgyzstan
Rest of former
Soviet Union
Armenia
Azerbaijan
Georgia
Iran
Turkey
Rest of West Asia
Egypt
Morocco
Tunisia
Rest of Norh Africa
Nigeria
Senegal
Rest of West
Africa
Rest of Central
Africa
Rest of South C.
Africa
Ethiopia
Madagascar
Malawi
Mauritius
Mozambique
Tanzania
Uganda
Zambia
Zimbabwe
Rest of Eastern
Africa
Botswana
South Africa
Rest of SACU

Producer

Consumer Income

Pop.

GDP

26.69
52.45
4.62
4.24
41.83
50.59
15.20
76.53
1332.95

72.40
46.51
5.64
5.77
31.29
43.54
20.30
69.01
1016.77

59.52
143.27
4.68
3.82
32.12
48.50
11.58
58.11
1464.48

7.24
4.60
0.32
3.11
7.78
9.81
4.54
21.79
143.90

357.54
250.05
15.71
8.99
24.57
21.96
33.93
74.42
569.84

217.62
5.89

126.61
8.15

159.47
3.05

46.99
4.22

60.98
2.60

70.96
161.61
5.18
132.88

68.06
134.85
5.71
94.27

56.35
157.05
4.66
128.90

14.29
14.84
5.20
37.40

44.98
44.35
2.21
20.20

3.38
24.18
2.43
299.80
163.34
909.16
120.29
31.86
18.38
127.64
39.92
4.15
19.85

4.29
26.86
4.67
301.86
192.71
707.86
101.71
38.01
18.61
110.65
38.16
5.91
34.51

2.47
21.59
2.36
340.16
153.30
1235.51
113.73
27.92
14.76
191.62
101.83
2.24
20.36

3.03
8.35
4.52
68.80
72.22
118.40
72.64
31.02
10.00
38.10
128.71
11.39
117.42

3.34
8.73
4.47
157.86
295.83
691.10
76.81
50.25
27.99
112.39
68.57
7.20
50.73

7.80

11.54

27.42

35.36

38.01

9.09

14.40

37.72

71.34

23.89

3.70
1.36
0.55
1.83
1.60
3.06
2.26
1.77
8.78
21.04

6.74
2.03
1.57
3.80
3.40
6.51
3.51
3.09
6.89
34.27

2.34
1.84
0.63
2.43
2.04
2.99
2.90
2.14
7.81
25.33

75.60
18.11
12.61
1.23
19.42
37.63
27.82
11.48
12.94
99.73

7.28
4.35
1.79
5.92
6.09
11.47
7.27
5.40
4.08
50.19

3.76
6.38
329.12
213.38
3.44
6.33
21731
21731

4.53
1.77
8.72
312.64
47.21
213.93
4.86
4.84
9.06
21731
6405
40962

Producer, consumer, and income-based responsibilities, population and GDP of GTAP
regions (year 2004) (Mt CO2, million, thousand million USD2004).

WIOD databases from European consortiums (Timmer, 2011; Tukker
et al., 2009) and the Asian IDE-JETRO (IDE, 2006).
6. Case Study
In this Section we present the results of the quantiﬁcation of
income-based responsibility, for 112 world regions (either countries or
groups of countries), using the Global Trade Analysis Project (GTAP) database. We compare the results with production and consumption-based
responsibilities.
6.1. Data and Methodology
We used the GTAP 7.1 database (Narayanan and Walmsley, 2008) to
build a full Multi-Regional Input–Output (MRIO) model (Peters et al.,
2011a) covering the whole world, for the year 2004. An input–output
(IO) model provides a description of an economy, for a given year


based on monetary flows between industrial sectors (Miller and Blair, 2009). The distinctive feature of a MRIO model is that it allows the analysis of the interactions and interdependencies between different sectors in different countries.

If \( \mathbf{e} \) is the vector of direct GHG emissions of every sector in the world economy, the vectors of upstream emissions embodied in final demand, \( \mathbf{u} \), and of downstream emissions embodied in primary inputs, \( \mathbf{d} \), are calculated as:

\[
\mathbf{u} = (\mathbf{e} \mathbf{x}^{-1}) (\mathbf{I} - \mathbf{z} \mathbf{x}^{-1})^{-1} \mathbf{y},
\]

\[
\mathbf{d} = \mathbf{v} (\mathbf{I} - \mathbf{x}^{-1} \mathbf{Z})^{-1} (\mathbf{x}^{-1} \mathbf{e}),
\]

where lowercase are vectors, uppercase are matrices, vectors are in column format, ‘ is transpose and ‘ is diagonal matrix. The source data for the calculations are: \( \mathbf{Z} \), the matrix of inter-sectoral transactions; \( \mathbf{y} \), the final demand vector; \( \mathbf{v} \), the vector of added value; and \( \mathbf{x} \), the vector of total output.

The derivation of Eqs. (1) and (2) can be found in Rodrigues et al. (2010), under an attributional interpretation, and in Oosterhaven (1996), under a consequential interpretation. The term \((\mathbf{I} - \mathbf{z} \mathbf{x}^{-1})^{-1}\) is the Leontief inverse (Leontief, 1936) and the term \((\mathbf{I} - \mathbf{x}^{-1} \mathbf{Z})^{-1}\) is the Ghosh inverse (Ghosh, 1958).

In an MRIO a country is defined as a set of sectors. So the production/consumption/income-based responsibility of a country is, respectively, the sum of \( \mathbf{u}_i \), \( \mathbf{d}_i \) and \( \mathbf{a}_i \) over the appropriate set of sectors \( i \). Both upstream and downstream emissions are conserved quantities, which means that the consumption-based and the income-based responsibility of the world are both equal to the production-based responsibility of the world, \( \sum \mathbf{u}_i = \sum \mathbf{d}_i = \sum \mathbf{a}_i \). This implies that the choice of the responsibility principle can be interpreted as an allocation problem in which a pie (the world total emissions) can be cut into different slices (the responsibility of each region).

6.2. Results

The responsibility of each region according to each principle is reported in Table 2. In order to facilitate the subsequent discussion we selected 15 representative regions for which we present figures. These regions are individual countries and two aggregate EU regions: EUR-15 (Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom) and EUR-10 (Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia). The latter is the group of countries that joined the European Union in 2004.

Fig. 2 presents the difference between consumption and production-based responsibility and the difference between income and production-based responsibility, as a percentage of production-based responsibility, for the selected regions. This figure illustrates the effect of the choice of the responsibility metric in the apparent contribution of a country to global emissions. The income-based responsibility of Norway (NOR) is 2.7 times higher than its production-based responsibility while its consumption-based responsibility is slightly smaller than its production-based responsibility. Venezuela (VEN), Australia (AUS), Russia (RUS) and Brazil (BRA) display the same pattern. The consumption-based and income-based responsibilities of Switzerland (CHE) are 2.7 and 2.2 times higher than its production-based responsibility, respectively. For Singapore (SGP), EU-15 and Japan (JPN) the relation between consumption and income-based responsibilities and production-based responsibility is the same as for Switzerland. South Africa (ZAF), China (CHN), EU-10 and India (IND) have lower values of consumption and income-based responsibility than production-based responsibility.

Fig. 3 displays the per capita producer, consumer and income-based responsibility of the selected regions. Per capita production-based values tell us how many tons of CO2 are generated inside the country’s border per inhabitant. Per capita consumption-based values indicate the emissions that are required to satisfy the final demand of goods and services of a citizen of that country. Finally, income-based values represent the emissions that are required to generate the average income of that country.

We see that citizens of wealthier economies are, on average, responsible for more CO2 emissions than citizens from least developed economies. We also see that in wealthier regions per capita consumer responsibility is typically higher than producer responsibility. This observation indicates that, in these regions, the (upstream) carbon embodied in imports exceed the (upstream) carbon embodied in exports. This phenomenon is particularly striking in small open economies, such as Singapore (SGP) or Switzerland (CHE), who rely strongly on international trade. On the other hand, for bigger economies, like United States or China (CHN), the indirect measures of responsibility are closer to the direct measure due to the predominance of the domestic effects. This occurs because consumer and income-based responsibilities are carbon trade balances (Kanemoto et al., 2012; Rodrigues et al., 2010; Serrano and Dietzenbacher, 2010), and in larger economies domestic trade outweighs international trade.

Fig. 4 shows the different responsibility metrics per unit of GDP. We find that developing economies have higher carbon intensities than more developed economies. The relations between producer,
consumer and income responsibilities are maintained. In the case of rich countries the differences are smoothed (for example, NOR). In the case of not so wealthy countries differences are sharpened (for example, VEN). Brazil (BRA) appears amongst more developed economies, an exception to this pattern. This may be explained by the fact that this country relies heavily on hydro power and ethanol. In 2004, the fraction of domestic electricity generated from hydro power was 82.8%, only surpassed by Norway (NOR) with 98.8% (IEA, 2006). These results show that developed economies are more carbon efficient, generating per each dollar GDP less emissions than developing economies.

With these results we see that whether the analysis is done for absolute, per capita or per GDP values, there is a group of countries that consistently displays higher income-based responsibility. These are Norway (NOR), Venezuela (VEN), Australia (AUS) and Russia (RUS). A similarity between all these countries is the fact that they are all fossil fuel exporters. Thus, under the income-based perspective, we see that countries which earn income from the export of fossil fuels (or other goods that enable emissions to occur downstream along the supply chain) are held responsible for the emissions generated abroad when the fossil fuel is burnt. Interestingly, using this indicator we find that Norway, a country whose electricity is obtained mostly from hydropower and which has an otherwise very ‘clean’ economy (Peters and Hertwich, 2006; Yamakawa and Peters, 2011), has the highest per-capita income-based GHG responsibility in the world.

7. Discussion and Conclusion

The aim of this work was to clarify and illustrate the information provided by the concept of income-based responsibility, i.e., the emissions enabled downstream by primary factors of production.

We suggest the use of income-based responsibility as a novel nomenclature to replace downstream responsibility, as a more intuitive term which also has a clear symmetry with the more popular metric of consumption-based responsibility. Thus, while consumption-based responsibility provides information on how GHG emissions are required to generate final demand, income-based responsibility provides information on how GHG emissions are enabled to generate income.

To illustrate this metric we present empirical values of income-based responsibility and compare them to direct emissions (the metric used in the Kyoto protocol and normally referred to as producer responsibility) and the alternative metric of consumer-based responsibility. We find that, for some countries, different metrics yield very different responsibility values, while for other countries the responsibility is very similar, irrespective of the principle chosen. Unsurprisingly, the
choice of metric is more important for smaller and more open economies than for bigger economies.

The most emblematic case is that of Norway. This is a rich nation that is considered one of the ‘cleanest’ countries in the world, with low producer and consumer responsibilities. However, its income is generated at the expense of enabling large CO2 emissions abroad through the export of fossil fuels. We have shown that income-based responsibility is able to quantify this type of situation, while the alternative production or consumption-based indicators cannot.

In climate policy, the analysis of a country’s performance from the point of view of all responsibility principles provides a more integrated understanding on how carbon emissions can be mitigated. Under a production-based perspective, mitigation options are circumscribed by domestic borders. Under a consumer-based perspective, importing countries can improve their environmental performance through the selection of clean suppliers from abroad. Under an income-based perspective, the symmetric possibility is open to exporting countries. For example, under the income-based perspective Norway can increase its environmental performance by deciding to sell fossil fuels only to countries with carbon efficient production chains.

Another application for income-based responsibility could be its combination with consumption-based responsibility in a single metric of shared responsibility (Gallego and Lenzen, 2005; Lenzen et al., 2007; Rodrigues et al., 2006).

Acknowledgements

We would like to thank the financial support of FCT through grant PTDC/AMB/64762/2006 and funding Pest-OE/EEI/LA0009/2011, FCT and the MIT Portugal Program through scholarship SFRH/BD/42491/2007 (to AM).

References


Chapter 6:

Jisun Jung and Hugo Horta. "Higher Education Research in Asia: A Publication and Co-Publication Analysis".
Higher Education Research in Asia: A Publication and Co-Publication Analysis

Jisun Jung (1) and Hugo Horta (2)*

A version of this chapter was accepted for publication in Higher Education Quarterly

Abstract

This study explores higher education research in Asia. Drawing on scientometrics, mapping of science and social network analysis, we examine the publications of 38 specialized journals on higher education over the past three decades. Our findings indicate a growing number of higher education research publications but a latent share of the worldwide higher education research as a specialized field since the 2000s. The higher education research community in Asia is heavily concentrated in a few countries and universities, resting on a relatively small number of core scholars who publish research in the international specialized higher education journals. In response to increasing challenges in Asian higher education systems, we suggest that the higher education research community in Asia needs to be expanded and include more regional and international collaborations.

1. Faculty of Education, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

2. Center for Innovation, Technology and Policy Research (IN+/IST), Instituto Superior Técnico, Technical University of Lisbon, Portugal.

* corresponding author, E-mail: hugo.horta@ist.utl.pt
Introduction

Higher education research has been expanding rapidly since the last quarter of the 20th century, and over the past few decades, this rate has increased even more (Teichler, 1996; Tight, 2004). This expansion was mainly triggered by the massification of higher education systems and related challenges this process brought to governments, university managers and other stakeholders (Trow, 2005). As higher education systems worldwide evolved and became more complex and globalized, a myriad of themes associated with novel challenges emerged. These are related to accountability, accreditation, governance, funding, diversification, internationalization and career and employability issues just to cite a few. More recently, themes associated with country and university global competitiveness, academic capitalism and talent wars involving the notions of world-class university, league tables and rankings, and internationalization have also brought about new research interests within higher education. All of these challenges have highlighted a number of higher education-related issues, from both theoretical and practical perspectives, at international, regional and national levels.

In this context, an increasing amount of research on higher education has ensued, which is reflected in the increasing number of academic degree programs, journals and publications, conferences, and evaluative policy and research reports that specialize in higher education themes (Altbach et al., 2006; Hutchinson and Lovell, 2004). In association, there has also been an increasing demand for expertise and appropriate data to achieve a more sophisticated understanding of the nature of higher education (Altbach et al., 2006). In this environment, supra-national institutions, such as the OECD and the United Nations, have compiled country data on higher education indicators as well as seeking to influence policymaking at regional and country levels (Huisman et al., 2007). However, there are clear differences between regions and
countries in terms of higher education research. For example, some countries have only basic statistics, rudimentary overview reports and incipient research while other countries have agencies that gather systematic information on higher education and have established academic research groups to analyze it (e.g. Schwarz and Teichler, 2000).

The study of higher education is closely related to the size, scale and substantive growth of the higher education system in each region of the world (Tight, 2007). Studies on higher education initially developed as a significant academic field in the United States, and this country’s body of research on higher education is larger and older than that of others (Tight, 2007). Besides having the largest research enterprise in higher education, the United States also benefits from well-developed training programs for researchers and practitioners (Altbach et al., 2006). Other Anglo-Saxon countries, whilst benefiting from English as the lingua franca of science, also have a well-developed body of research on higher education. This has contributed to the incorporation of Anglo-Saxon academic traditions into higher education studies, influencing higher education research worldwide.

As the massification of higher education gradually occurred in other countries, this led to more diverse research in higher education being undertaken. Tight (2004) has shown that European academics have a diverse array of research interests and methods in higher education research, differentiating these countries from the research performed in the United States. In addition, as higher education systems have developed worldwide, higher education courses have been added to the curriculum of several disciplinary areas and research institutes focused on higher education have been established in several places (Jones, 2012; Altbach et al., 2006).

While the evolution of higher education research has been extensively analyzed in North America, most notably in the United States, and in Europe, little attention has been given to the
evolution of higher education research in Asia. This is possibly explained by the only recent, but accelerated, expansion of higher education and research productivity in some Asian countries, especially in China (Postiglione and Jung, 2013; Chen and Hu, 2012). Currently, an analysis of higher education research in Asia is more relevant than ever, as the Asian Pacific Higher Education Area is in the making (Chao, 2011). This requires Asian countries to know more about one another, as they are in different stages of maturity, and thus face both similar and dissimilar challenges (World Bank, 2000). Therefore, the purpose of this article is to contribute to the understanding of the evolution of higher education research in Asia. We ground our analysis on literature, using Tight’s (2004) definition of higher education community, combining it with scientometrics and the mapping of science methods, to offer a richer analytical framework. In this context, we have examined articles published in specialized journals on higher education from 1980 to the present.

**Higher education as a field of study**

Most analyses of higher education research query the extent to which higher education is a mature and independent academic discipline. More than often, higher education research is considered immature, and little evidence exists defining it as a ‘discipline’ on its own right. Rather, it remains exclusive to a number of people with interest in the subject (Pierce, 1991). Others describe higher education research as enjoying only ‘part-time involvement’ from scholars who have their own disciplinary theoretical perspectives that are applied to the specific issue of higher education (May, 1997). Instead of ‘discipline’, higher education is described by scholars as a ‘field’ (Altbach et al., 2006). Tight (2004) reinforced this idea by stating that higher education departments or centres could be called ‘fields of study’ and ‘practice’, rather than
disciplinary poles. The term ‘field’ is a geographical metaphor used to describe categorized and classified knowledge and skills (Burke, 2000). Fields are constructed over time with ‘accumulated knowledge, paradigms and academic capitals, recognized by academics or other social groups’ (Chen and Hu, 2012, p. 656). As a field of study, higher education research is represented by academic degree programs at the post-graduate level, higher education institutes and centres, and by specialized journals (Altbach et al., 2006).

The term ‘academic field’ is used loosely and its boundaries recognized by a group of people operating as a community who identify themselves with it and work within it (Chen and Hu, 2012). However, higher education is an object-focused area based on a multidisciplinary approach since researchers from several disciplines often deal with topics related to the enterprise of higher education (Altbach et al., 2006). Higher education research is also characterized by its blurred distinction between practitioners and researchers, raising tension between scientific reasoning and professional problem solving (Harland, 2012; Teichler, 1996). This frees higher education researchers from making their theoretical perspectives explicit or become involved in the broader aspects of the theoretical debate. In this framework, Tight concluded that ‘higher education research is not a single community of practice but, rather, a series of, somewhat overlapping, communities of practice’ (2004, p. 409).

In this milieu, the field of higher education research was built through the establishment of associations and degree programs but mainly through the organization of conferences and seminars, the publication of books and the setting-up of specialized journals. The number of higher education journals has expanded over the past three decades and many now circulate internationally (Budd, 1990; Altbach et al., 2006). Academic journals, particularly international ones, encourage the widespread and efficient communication of ideas, stimulate discussion and
allow scholars to share their findings (Hutchinson and Lovell, 2004). Through scientometrics methods, they permit the identification of research patterns geographically and over time, as they provide an unobtrusive, non-reactive measure of formal communication practices within a scholarly discipline (Smith, 1981).

Our study is not the first to examine the field of higher education using these methods. Until recently, most analyses have focused on US higher education journals. For instance, Silverman (1987) classified the main research topics in higher education through an extensive analysis of 1,103 articles. Budd (1990) observed the characteristics of the research literature, its major contributions on subsequent research and the influence of journals and authors on the field. Milam Jr.’s (1991) analysis of five higher education journals concluded that higher education researchers need to broaden their approach to include alternative paradigms for interpreting higher education phenomena. Several studies have performed content analyses of articles in higher education journals (e.g. Hutchinson and Lovell, 2004). Others have identified a shift in the research from a focus on individual countries to a comparative and international focus (Teichler, 1996). Selecting 17 specialized higher education journals from outside North America, Tight (2004, 2008) analysed themes, methodologies, levels of analysis, location and other identifiable author characteristics. Similarly, Ross (1992) examined the journal *Higher Education* in terms of its authors’ locations, changes in main research topics and citations while Huisman (2008) analytically described the contributors (and their affiliations) of *Higher Education Policy*. 
Higher education research in Asia: brief literature overview

Simon Marginson (2011), quoting Times Higher Education, stated that ‘higher education systems in North America and Western Europe are watching the emerging Asian systems with a mix of excitement and apprehension’ (p.588). The Asia-Pacific region has the fastest-growing higher education market in the world, sustained by rapid economic growth (Shin and Harman, 2009). According to UNESCO annual statistics, enrolment in tertiary education in East Asia and the Pacific was approximately 47 million in 2007, three times the enrolment recorded for 1991, at 14 million. The enrolment growth rate in South and West Asia is also swelling with 20 million students currently enrolled. This is higher than in any other region worldwide.

Massification of higher education in Asia, due to its high enrolment scale, has transformed the entire regional higher education landscape, raising issues of governance, financing, quality, curricula, faculty, and student demographics. Based on these issues, Shin and Harman (2009) identified major challenges for the 21st century Asia-Pacific higher education system, including further massification, privatization, accountability and governance, internationalization and ranking and world-class universities. Postiglione (2011), pointing out the effects of the global recession on higher education in East and Southeast Asia, underlined the need to continue reforming governance and administration, access and equity, internal and external efficiency and regional collaboration. This requires an overarching effort from higher education research, taking into account that the development of higher education itself is complemented by an increase in interest within the academic community, which may serve as a stimulus for enhancing the creation of a community with a common identity (Teichler, 1996).

In this process, the importance of fostering an international higher education research community, more than a nationally focused community is critical, as the higher education
challenges faced in Asia are mostly shared regionally and with other countries around the world, and require mutual learning. Several scholars have argued that greater effort is needed to consolidate and strengthen higher education research in Asia with an international focus (e.g. Chen and Hu, 2012). Arimoto (2000) proposed this as the right time for Asian higher education scholars to reach out to the worldwide network of higher education researchers. He reinforced this need by assuming the minor participation of scholars in higher education research and by describing their integration in the international higher education community as limited (Arimoto, 2000). The key underlining question of Arimoto’s work is as follows: ‘Does a higher education community exist in Asia?’

Several internationally recognized higher education institutes exist in Asia, such as the Research Institute for Higher Education established in 1972 at Hiroshima University and the Research Centre of Higher Education Development established in 1978 at Xiamen University. Besides these, international organizations like the Asian Development Bank, have continuously contributed to higher education research by collecting and analysing macro-data. Still, the institutionalization of higher education research in Asia continues to be described as internationally underdeveloped (Chen and Hu, 2012; Arimoto, 2000) and the question concerning the existence of an Asian higher education research community that is internationally visible remains unanswered.

**Methodology**

This study draws on the Scopus database, the largest abstract and citation database of peer-reviewed research literature increasingly used in scientometrics studies focusing on the social sciences and related fields of knowledge (Norris and Oppenheim, 2007). Data were
gathered during October 2012. We selected all articles published from 1980 to October 2012, having an institution affiliation based in Asia, from journals containing the words ‘higher education’ or ‘tertiary education’ in the title. This amounted to 514 articles published in 38 journals, identified in Appendix 1. In doing this, we followed Tight’s (2003; 2008) methods in choosing journals focused exclusively on higher education research. These journals consist of core identifiable journals that are part of and define the higher education research community (see Bayer, 1983).

The choice of articles with an Asian affiliation is clear, as the study focuses on higher education research in Asia, but a definition of the Asian region is required. For this purpose, we used the United Nations classification that includes countries located in Eastern, Southern, South-Eastern and Western Asia (see Appendix 1). Some countries listed as belonging to Western Asia, such as Israel or Cyprus, were not included in the analysis, as they are member or associate states of the European Union, strongly engaged with the European Research and Higher Education Areas and have access to the Framework Program and other European research funds. Their inclusion would bias the findings vis a vis other Asian countries. Based on this data we have undertaken trend analysis, but also apply mapping of science methodologies following the work by Bornmann et al. (2011). We further complement the scientometrics analysis by performing social network analysis.

As is appropriate in order to perform social network analyses, a collaboration matrix for each article was created, grounded on institutional affiliations and displayed as a square array of measurements. The data were set as a one-mode affiliation network based on article co-authorship coded further to provide country information. The gathered publication data covered a total of 369 institutional affiliations, 245 located in Asia and 124 located in other regions of the
world. This allows the analysis of co-authorship patterns among institutions but also provides information about within and outside country collaborations. Since the data is based on all the collaborations that Asian institutions have among themselves and others in the field of higher education research, we have a complete relational overview of the population and apply full network methods. This is important because it allows the study to sustain the measurement of key structural network analysis concepts related to centrality measurements. Overall, this is significant to better understand the centrality of the most productive universities. The social network analysis is performed through the use of UCINET (Borgatti et al., 2002).

**Scientometrics analysis of higher education in Asia**

**The evolution of higher education research across world regions since the 1980s**

The number of publications in higher education journals has been increasing worldwide, marked by two periods of growth after a latent period of publications during the 1980s (Figure 1). The first period of growth began in the early 1990s followed by a second period of exponential growth starting in the early 2000s. These periods reflect the growing interest in higher education themes but also the increasing number of specialized higher education journals emerging in the past several decades. Figure 1 shows that these periods were driven mainly by higher education researchers based in North/Central America, essentially the United States, and in Europe. These regions of the world dominate higher education research, the only difference being that more publications are currently being produced by researchers in Europe than in the United States.

The dominance of Europe and North America can be understood by the fact that higher education research has greater maturity and a critical mass of researchers, associated with earlier periods of higher education massification. The massification of higher education in the United
States has been in full swing since the 1960s (Gumport et al., 1997), which explains the dominance of the US in the 1980s higher education research in terms of publications. The massification process led to earlier significant reforms and structural changes (and new challenges) in these higher education systems, further boosting higher education research. In this sense, it is likely that much of the second period growth in publications by European-based researchers is related to the Bologna process. Another reason for the sustained growth and worldwide prevalence of higher education research in both North/Central America and Europe is that the majority of higher education journals are based there and most chief editors are from these regions (Tight, 2003).

In the 2000s, a growing number of publications from the Oceania could be noted, mostly driven by Australian-based researchers, and less substantial growth of higher education research in Asia could be observed. The number of publications in the field by researchers based in Asia more than tripled between the 1997–2001 and 2007–2011 quinquennia, but in relative terms the research intensity of Asian-based research continues to evolve very slowly and is not overly different in 2007–2011 than it was in 1997–2001 (between five and seven percent of the world’s higher education research). This indicates that higher education research in Asia is still incipient, not yet having reached the level of growth that other world regions have already reached. In explaining this trend, one should not ignore the language issue and the known difficulties of non-native English speakers to publish in English, the lingua franca of modern science. This is compounded in the field of social science, as countries have a traditional practice of publishing in national scientific outlets in their own languages. Also, at a national level dynamic and productive higher education research communities might be present that do not publish internationally. However, state-of-the-art knowledge flows in the sphere of international research
networks and publications, and it is mostly through them that national policies are informed. In modern science, higher education research communities need to be considered as those that contribute to the global knowledge pool of the field (Arimoto, 2000).

**FIGURE 1 HERE**

**Higher education research output in Asia by country: the broader picture**

When analysing higher education research in Asia, several differences can be observed between countries. Table 1 shows countries that account for at least 10 publications during the period 1980-2012. These countries account for 90% of all articles published in Asia in the specialized literature, but even amongst them there are substantial differences. Researchers based in Hong Kong publish almost twice as many articles than those based in Japan, which ranks second in terms of the countries with greatest production, and over five times more articles than South Korea, ranking eight. The difference from Japan, ranking second to Malaysia, ranking third, is almost twice as well. This suggests that there are different stages of development of higher education research communities in Asia, country-by-country, with higher education research being much more developed in some countries than in others. Differences are also perceived in the authorship patterns. Researchers in Hong Kong, Singapore, China and Taiwan lean more towards a more collaborative research effort while those in Japan and India show a balance between single-authored and co-authored articles. Still, the overall high percentage of single-authored publications (43%) suggests a substantial degree of isolation of Asian higher education scholars, especially when compared with collaboration rates in other areas of the social sciences (Larivière et al, 2006).
Co-publication patterns of Asian higher education research: not regionally integrated but internationally linked

The co-authorship analysis does not provide evidence of any clear collaboration trend among authors working in different Asian countries. A more complex picture emerges from the data. In some countries domestic co-authorship is dominant, such as Saudi Arabia, Iran, Hong Kong and India. Scholars in a number of Asian countries rely mostly on international collaborations, notably South Korea, China and Singapore, while in Japan and Malaysia there is more of a balance between domestic co-authorship and publication with international peers. However, the analysis of international co-authorship as a whole indicates that collaboration within Asian countries is low when compared with the collaboration with other regions outside Asia, suggesting that higher education research in Asia is not regionally integrated. This points to an absence of a solid Asian higher education researcher network. However, it also indicates that Asian scholars already have a degree of international connections with the rest of the world, even if, as a whole, it is limited, as suggested by Arimoto (2000).

To better understand the collaboration of Asian scholars with others based in other regions of the world, a social network analysis was carried out. Figure 2 shows that most Asian scholars’ collaborations in the field are conducted with peers based in English-speaking countries. This is to be expected due to the maturity of the field in these countries. Still, reliance on co-authorship with peers in these contexts is striking. The degree of network centrality of the US (#27=14) is greater than of any of the Asian countries. That of Australia (#30=12) is similar to
the levels of most central Asian countries in terms of the number of collaborations (Hong Kong: #1=13; Japan: #3=11). The US also has the greatest betweenness centrality (182) of the network, meaning that this country has a greater role in linking Asian countries’ higher education scholars than any other Asian country. Australia has the fourth highest betweenness centrality (106), highlighting the role of this country as well in linking the Asian higher education research community. Such reliance on the US and Australia may be associated with the fact that these countries are where Asian higher education scholars have most frequently undertaken their studies, as the flow of Asian students to these countries is well documented. Interestingly, Hong Kong, with the second highest betweenness centrality (160), has a large international community of scholars and research that is strongly influenced by Anglo-Saxon countries. Japan has the third highest (128).

FIGURE 2 HERE

Engagement and participation in Asian higher education research: concentrated and limited

In order to obtain a more in-depth analysis, we performed an analysis focused on institutional affiliations. Figure 3 shows the geographical distribution of higher education research in Asia during the period of 1980–2012. Our mapping is based on all publications combining countries with institutions (overwhelmingly universities). The analysis shows a skewed distribution of universities (and countries) actively engaged in higher education research.

As Figure 3 suggests, it is in Eastern Asia, comprising China, South Korea, Japan, Hong Kong and Macao, that universities are more active in higher education research. The only
exceptions are Mongolia and North Korea, with no publications. Eastern Asia accounts for half of the total universities in Asia working on higher education research, suggesting that the higher education research community is highly concentrated in Eastern Asia. This is expected, as this is the Asian region with the highest cumulative investment in research in development and it holds the majority of mature higher education systems. South Eastern Asian countries, mainly through the number of Malaysian universities, have 21% of the total Asian institutions involved in higher education research while Southern Asia, mainly due to Indian and Iranian universities, accounts for 19%. Western Asia has less institutional engagement in higher education research. No affiliations based in Central Asia, which includes countries such as Kyrgyzstan and Turkmenistan, were identified.

FIGURE 3 HERE

At an incipient stage of the evolution of higher education research in Asia its community is thus concentrated in Eastern Asia, both in terms of publications (accounting for at least for 60% of all articles published in the region; Table 1) and institutions (Figure 3). The analysis also indicates a diminished regional collaboration focus. These findings raise questions regarding the intensity of this community as well. From 1980 to 2012, the analysis shows that 66% of the 244 institutions engaged in higher education research only published one article and 15% published two articles. Therefore, 81% of the institutions still have a low level engagement in higher education research and it can be argued that their infrequent contribution to the higher education community indicates that the field in Asia is smaller than one might realize at first. Additionally, 36% of institutions with one publication are single-authored publications, indicating that these
publications were more a single occurrence than a frequent engagement in higher education research. In fact, only nine institutions (3% of all institutions) have published at least 10 articles in higher education journals since 1980, reinforcing the idea that the Asian higher education community is at a more embryonic stage than the country analysis suggests.

**Looking at those producing the most: few institutions on the shoulders of fewer scholars**

Figure 4 illustrates our analysis of the whole co-authorship network of Asian institutions concerning higher education research, focusing the analysis on the nine institutions that have published at least 10 articles in the past three decades. These are located in the upper left side in the figure. Regional clusters are identified for North America, Europe Oceania and Africa (Institution #348). No collaborations were found with South America.

Not surprisingly, the top five institutions in terms of publications are universities based in Hong Kong since this Special Administrative Region of China also accounts for the largest number of publications. Two universities based in Singapore, one located in Japan and the other in Taiwan, complete the group. Also expected is the top five institutions’ degree of centrality in the network. The University of Hong Kong (#21), while not having the largest number of publications, is a key institution to higher education research in Asia. It has the greatest number of ties (co-authorships) and a degree centrality of 18. This institution also has the highest betweenness centrality (2722), thus indicating its critical role in linking other institutions in Asia, influencing their collaboration and possessing a greater ability to foster regional alliances.

The other institutions that have published at least 10 articles in the past three decades also have high levels of network centrality. Still, some variations are found in terms of the degree of centrality [ranging from 10 for The Hong Kong Polytechnic University (#48) to 3 for Hiroshima
University (#54)], and betweenness centrality [ranging from 1655 for The Hong Kong Polytechnic University (#48) to 360 for the Chinese University of Hong Kong (#44)]. The only exception is the National Taiwan University (#16), in which higher education publications were developed without ties to other institutions. This highlights the existence of two types of higher education research communities in Asia. The first case concerns institutions that focus their publication efforts on co-authorship across institutions. The second case refers to institutions focusing on co-authorship within their institutions only. The latter type entails some degree of isolation but also the possibility that the sum of the institutional publications result from several authors producing single or infrequent publications within the field. This may relate to arguments made by May (1997) and Pierce (1991), underlining the importance of looking at the scholars producing higher education at these universities.

FIGURE 4 HERE

An analysis of the institutions’ publications by author revealed that the universities’ publication rapport depends on a few key scholars within these universities. Faculty in the most highly productive universities in higher education research in Asia, the Polytechnic University of Hong Kong and the Chinese University of Hong Kong, have published 27 and 23 articles, respectively, since 1980. Among these, eight (30%) and 11 (41%) articles were co-authored by a single scholar, suggesting that these universities owe, to a large extent, their higher education publications to one scholar. Moreover, this scholar is the same in both universities and the dependence of the institutional higher education research rapport on this single scholar is easily discernible. Since this scholar moved from the Polytechnic University of Hong Kong to the
Chinese University of Hong Kong, the latter university has been credited with an increasing number of higher education articles, while in the former has been credited with fewer. This is observed in other universities as well, including the National Taiwan University and the Hiroshima University where 60% and 39% of the university publications were authored by a single scholar. This indicates that a significant proportion of Asian higher education research rests on a few individual scholars rather than on a few universities.

A slightly different situation is identified at other universities. The University of Hong Kong and the National University of Singapore, for example, have a wider range of scholars publishing in higher education. They rely substantially on international co-authorship and collaboration. However, most of these scholars have published less than three articles in the last three decades, highlighting either a lower level of productivity or infrequent productivity, thus underlining the fragility of the higher education research community in Asia.

**Discussion and conclusion**

As the relevance of higher education increases in Asia, its higher education research community is still at an emergent stage of development when compared to other regions worldwide. The analysis shows that in spite of the growing number of publications produced by this community, its relative contribution to the world pool of knowledge in this field has remained stable. This raises concerns regarding the condition of the field in Asia, posing the question: “Does a higher education community exist in Asia?” Our findings indicate that this community does exist. It is mostly concentrated within a few countries and universities, resting on a relatively small number of scholars who publish frequently in the international specialized higher education literature. This is similar to the characterization of the European higher education literature.
education research community in the 1990s and can be considered part of the development process of the field (Teichler 1996).

The Asian higher education research community is concentrated in Eastern Asia, the region which also holds the majority of post-massification higher education systems. Within this region, Hong Kong universities play a pivotal role not only in terms of their number of publications but also in fostering regional collaborations. The fact that the Asian higher education research community is concentrated in a region with these characteristics is similar to other regions of the world. Higher education research communities are larger in post-massified higher education systems. These regions also have in common a greater cumulative investment in research and development that has fostered the consolidation of higher education research communities. The growing critical mass of academics and researchers also permits greater disciplinary research specialization and thus more scholars focused on higher education research. Additionally, massification and post-massification processes bring about a complex and varied range of challenges, raising awareness of governments and universities for the need for establishing (or at least promoting) a higher education community.

This scenario still seems to be underway in Asia (Chen and Hu, 2012), in particular when certain countries like China are rapidly developing their higher education systems and Asian countries are attempting to integrate their higher education systems at a regional level to create an Asian higher education area (Chao, 2011). This is even more important when the analysis has shown that countries are in different stages of developing higher education research. In order to provide a much stronger contribution to the field, the higher education research community in Asia needs to be strongly reinforced and further articulated. Our analysis has shown that regional collaboration among scholars from different Asian countries is sparse and that the main
‘promoter’ of higher education collaboration among Asian countries is the US, with Australia playing a relevant role as well.

The fostering of these linkages by the US and Australia is positive, mainly because the institutional analysis shows that the collaboration is carried out with many US and Australian universities. This presupposes a variety of knowledge sources and flows that enriches higher education research in Asia. However, at the same time, it demonstrates that a cohesive Asian higher education community is not present and this community needs to be fomented. A cohesive community draws on frequent collaborations and a sense of belonging, which are key factors to consolidate the current higher education research community. Teichler (1996) noted that many European higher education researchers consider colleagues all over Europe to constitute their academic community. The focus is not countries or institutions but on the regional and international higher education research community. This becomes even more critical when our findings indicate that a substantial number of contributions to the field tend to be single contributions rather than frequent sets of contributions, which better define a consolidated research community. Several of the contributions are single-authored, underlining a lack of articulation with other scholars interested in higher education research. The findings indicate few truly active universities engaged in higher education research. More striking is the fact that the higher education rapport of these universities is highly dependent on a few scholars who arguably represent the core of higher education research in Asia.

In order for this core higher education research community to grow and better inform policymakers in Asia at a dynamic period of growth but also of multiple challenges, public policies and funding promoting the development of the field are required. However, this should not be only a government effort; it should also be sustained by the universities through the
creation of positions for higher education research scholars. An important lesson drawn from the evolution of the higher education research in the US is that universities in this country understood the relevance of fostering higher education research associated with the development of higher education educational programs (Altbach et al., 2006). Such programs could be an effective way to sustain higher education research at Asian universities, contributing decisively to the development of a higher education research community. They could also lead to the creation of a professional body of policymakers and university managers informed by the most up-to-date knowledge in the field, which could contribute decisively to the further development of Asian higher education systems and institutions.

We conclude the article by pointing out several limitations to our study that need to be considered in future research. First, certain types of publications, such as books, were not included in the analysis. Second, we acknowledge the existence of national higher education research outlets and education journals focused on Asia (e.g. Asia Pacific Education Review) where higher education-related articles can be found. These were omitted, as they did not fit with our methodological referral (Tight 2004) but will be pursued in future studies, which will include a study on higher education research themes.

References


Figure 1 - Evolution of higher education research, 1980-2011

a) Overall number of publications

b) Publications by region

- North/central America
- South America
- Europe
- Africa
- Asia
- Oceania
Figure 2 – Co-authorship network in higher education research of scholars based in Asian countries, 1980-2012
Figure 3 - Higher education research in Asia in 1980-2012

Figure 4 - Higher education research network in Asia
Appendix 1. Journal list:


Source: Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings

http://unstats.un.org/unsd/methods/m49/m49regin.htm

Central Asian countries: Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan

Eastern Asian countries: China, Hong Kong, Macao, Japan, South Korea, North Korea, Mongolia, Taiwan

Southern Asian countries: Afghanistan, Bangladesh, Buthan, India, Iran, Maldives, Nepal, Pakistan, Sri Lanka

South-Eastern Asian countries: Brunei Darussalam, Cambodia, Indonesia, Laos, Malaysia, Myanmar, Philippines, Singapore, Thailand, East Timor, Viet Nam

Western Asian countries: Armenia, Azerbaijan, Bahrain, Georgia, Iraq, Jordan, Kuwait, Lebanon, Occupied Palestinian Territory, Oman, Qatar, Saudi Arabia, Syria, United Arab Emirates, Yemen.
Chapter 7:

Technology Adjacency: Using Science to Create New Markets?

Daniela S. Couto, Luis Perez-Breva, Charles L. Cooney and Manuel Heitor

A version of this chapter was submitted for publication in Industrial and Corporate Change

Abstract

Through the study of drug-device and biological-device combination products, we identify a pattern of technology innovation used by science-based firms to expand to distinct markets, creating new and uncontested market spaces. We name this new pattern of technological innovation as “technology adjacency”. Evidence from our work also suggests that this process strongly depends on collaborative strategies between different stakeholders, in particularly physicians. This is supported by detailed case studies associated with technologies to restore blood vessels function and complemented with a case study of Boston Scientific Corporation in order to facilitate our understanding of the phenomenon of technology adjacency and its impact on corporate change and growth.

We use the current theoretical understanding of technological innovation and of knowledge management to discuss the dynamics of technology adjacency. We argue that science-based firms learn by using technology adjacency to integrate diversified technological features, typical of distinct markets, into new products, leading to the creation of new market spaces. An example are drug-eluting stents, in which medical devices firms added features from the drug space into their stents, created the drug-device combination space and expanded their potential market. Our analysis also suggests that science-based firms can use technology adjacency strategically to access new markets and penetrate into areas beyond their original area of expertise in a process of continuous adaptation to market opportunities and systematic research collaborations and stakeholder engagement. The success of technology adjacency depends on the role physicians in adopting new technologies. Physicians facilitate access to markets and are key elements in building a process of societal trust.

Technology adjacency is emerging in association with convergence of technologies in life sciences, physical sciences, and engineering. It requires new expertise and skilled experts to develop the combination of technologies arising from distinct areas of knowledge.
1. Introduction

This article introduces the concept of technology adjacency in corporate strategy and industrial dynamics and is of particular interest for science-based firms and start-ups. It builds on previous findings, as published by Couto et al. (2012) and explored in detail by Couto (2011), and examines detailed case studies in bioengineering and related business ventures. We examine the contemporary role of science-based firms drawing from conceptual approaches of knowledge management to economic growth. Learning as a process of knowledge accumulation is a key driving force behind growth. In addition, our analysis suggests that the functions that society commonly attributes to the university are beginning to be shared among a wide range of institutions in the context of the knowledge-based economies (e.g., Nowotny et al., 2001), with particular emphasis for science-based firms. They use technology adjacency to learn. Firms may integrate diversified technological features from distinct markets into new products, leading to the creation of new market spaces.

Figure 1 depicts the concept of adjacent vertices, angles, and markets. We extrapolate from graph theory the concept of adjacency. An adjacency relation in mathematics is any relation which is irreflexive and symmetric, Foldes (1994). In graph theory, an adjacent vertex $\alpha$ in a graph is a vertex that is connected to $\alpha$ by an edge. Neighborhood is an interesting mathematical concept intrinsically associated with the notion of adjacency. Using the same example, the neighborhood of $\alpha$ includes all the vertices adjacent to $\alpha$ and all the edges connecting such vertices.

![Figure 1: Graphical explanation of adjacency relation in (a) vertices, (b) angles, and (c) markets. (a) The graph represents a set of 5 vertices, $b$ and $c$ are adjacent vertices of $a$ and together with link 1 define the neighborhood of $a$. (b) The angles $\alpha$ and $\beta$ are adjacent angles because they share the same plane, a common vertex and a common side. (c) A firm operating in the market $a$, with $b$ as an adjacent market, may develop a new product combining features from $b$ in $a$, intersecting both markets. The potential market for the firm starts to be $a$, but expands to become $a$ and $b$ combined.](image-url)
In geometry, adjacent angles are two angles in a plane, which share a common vertex and a common side but do not overlap. In corporate strategy, we define adjacent markets as those sharing the same space, without intersecting. Medical devices and drugs are an example; they operate simultaneously in healthcare but do not traditionally intersect. Biomedical firms developing novel drug-device combination products created new market spaces at the intersection of medical devices and drugs, or adjacent markets.

Scientific convergence of distinct scientific areas and related technologies is the basis for recognizing opportunities at intersection of adjacent markets. Historically, convergence of multiple areas of expertise enabled firms to capture blockbuster markets with the development of drug-device products, namely drug-eluting stents (Ratner, 2007). However, this seems to be the leading example. Several other collaborations set to combine complementary expertise failed to reach clinical practice. For example, Medtronic and Genzyme started a joint venture in 2004 - MG Biotherapeutics - to develop new therapies to treat ischemic heart failure (Medtronic, 2004a). They co-sponsored a clinical trial to test autologous cell therapies from Genzyme and intended to use a catheter developed by Medtronic to deliver the cells directly to the heart (Medtronic, 2004b; Genzyme, 2007). However, the cells used generated major side effects and the phase II trial was prematurely terminated.

A distinct example in computer technology is the Apple’s Newton platform, which has been discussed in the technical literature as a typical example of scientific convergence in telecommunications (e.g. Conceição et al., 2001). The Newton platform was the first tablet platform developed by Apple. John Sculley, Apple's CEO, defined it in 1993 as "the defining technology of the digital age. . . a focal point for the coming convergence of. . . computers, communications, and consumer electronic" (Newton, 1993). This was the year in which the MessagePad – the first system using Newton platform – was launched to the market. Convergence of computers, communications and consumer electronics became, in fact, a significant part of the telecommunications sector, but only years later with the introduction of the second generation of the tablet platform. The second-generation platform – iOS – was used in the iPhone, iPod Touch and iPad. The Apple’s Newton platform provides an example of a convergent technology that emerged before customers were prepared to adopt it. This example illustrates the critical role end users play in adopting new technologies.

Scientific convergence has been increasingly considered in many research fora. A recent MIT White Paper focuses on the emerging convergence of life sciences, physical sciences, and engineering as a critical step to foster improved forms of public health, environment, and economic prosperity (Sharp, 2011). It considers convergence as the next paradigm shift in technological innovation (see also, Hockfield, 2009) and uses multiple examples of research at
MIT to discuss their potential impact in health and economic prosperity. Such examples are of tissue engineering, microelectrical mechanical systems, biomaterials, gene and protein delivery, targeted medicine, drug-device combinations and high-throughput devices (Shmulewitz and Langer, 2006).

According to Conceição and Heitor (2003), convergence of scientific knowledge needs an extended learning period and further development to have a significant economic impact (e.g., Conceição and Heitor, 2002). Drug-eluting stents are the only notable exception (Ratner, 2007). They required a short learning period and rapidly reached the market. Tissue engineering is an example of scientific convergence that is requiring an extensive scientific learning period before advancing to market. In the early 1990s the field was expected to play a significant role in clinical development within a decade. Today, scientists recognize that it still requires substantial accumulation of new knowledge in order to deliver to its initial promise.

This paper makes strides in filling a blank space in connecting technological innovation with business strategy. We introduce the notion of technology adjacency as a strategic tool for science-based firms to identify new markets and create new opportunities for growth. In addition, we identify the interaction of scientific knowledge in the form of both ideas and skills as the key elements for the convergence of multiple scientific areas that, ultimately, enable the recognition of adjacency markets by firms.

The paper is organized as follows. We start by briefly exploring our theoretical background. We look at theories of economic growth and emerging conceptual approaches to knowledge management. We then examine the evolution of medical procedures and technologies to restore blood vessel function. These examples support our discussion about scientific convergence in enhancing the function of the standard of care. Procedures to improve blood function were enabled by the use of medical devices. Medical devices evolved by incorporating technical features from other adjacent market spaces, such as drugs and biologics. Drug-device combination products, whose function is enhanced with drug delivery, later replaced pure medical devices. Biological solutions are currently being developed with cells or other biologics enhancing the functionality of the combination product in repairing the vessel. Also, engineered blood vessels are being developed for severe cases that cannot be regenerated or repaired. This path of technological evolution was enabled by the scientific convergence of life sciences and engineering. In turn, it enabled firms to expand to adjacent markets and creating new market spaces within. Companies evolved from pure medical devices, to drug-device, followed by biological-device and tissue engineering.

To make the bridge from scientific convergence to technology adjacency and firm strategy, we examine in detail the case study of Boston Scientific Corporation. We assess the impact of
expanding to adjacent markets in firm growth, as well as the way the process is influenced by external factors, namely stakeholder’s engagement. We analyze growth due to penetration in drug-device market and compare with other firms in the sector. We also discuss the strategy of Boston Scientific in incorporating ideas and skills from other firms through collaborations and acquisitions. In addition, we illustrate with examples the role played by physicians in the overall process of technology adjacency and market penetration.

The paper concludes highlighting possible strategies for firms to capitalize on scientific convergence, identifying adjacent markets and creating new market spaces for firm growth.

While our work draws from examples particularly associated with combination products in biomedical applications, our analysis suggests that our findings are more general and broadly apply to next generations of combination products, as well as other technological sectors.

2. Industrial Evolution Through Technology Convergence and Technology Adjacency

Most of the studies that aim to explain the dynamics of industrial innovation are based on the PLC model, however it presents serious limitations: focus on single product and homogeneous industries. Here, we aim to show that convergence and adjacency are two concepts that occur across boundaries, mixing knowledge from distinct product markets. We will introduce adjacency as a strategic tool to explain knowledge and market convergence.

The dynamics and patterns of industrial innovation are traditionally explained by the Product Life Cycle (PLC) model, in which evolution of industries is described as a sequence of stages: an initial phase of higher diversity in product design and process flexibility, where the emphasis is on product performance; a transitional phase in which the emphasis is placed on process innovation and optimization (increasing production volume), rather than product; and final phase of lower growth and eventually decline, where the emphasis is on cost reduction and improved quality, with large-scale, automated, and highly-specific production (Abernathy and Utterback, 1978; Klepper 1997). While PLC model captures the underlying behavior of many industries, different authors have shown its inability to illustrate the behavior of others, such as biotechnology, laser industry (Klepper and Thompson, 2006) and semiconductors. These are hypercompetitive and heterogeneous industries that present a deviant behavior to the PLC model. The PLC model tends to assume that industries are single-product and homogeneous. Klepper and Thompson (2006) argue that the way we typically define industries in empirical work, often by SIC code, eliminates variability and heterogeneity in firm activity, which presents serious limitations in explaining industry dynamics. The same authors construct an alternative model with data from the laser industry. Here, industry is defined as a collection
of submarkets and these define the dynamics of industry evolution (firm entry, exit, and growth). The central force for change in the model proposed by Klepper and Thompson (2006) is the creation and destruction of submarkets and firm expansion and contraction in function of submarket creation and destruction. Other authors have explained industry dynamics based on submarket dynamics, such as for the disk drive (Christensen, 1997), lamp bulb, typewriter (Utterback, 1994), transistor, automobile (Clark, 1985), semiconductors (Henderson & Clark, 1990).

We discuss the limitations of the PLC models and introduce convergence across market boundaries. We highlight previous studies on convergence found in the literature. Our goal is to show how adjacency articulates with convergence across markets and present it as a new strategic opportunity for knowledge integration and market space creation.

There is, however, a similar limitation underlying submarket analysis when compared to the PLC model, the definition of “submarket” is typically restricted to one or more submarkets at a given industry. Authors typically isolate one industry, or the submarkets within, to study the underlying dynamics, but few studies have explored dynamics across sectors or industries. Rosenberg (1976) introduces the concept of technological convergence by which different industries come to share similar technological bases, disregarding boundaries across products markets, industries, or submarkets. He explains the process by which industries apparently unrelated in the terms of the nature and uses of the final product became very closely related on a technological basis, namely by sharing the same processes and having common challenges in production. Sewing machines and bicycles are an example of technological convergence for machine tool producers, such as producers of milling and grinding machinery. Producers develop tools that serve at the same time sewing machines and bicycles, following the same example. Technological convergence thus allows a higher degree of specialization at higher stages of producing machinery, taking advantage of economies of scope.

Technological convergence in the electronics sector has been extensively discussed over the last decades. Gambardella and Torrisi (1998) approached technological convergence and relate it with business diversification in five sectors: computers, telecommunications equipment, electronic components, other electronics, and non-electronic technologies in the U.S. and Europe during 1984-1992. Diversification is measured by the number of new subsidiaries and acquisitions, or joint venture and other collaborative agreements that show external growth operations. The authors suggest that the forces that drive technological convergence do not coincide with those that drive convergence in industries and product markets. Moreover, many firms that have made considerable attempts to cross industry borders and converge two or more
product markets have failed. They found that during the 1980s many of these companies have focused on fewer industries, while preserving a high degree of technological diversification.

Christensen (2011) explores in detail the IP security sector from 1989-2004 and provides an explanation of the evolution of industrial dynamics by studying convergence across product markets. The author identifies critical periods where reshufflings of boundaries occur across product markets and explains the evolution of industrial dynamics. During the formation of the IT security sector in the late 1980s and early 1990s, the sector has shown typical Schumpeterian features. The market was highly dynamic in terms of number of entering firms, flexibility in product design, and populated by start-ups developing technology-based products. By the mid 1990s, the first dominant designs emerge, at the same time as firm entry decreases and intensifies the number of exits. During this same period, complementary convergence proliferates first into lower-order convergence across product market niches by unifying and consolidating industry and then increasingly into higher-order.

Christensen (2011) also distinguishes between lower-order and higher-order integration, while respectively referring to integration of products, technologies, or activities that share some functionality in common, and to integration of one or more products or components into an integrated system or platform. Lower-order is typically associated with development by start-ups or young ventures and it typically requires technology integration and substitution that results in better performance or lower costs. For example, through lower-order integration, the security integrators and security specialists were competing on the same space in 2000s. On the other hand, higher-order convergence is associated with larger incumbents that develop a new solution to gain a secure position in the sector. For example, large incumbents like Cisco and Microsoft operating outside the security sector decided to integrate security functions. As a result, during the 2000’s firms like McAffee and Symantec that grew by lower-order convergence to be large integrator firms were competing with large incumbents like Cisco and Microsoft. In both cases, mergers and acquisitions were the central means for realizing convergence in the IT security sector. Mergers and acquisitions across different product markets (portfolio broadening) suggest a change from technology-based strategy to a lower-order convergence oriented towards the market. In the case of higher-order convergence, mergers and acquisitions make dominance broader and stronger by developing an integrative platform with security offerings. The IT security study provides a descriptive analysis of such trajectories of convergence at product market and sectoral levels and how these trajectories are reflected in firm strategy.
We discuss how to capitalize on opportunities uncovered by convergence and adjacency. How is learning an essential part of the convergence process and, ultimately, of the adjacency process.

Technological convergence for knowledge-intensive firms is highly dependent on integration of substantially different knowledge pools. This is often a problem for more mature markets since larger companies and incumbents present a higher level of specialization and strongly vertical. Larger firms that aim at market dominance must be able to reconcile greater depth and breadth (von Tunzelmann, 1996), however breadth within only their sector of activity, presenting limited flexibility in integrating knowledge outside of their area of expertise. Christensen (2011) defends that when knowledge pools from dissimilar areas are required for the integration of two or more product, then the integration process is more likely to occur through a merger and acquisition than through internal development. This is even more evident for larger firms that have limited flexibility in expanding to areas outside their core expertise, while is current practice to integrate know-how and human resources from an acquired company.

Learning, or the accumulation of knowledge following Solow (1997) emerges as the driving force behind increased efficiency and ultimately leads to economic growth. Arrow (1962) emphasis the role of informal ways of learning, “learning by doing”, as the basis for explaining that experience in the use of capital leads to an increment in the knowledge used in production. Arrow drew up a relatively simple model in which workers in a firm learn by means of production, thereby increasing firm’s productivity. It should also be noted that in this model, knowledge is accumulated only in the form of skills. But, as discussed extensively by Conceição and Heitor (1999), the contribution of the new economic growth theories has been posed to extend this reasoning to other types of learning, as well as to the accumulation of ideas. In particular, Lucas (1988) analyzed the accumulation of knowledge in the form of skills, but this time putting forward education as a formal learning process. In turn, Romer (1990) and Grossman and Helpman (1991) constructed models in which the accumulation of ideas results from effort put into research, another formal learning process. In addition, Foray and Lundvall’s analysis (1996) placed particular emphasis on the formation of networks of personal and professional contacts, which result from processes of social interaction. In this context, Conceição and Heitor (1999) summarized how these contributions fit into a framework of possibilities, which relates the accumulation of knowledge to the different kinds of learning that can lead to that accumulation.

According to Soete (1996), ideas and skills are no more than two sides of the same coin, two essential aspects of the process of knowledge accumulation. In other words, many good ideas are useless if the skills needed to use them do not exist. Studies by Pavitt (1987), Nelson
Nelson (1997) describes various circumstances, in which individuals, firms, universities, and other institutions have made use of their skills in order to increase their accumulation of knowledge, acquiring further skills as well as ideas.

The main implication of this argument is that the interdependence between ideas and skills casts doubt on the idea that the market supplies the necessary incentives for the production of skills. On the other hand, we argue that new ideas for product markets result from the interaction of knowledge-based ideas and skills. Although to a great extent skills result from the innate characteristics of an individual or from the history of a corporation, an institution or a country, they also depend on the learning processes (education, research, experience, social interaction) in which these entities are involved (North, 1990).

We propose here a model that illustrates knowledge integration through multiple stages of product development building on model presented by Conceição and Heitor (1999). These authors illustrate and propose a simplified framework to model the dependency between ideas (i.e., “software”) and skills (i.e., “wetware”), suggesting that it is through this interaction that new knowledge is generated and adapted by science-based firms. We consider that it is the

---

**Figure 2.** Representation of technological adjacency and creation of new market spaces. Basic knowledge is used as the basis for applied knowledge and product development until it the new product reaches the market. Note that knowledge is constantly iterated with constant feedback loops from basic knowledge until the product is fully development. In this example we represent two silos of development of medical device and pharmaceuticals. They typically do not intersect and share very little in common, but we here illustrate an example of technological adjacency that integrates a medical device with a drug, using know-how from both technologices (applied knowledge) and integrate it with basic knowledge. There is a technological link for integration of two adjacent technologies, reshuffling boundaries, and creating a new market space of drug-device integration.
basis for knowledge integration through the different stages of product development. We here intend to illustrate how technological adjacency occurs and creates new market spaces at the intersection of traditionally dissimilar industries, such as medical devices and pharmaceutical. Knowledge integration is not represented through every stage of product development in Figure 1 for simplicity. However, we consider that through the entire process of product development it requires basic and applied knowledge. The same way, technological adjacency only occurs when a link (idea or technical evidence) uncovers a possible technological adjacency between two or more sectors or industries that have little in common. Once there is a technological link that shows possible technological adjacency, development occurs with integration of basic knowledge (represented at the top of our sketch) and applied knowledge from each silo. Once integrated and developed, the new product resides in a new market space. Firms can rely on technological adjacencies to expand and create new markets spaces.

Which are the key agents in this process? Young firms have certainly higher flexibility to integrate knowledge from both adjacent spaces and develop new market spaces at the intersection of adjacent markets. Young firms are the key elements for dissemination of new knowledge integration. On the other hand, as we have previously discussed, integration process is more likely to occur through a merger and acquisition and that is possibly the more likely path for incumbents or larger firms. In the discussion section we discuss this topic in detail.

3. Methods
We focus our study in science-based firms and start-ups involved in novel medical technologies in drug-device combination products, cell therapies, and tissue engineering to restore blood vessels function. These three areas constitute the research setting of our experiment. Our ultimate goal is to understand how firms develop the necessary skills and learn to access and create new markets, capitalizing on technology to evolve from the current standard of care. In particular, we explore how these firms use ideas and skills associated with distinct markets to move from medical devices, to combination products and to cell therapies, and, finally, to tissue engineering products.

3.1 Data and Sources
Data for this analysis was collected during a 26-month period. It was conducted from the winter 2009 through the spring of 2010 (see, for details, Couto 2011). The data collected consists of a comprehensive longitudinal dataset covering the historic evolution of combinations products, in particular drug-eluting stents, cell therapies, and tissue engineering. Our dataset includes data from multiple sources, such as interviews, published data in
international journals and other publicly available information, SEC files, firm annual reports, databases, deals and other agreements between firms or other organizations, and clinical trials.

**Expert Interviews**

We interviewed experts from three areas - combinations products, cell therapies, and tissue engineering. In combination products, we interviewed medical doctors with expertise in the field of interventional cardiology and imaging, scholars with scientific expertise in combination products, and industry experts from the scientific and regulatory viewpoints. The questions were focused on the evolution of technologies to restore blood vessel function, with particular emphasis on drug-eluting stents, and the evolution of Boston Scientific Corporation and its role in the development of drug-eluting stents.

The interviews with medical doctors aimed to provide perspectives on the evolution of procedures from cardiopulmonary bypass to angioplasty, and related medical devices. Also, we discussed in detail the evolution from stents to drug-eluting stents and related criteria used in clinical assessments to decide whether to implant a stent or a drug-eluting stent. The medical doctors interviewed included David Sosnovik, Professor at Harvard Medical School and Director of Cardiovascular Molecular Imaging Research at Massachusetts General Hospital, Douglas Drachman, an interventional cardiologist at Massachusetts General Hospital, and Farouc Jafer, also an interventional cardiologist at Massachusetts General Hospital and Professor of Medicine at Harvard Medical School.

The research scholars interviewed included Elazer Edelman, a cardiologist and professor at Harvard Medical School, coordinating a laboratory at MIT focusing on developing new solutions to regenerate blood vessels. A patch with cell therapies to apply together with a stent was developed in this laboratory and Elazer Edelman helped forming Pervasis, a spin-of using a cell-based therapy patch to treat stenosis.

To support the case-study of Boston Scientific we interviewed a group of experts that complemented each other's perspectives. In particular, we discussed all the data collected from our previous research and the role of Boston Scientific in developing drugs-eluting stents with James Barry, the Scientific Director at Boston Scientific. We also discussed regulatory topics with Tony Blank, the Vice-President for Cardiovascular Regulatory Affairs at Boston Scientific Corporation, while related issues about the firm formation and historic and technological overview of the stent space were specifically addressed with its founder, John Abele.
Deals and Agreements

Data on R&D agreements was obtained from Recombinant Capital (Recap), a private database that tracks and analyzes deals in the healthcare area, particularly from the biotechnology and pharmaceutical fields, as used in prior studies (Lane and Lubatkin, 1998, Stuart et al., 2007; Rothaermel and Thursby, 2007). Our analysis was extended to agreements for medical devices firms and included a comprehensive list of data in terms of SEC and FDA filings, press releases, industry conferences, and industry contacts, among others.

The Recap database has the most comprehensive agreement data as it includes not only the agreements between pharmaceutical and biotechnology firms, but also other organizations, such as government institutions and universities (Schilling, 2009). Validity of the data from Recap was previously checked by Schilling (2009) against other databases such Securities Data Firm, MERIT-CATI, CORE, and Bioscan and it showed the most comprehensive list of results.

Secondary Data Sources

Primary research from databases was supplemented by extensive analysis of technical journals, technical and business evidence from firm websites or SEC files, regulatory guidelines and reports published by the FDA and EMA, as well as the following secondary data sources.

Clinical Trials. Clinical trial data was collected from ClinicalTrials.gov, a publicly available database of clinical trials registry. Wood (2009) considered it the most complete web-based registry of clinical trials. It was created in 2000 by the U.S. National Institutes of Health, through its National Library of Medicine, in collaboration with the FDA, in response to Section 113 of the 1997 Food and Drug Administration Modernization Act. Section 113 of this Act specifically requires the U.S. Department of Health "to establish a registry of clinical trials for both federally and privately funded trials of experimental treatments for serious or life-threatening diseases".

It should be noted that ClinicalTrials.gov database was developed to make the results of clinical trials publicly available to benefit patients, physicians, researchers, firms, and investors (Wood, 2009). Patients can now search for new clinical trials for their clinical conditions. Physicians and researchers can keep up-to-date with the last clinical developments and innovative technologies that reach the clinical setting. Firms and investors have an easy platform to track clinical developments in clinical areas of interest.

Mandatory elements for registration were initially trial's title, condition studied, trial design, and the intervention studied. The FDA Amendments Act, enacted in 2007, required additional elements for registration, such as the trial start date and to report primary and secondary
outcomes within one year of trial completion. Ross (2009) demonstrates that nearly 100% of the completed trials registered until 2007 at ClinicalTrials.gov reported all data elements mandated. An exhaustive list of the mandatory and optional data elements is available at Wood (2009).

**Firm data.** Corporation SEC files were taken from the U.S. Securities and Exchange Commission database, which considers all the SEC files of the public firms registered in the U.S.. We were particularly focused in the 10-k Forms, corresponding to the annual report submitted by the firms. Their analysis was complemented with Annual Reports available through the internet, as well as with press releases, and other information at investors-related links. This information is particularly useful for firms that were recently founded and have no SEC files registered or annual reports publicly available.

**FDA and EMA regulations.** The websites of the U.S. Food and Drug Administration (www.fda.gov) and the European Medicines Agency (www.ema.europa.eu) provided useful information about the historical perspective on the evolution of the agencies, at the same time as regulation evolved. Both agencies compile in their websites all the information regarding new directives, guidelines for their staff and industry, and guidance of how to proceed during pre-clinical and clinical testing.

### 3.2 Research Methodology

The data collected about products and firms and the interviews performed were crucial to understand the evolution of technologies to restore blood vessels function. We focused our analysis on firms entering the drug-device combination from the previous generation of single-regulated products. We collected data from firm's SEC files, public available data, deals, and clinical trials that involved the firms in the period from 1993 to 2010. For Boston Scientific, in particular, collected data was complemented with interviews to key experts of the firm and a visit to the firm R&D facilities.

We listed and analyzed all the drug-eluting stents approved by the FDA that required “New Drug Application” submission. This resulted in 12 drug-eluting stents with market approval in the U.S.. Information on approval of drug-eluting stent was collected from the FDA website, and re-examined at the firms websites and financial reports. All the clinical data was retrieved from the ClinicalTrials.gov database, when searching for each individual sponsor. It was then complemented with data from scientific publications. All the deals related with the sponsor firm for each combination product were searched in the Recap database and firm websites.
4. The evolution of therapies to restore blood vessel function

Figure 3 illustrates the evolution of medical procedures and technologies to restore blood vessel function and suggests how technology adjacency enables the development of new medical solutions with improved features that solve limitations of the current standard of care. The evolution highlights two distinct procedures, three generations of stents (bare-metal, drug-eluting, biological) and the introduction of engineered blood vessels.

Adjacent technologies bridge distinct product designs by including adjacent technological features. For example, bare-metal stents with coating make the bridge between bare-metal and drug-eluting stents. The inclusion of a polymeric coating to the bare-metal stent results in a design closer to drug-eluting stent, which has a drug enclosed in a polymeric coating. The polymer itself when coating the stent improves its application inside the blood vessel. By including adjacent technologies to previous product designs, firms bridged devices to assist medical procedures, medical devices, drug-device combination products, and biological-device combination products. We argue this is a stepwise process towards the introduction of tissue engineering products, with new solutions in clinical development to restore blood vessels flow including biological stents and engineered blood vessels.
Cardiopulmonary bypass is a procedure applied to restore blood flow by placing an alternative vessel to bypass a stenotic vessel. Stenosis is an occlusion due to the formation of atheromatous plaque inside a blood vessel. Cardiopulmonary bypass was enabled by two key developments: the introduction of open-heart surgery, in 1952; and extracorporeal circulation support. Open-heart surgery was one of the most important medical advances of the XX century (Gravlee, 2007). Extracorporeal circulation was enabled by the development of different oxygenators and pumps. Three key generations of pumps marked the improvement of the cardiopulmonary bypass procedure, namely: i) the multicam-activated Sigma motor pump; ii) the roller pump that demonstrated higher reliability; and iii) the centrifugal pump introduced in 1968 by BioMedicus, Inc..

Physicians started adopting open cardiotomy in 1956 for cardiopulmonary bypass and restore blood vessel function. This procedure is still applied today worldwide. For less severe cases, however, alternative procedures were developed, such as angioplasty. Angioplasty enabled the treatment of blood vessels by clearing the inside of the vessel, instead of placing an alternative vessel to bypass the damaged artery and restore blood flow.

In 1974, Andreas Gruentzig performed the first balloon angioplasty in humans to treat coronary artery stenosis (Rapaport, 1983). The concept of remodeling the artery was introduced by Charles Dotter in 1964 (see, for details, Dotter and Judkins, 2005), but angioplasty was only performed in humans a decade later when two distinct technologies enabled it: the steerable catheter and the arterial wall expanding device. Angioplasty became then the most applied medical procedure in the world (Angioplasty, 2011). According to the same organization, in 1997 only, over one million angioplasties were already performed worldwide. To support angioplasty, a large number of interventional devices were invented and developed between the late 1980s and early 1990s, such as rotational atherectomy devices (Rotablator), intravascular ultrasound, and stents. In 1994 the Palmaz-Schatz stent, introduced by Cordis, was approved by the FDA for use in the U.S.. The stent became the most used technology, or the dominant design according to Utterback and Abernathy (1975), for the treatment of coronary vascular disease and restoring blood vessel function.

Within the past two decades, two different generations of stents were launched into the market. Stents are simple solutions - less invasive and highly deliverable - when compared with previous technologies and procedures, such as cardiopulmonary bypass. A steerable catheter and a driving system with an arterial expanding device are required to deploy the stent. The stent is expanded once inside the occlusion spot to dilate the blood vessel and improve blood flow.
In order to improve delivery of these stents, coatings were applied to the surface of the metal, such as heparin (Serruys et al., 1998). With the development of a stent with a coating became apparent that the metallic frame could load compounds to improve delivery and prevent in-stent restenosis. Though bare-metal stents showed to be effective in reducing the rate of stenosis when compared to balloon angioplasty, in-stent restenosis was a major complication after stent deployment (Serruys et al., 1994; Fischman et al., 1994). In other words, patients with stents may develop stenosis inside the stent, occluding the vessel once again.

Less than one decade after the introduction of the first bare-metal stent, the FDA approved the first drug-eluting stent. In 2003, the second generation of stents was introduced. Drug-eluting stent combines a drug enclosed in a polymer that coats the metallic framework. Adding new features typical from the pharmaceutical space into a bare-metal stent enabled the emergence of a new medical solution. Scientific know-how from drugs and medical devices converged in developing drug-eluting stents and improving stent functionality by delivering drug in situ. The goal of this new solution is to prevent in-stent restenosis. Firms capitalized on the convergence of scientific know-how from drug-device combination.

Firms that were able to converge know-how from drugs and devices and develop a stent, advanced through adjacency from the medical device space to the drug-device combination space. These firms created a new market space intersecting drugs and medical devices for the treatment of coronary artery disease. According to Kim and Mauborgne (2005) this maybe considered a “blue ocean”, in which firms create a new and uncontested market space, building competitive advantages and going beyond the competition within established drug and medical device boundaries. As a result, they are in a vantage point to further expand within the drug-device combination space and develop new solutions for other clinical applications.

The drug-eluting stent complemented the mechanical action of containing the blood vessel occlusion with high deliverability and drug release to inhibit restenosis. The new generation of stents overcomes, therefore, a major problem of the related first generation (Moses et al., 2003; Morice et al., 2002). Since 2003, drug-eluting stents were developed for patients with a broader range of complications associated with stenosis. Currently there are drug-eluting stents approved to treat patients with multiple lesions, small vessels, long lesions (Turco, 2008), and diabetes (John et al., 2009; Mahmud et al., 2009).

The cardiac stent market grew from $1.5 billion in 2001 to $3.3 billion in 2005. In 2006, however, the market penetration for Boston Scientific Corporation decreased from 80% to 60%. This was due to the divulgation of specific studies, namely SCAAR and COURAGE (see Lagerqvist et al., 2007), showing increased mortality for drug-eluting stents when compared to bare-metal stents. For example, SCAAR clinical study showed that after six months, drug-
eluting stent have a 0.5%/year increased mortality versus bare-metal. According to our interviewees, the comparison was established between the overall number of patients with bare-metal and drug-eluting stents ever enrolled in clinical trials. Patients with single lesions and single vessel stents were compared with a population of patients with diabetes, small vessel, long lesions and multi-vessel lesions that used drug-eluting stents. When similar populations were compared (single lesions, single vessel), for both bare-metal and drug-eluting stents, the results were significantly different. In the updated SCAAR study from 2003 to 2005 (also adding one more year of data), the adjusted death curves were equal for both types of stents. In 2006, the adoption rate of drug-eluting stents over total stents significantly decreased and today is around 70% for drug-eluting stents.

The physicians interviewed throughout our analysis were consistent in that physicians massively adopted the new generation of stents as it had showed improved results in clinical studies. After 2006, however, the adoption further decreased as a consequence of the SCAAR and COURAGE publications. After this controversy, physicians evaluated more carefully the need for more expensive drug-eluting stents, versus the bare-metal.

It should be noted that the criteria to chose between bare-metal and drug-eluting stent depends on the patient medical history, clinical condition, and capacity to comply with the medication requirements. Patients administered with drug-eluting stents have to comply with antiplatelet medication for a period of time so that the site of stent placement heals. Otherwise, the patient has increased risk of thrombosis. If the patient has a track record of poor compliance with previous medications, physicians would not recommend drug-eluting stents once antiplatelet medication is needed. Also, patients that cannot take antiplatelet medications are likely to get a bare-metal stent.

Though there is evidence of improvement of in-stent restenosis, major complications resulting from neointima thickness were evident in clinical trials using the new generation of stents. Biodegradable stents are an intermediate design between drug-eluting and the new generations of biological stents. It presents potential advantages: stent disappears from the treated site include avoid in-stent restenosis, reduces or abolishes late stent thrombosis, facilitates repeated treatments (surgical or percutaneous) to the same site (Ormiston and Serruys, 2009). Also, biodegradable stents have a potential pediatric role in allowing vessel growth and do not need surgical removal.

Examples of biodegradable stents currently being tested in clinical trials include the following: i) the Igaki-Tamai stent, by Igaki Medical Planning Firm (Japan), constructed from poly-L-lactic acid; ii) a tubular magnesium WE-43 by Biotronik (Germany); iii) the BVS everolimus-eluting stent by Abbott Vascular (U.S.); iv) the REVA by Reva Medical Inc (U.S.) from an
absorbable tyrosine-derived polycarbonate polymer that metabolizes to amino acids, ethanol, and carbon dioxide; and v) the bioabsorbable therapeutics stent by Bioabsorbable Therapeutics Inc (U.S.), a fully bioabsorbable sirolimus-eluting stent that also releases salicylic acid (Ormiston and Serruys, 2009).

When considering the development of biodegradable stents, Renu Virmani, a medical doctor at CVPath Institute (Gaithersburg, US) mentioned that "…these stents won't be for everyone. There will probably be 10% to 20% of the population in whom they'll work… . But overall, I'm skeptical. This is a thick-strut [stent], and yes, it will dissolve in two years, but after two years even bare-metal stents have healed." (see O'Riordan, 2010). This leads us to consider that experts are still skeptical today about the potential benefit of biodegradable stents. Besides, they don't seem to solve problems associated with neointima thickness.

Other solutions are being developed to treat stenosis, at the same time as avoiding neointima thickness and in-stent restenosis. New solutions include combination of drug-eluting stent and a patch of cell therapy applied outside of the vessel to avoid neointima thickness (www.pervasistx.com), and biological stents that regenerate the injured vessel from the inside. For example, OrbusNeich has developed Genous Bioengineered stents to promote rapid endothelialization limiting the number of stent thrombosis (OrbusNeich 2008, 2010). This stent has a layer of anti-CD34 antibodies to attract circulating endothelial progenitor cells aimed to promote faster and more complete cell endothelialization and vessel healing after stent deployment.

Another example is that associated with Cytografix Tissue Engineering that is developing engineered blood vessels for more severe cases that still require cardiopulmonary bypass surgery (Cytografix Tissue Engineering, 2009). These new solutions are today in clinical trials. Engineered blood vessels are an alternative to the patient's own vessels that are today typically harvested from his leg to place as the bypass (see, for details, www.cytografix.com). They result from the convergence of material science, biology and emerging engineering tools, which has enabled the development of three-dimensional blood vessels.

It should be noted that, today in the market, there are two procedures to restore blood vessel function - cardiopulmonary bypass and angioplasty, applied according to the severity of the vessel lesion. Angioplasty has become widely adopted and more competitive as enabling technologies, such as stents and drug-eluting stents, were developed and introduced in the market. Firms operating in the medical device space realized that the combined effect of drugs and medical devices could enhance the final product in preventing restenosis.
Our analysis has shown that drug-device integration emerged as a plausible solution to solve bare-metal limitations and also as an attractive market that goes beyond traditional market boundaries in drug and medical device. The same is happening with stents integrating biologics and engineered blood vessels. The integration of new features from distinct technological fields enables the development of improved products at the intersection of two or more markets. Firms integrated existing knowledge from drugs, devices and polymers, bridged them with drug-eluting stents and moved through adjacency from the first generation to the second generation of products. The same reasoning is today being applied to the development of new generations of biological stents. Boston Scientific case study illustrated an example of a firm that moved through adjacency. This case study is described in detail in the next section.

5. Technology adjacency fostering corporate change and growth: the case of Boston Scientific Corporation

Our analysis of Boston Scientific Corporation, BSC, provides the necessary evidence to complement the analysis above and to facilitate our understanding of the phenomenon of technology adjacency, its impact on corporate change and growth, as well as the way it is influenced by external factors, namely stakeholder engagement. It bridges scientific convergence to firm strategy and suggest the impact of expanding to adjacent markets to foster firm growth and its dependence on the crucial role played by physicians. We focus on drug-device markets and compare firms in the sector, as well as the leading strategy of Boston Scientific in incorporating ideas and skills from other firms through collaborations and acquisitions.

A brief historical background

John Abele and Pete Nicholas founded Boston Scientific Corporation in 1979 with the goal of developing less invasive medical devices for surgery and other medical procedures (BSC, 2005). BSC has become, at the same time, the holding firm of Medi-Tech and introduced during the 1970s the steerable catheter and the arterial wall expanding device used to perform angioplasty - instruments used by Andreas Gruentzig (German physician, University Hospital in Zurich, Switzerland) when performing pioneering experiments. It was under this context that BSC became the leading firm in the interventional cardiology area, inheriting Medi-Tech’s reputation in developing innovative medical devices for non-invasive procedures.

During the 1980s and early 1990s, BSC expanded into different business areas through the acquisition of multiple firms, from which resulted three key divisions: cardiology, peripheral vascular products, and gastrointestinal and pulmonary endoscopic accessories. In 1995, BSC
acquired Scimed Life Sciences, forming the sixth division of BSC at the time, in interventional cardiology. BSC became then a leader in interventional cardiology. By the end of 1997, BSC had acquired nine other firms relevant to the firm pipeline (BSC, 1998). According to our analysis from Recap.com database, in 2009 BSC had already acquired 24 firms.

Acquisitions have always played a key role in the strategic expansion of the firm, such as the acquisition of Guidant (together with Abbott Laboratories) in January 2006 to form the cardiac rhythm management division. Today, BSC has established nine divisions: cardiac rhythm management, gastroenterology, interventional bronchoscopy, interventional cardiology, interventional radiology, oncology, pain management, urology, and women's health (www.bostonscientific.com).

In 1995, BSC made an exclusive license, supply, and development agreement with Medinol to first access the coronary stent space. Medinol had developed a flexible stent design - NIR stent - with improved cost-efficient manufacturing process (margins of up to 80%, compared to 35% for the competitors) (Suwanvanichkij et al., 2006). Under the agreement, Medinol would become BSC's primary source of stents for at least ten years, help building an alternative production line for BSC, training manufacturing personnel to the alternative production line, as well as training marketing and sales personnel on NIR® stent (Suwanvanichkij et al., 2006). In exchange, BSC would bundle the NIR stent with balloon-delivery systems, obtain regulatory approval in the U.S. (and elsewhere), and market the devices through its global sales channels. BSC also purchased a 22% interest in Medinol. Under this agreement, both parties should disclose to each other all inventions, ideas, and improvements relating to stent development during the term of the agreement, and jointly participate in a development program for future stents, including sharing intellectual property between other benefits.

BSC started to develop their new line of drug-device combination stents and Medinol claimed they that it was against their agreement. Court disputes were constant. Medinol accused BSC of capitalize on their skills in manufacturing, product development, marketing and sales to develop a new line of products. Finally, BSC ended up acquiring Medinol in 2001 after long disputes and negotiations.

Around 1997, BSC started to develop a new line of drug-device combination stents - the Taxus platform - for reducing coronary restenosis (BSC, 2004). This is associated with establishing strategic research collaborations intended to add complementary technologies to improve the new platform. A license agreement was established with Angiotech Pharmaceuticals in 1997 for the use of paclitaxel and a supply agreement with Natural Pharmaceuticals to provide the drug. In addition, the acquisitions of Interventional Therapeutics and Quanam Medical Corporation in 2001 had an important role in the development of drug-eluting stents. This is
because the former had extensive expertise in interventional cardiology in invasive medical devices, while Quanam Medical was a manufacturer of medical devices specialized in drug delivery systems, with a vast portfolio of biomaterials for drug delivery integration. BSC integrated new features from distinct areas, namely pharmaceuticals and biotechnology, into its device platform to improve product performance. The firm is currently positioned in a new market space at the convergence of drugs and medical devices.

**Integrating skills towards a knowledge-based strategy**

Interviewing James Barry (Scientific Director, BSC) provided a clear illustration of the relevant role played by a diversified group of skills and skilled people in BSC’s corporate strategy. In particular the role played by physicians in the improvement of medical devices has become significant in the transition from medical devices to drug-eluting stents. For example, before the agreement with Angiotech, James Barry discussed with a good friend and interventional radiologist working at Angiotech (who implanted esophageal drug stent in patients that had cancer) the possibility of integrating drugs with medical devices as a system for controlled drug delivery. The idea was simple and based on containing cancer with a stent. A cancer drug can be placed into a stent to prevent tumors for growing back inside the stent. The solution does not provide a live saving solution, but improves patients’ quality of life. James Barry made an analogy for the treatment of coronary artery disease: if tumors can be stopped from growing into the device, also artery tissue can be stopped from growing into the stent. At this time, Angiotech tried to put a drug into a stent, but every time, in the animal tests, the results were terrible blood clots. They quickly realize that they couldn't do it anyway, because they didn't have a stent. According to James Barry ”Angiotech was looking for a device partner at the time. We kind met in a meeting and we started to think it together”. BSC and Angiotech ended up collaborating for a long period in combining drugs with medical devices.

This brief story helps us understand the evolving strategy of BSC towards incorporating different skills and using the resulting institutional capacity to create and aces to new market through technology adjacency. Only during 2003, BSC invested more that $350 million in approximately 25 strategic initiatives. In 2004, BSC invested more than $1.2 billion in approximately 50 new and existing strategic alliances. All these initiatives were aimed to expand BSCs product offerings by adding new or complementary skills and the firm technological capacity (BSC, 2004).
It should be noted that BSC established collaborations mainly with firms supported by them in their initial stages, always including the possibility of acquiring those firms if the technology was successfully integrated within BSC’s final products and markets. James Barry identified this as the BSC growth strategy. Examples of this type of collaboration are the agreements with Reva Medical and Labcoat. With Reva Medical, BSC agreed on making an equity investment in November 2004 with an exclusive option to acquire it afterwards. With Labcoat, BSC started by establishing an initial collaboration, with the final acquisition being accomplished in January 2009.

Both firms, Reva Medical and Labcoat, shared common appealing technologies to integrate within BSC’s product pipeline. REVA was developing a balloon-expandable bioresorbable drug-eluting stent that combines distinctive geometry and resorbable polymer properties. Labcoat was developing a novel technology to coat drug-eluting stents that uses precisely metered droplets of a biodegradable polymer and drug formulation to create a thin coating confined to the outer surface of a coronary stent. It was designed to significantly reduce the amount of polymer and drug to which the vessel wall is exposed, while minimizing polymer and drug on the inner surface of the stent where endothelial cell growth is required for healing. Once the drug has been delivered, the biodegradable coating resorbs, leaving behind only the bare-metal stent. This technology was tested in JACTAX Stent, using BSC bare-metal stent and Labcoat’s coating technology. It is currently integrated into BSC products.

It should also be noted that BSC assumed a key role in developing drug-eluting stents; it has three distinct drug-eluting platforms - Taxus, Promus, and ION. From 2004 to 2011, BSC launched 11 different DES in the European and U.S. markets, with at least five of them being incremental improvements over the Taxus platform. Promus and Promus 2.25 are exactly the same as Abbotts Xience V Everolimus and Xience Nano Everolimus, respectively. They resulted from a previous agreement, in which Boston Scientific acquired Guidant Corporation. ION Paclitaxel-Eluting is a new platform developed by Boston Scientific and approved in the U.S. in April 2011.

**Expanding to new markets and the role of physicians**

Our analysis suggest that BSC has explored new ways of introducing combination products that may bring improved medical solutions, as well as facilitates expansion of the firm to new markets. Recently, the firm established investment agreements in emergent biotechnological fields, including gene therapy with Corautus, and cell-based therapies with Osiris Therapeutics and Theregen. This evolving path is gaining from the accumulation of knowledge by the firm
and is following previous “learning” procedures: BSC invests in the firms, supporting development, sometimes buying firm’s equity, and assuring an exclusive right for a future acquisition or license agreement. The agreements are settled to secure a possible future position in getting access to the technology if it is attractive to incorporate in their product portfolio. As the technologies advanced in development, BSC takes the decision about licensing or acquisition the technologies. Currently, BSC is exploring potential candidates to deliver new solutions to treat coronary artery disease.

We note that, besides Reva Medical and Labcoat, BSC made strategic agreements with Couratus and Osiris Therapeutics in 2003 and with BioSeek in 2006 with the purpose of securing the option of acquiring those firms or rights. The agreement with Corautus included development, license, equity, and distribution for the application of VEGF-2 gene therapy for vascular diseases. With Osiris Therapeutics, BSC established a co-development, equity, license, loan, and settlement agreement for access the application of Provaceel in myocardial infarction and chronic ischemia. In addition, BSC made collaboration, license, and research agreements with BioSeek to secure the right to use BioMAP platform for drug discovery. BioMAP incorporates predictive human cell-based disease models to help selecting new drug candidates.

The analysis suggests that the expansion strategy adopted by BSC, since it was formed in 1979 and including the transition from bare-metal to drug-eluting stents, enabled growth. Revenues increased from $1.8 million in 1979 to $73 million when it went public in 1992, $3.5 billion in 2003, and $5.6 billion in 2004. Nowadays, BSC has revenues superior to $8 billion. The evolution of net sales since BSC went public until 2009 is illustrated in Figure 4.

The release of drug-eluting stents to the market, in 2004, is associated with BSC’s fastest net sales increase since it went public. This was the year when TAXUS was released into the U.S. market and adoption was increasing in Europe after its launch in 2003. In 2004, approximately 38% of BSC net sales resulted from sales of the TAXUS stent system. By the end of 2004, TAXUS had 56% of market share in the U.S. and BSC’s overall intra-coronary stent revenues increased from 11% in 2003 to 53% in 2004 (BSC, 2004). Overall net sales and partial revenues due to drug-eluting stents and interventional cardiology are also illustrated in Figure 4.

![Net Sales (Billions) vs Time (year)](image-url)
Figure 4: Impact of drug-eluting stent development in Boston Scientific Corporation net sales. The net sales of the firm are represented in billions of U.S. dollars in time. Firm net sales are segmented per drug-eluting stent (DES) and overall interventional cardiology (IC) except for DES. Taxus Express was launched in 2003 in Europe and in 2004 in the U.S.

The data of Figure 4 should also be considered in terms of related firms and it should be noted that the stock price of the four leading firms in drug-eluting stents - BSC, Johnson & Johnson, Abbott, Medtronic - show the stock appreciation of BSC when the first DES was released into the European and U.S. markets. The transition from pure medical devices to drug-device combination products had a significant impact on BSC’s growth. This shows that beyond the fact BSC grew at the fastest pace during this period, also its stock prices increase substantially when compared to other players in the field.

To provide further evidence for our main argument, Figure 5 illustrates the stock price of the four leading firms developing drug-eluting stent for cardiovascular applications and compares firms with very different market capitalizations. As of August 31, 2011, Boston Scientific has a market capitalization of $16 billion, Medtronic of $37 billion, Abbott Laboratories of $80 billion, and Johnson & Johnson of $180 billion. However, at the peak in 2004, Boston Scientific’s stock price increased to 349% its price, which further supports the evidence that this was the period of the firm highest growth rate. The launch of drug-eluting stents to the market had a positive effect in BSC expansion, though it is the smallest player in the field.

Figure 5: Stock price of Boston Scientific Corporation, Johnson & Johnson, Abbott, Medtronic, and Dow as the reference stock price, from January 1, 2000 to December 31, 2010. Data of stock price was obtained from google finance (www.google.com/finance) assessed online on August 30, 2011.
To sum-up, our analysis suggest that the integration of new technological features from adjacent markets enabled BSC to develop improved products to restore blood vessel function and enter larger markets at the convergence of two or more fields of expertise. In particular the transition from bare-metal to drug-eluting stents enabled BSC to bridge the drug-device space, resulting in a new market. BSC used a strategy of licensing and acquiring technologies and firms to add adjacent technological features to their stents platforms. This strategy had a positive impact in firm growth and expansion. In particular, the transition from bare-metal to drug-eluting stents had the highest impact in firm growth since it was formed in 1979.

However, we argue that understanding BSC’s technological development requires an in-depth perception of the crucial role physicians played in the process of market access and growth, namely in the conception, design, and adoption of stents. Andreas Gruentizg performed the first angioplasty procedures with medical devices from BSC. This established a technological standard for angioplasty procedures (Rapaport, 1983). In addition, James Barry’s interaction with his friend and interventional cardiologist at Angiotech resulted in the idea for combining drug into a device for coronary artery disease. Last, but not least, the role of physicians is also associated with the decrease in the adoption of drug-eluting stents after SCAAR and COURAGE studies, which generated massive discussion among the medical community and increased medical concerns in prescribing drug-eluting stents.

Our conclusion is that the process of technology adjacency, which we identify as a strategic tool that facilitates science-based firms to expand to distinct markets, creating new and uncontested market spaces, is strongly influenced by the crucial role physicians play. Although new solutions to improve stent technology are continuously being developed, including biodegradable materials or cell therapies, our fieldwork has shown an increasing skepticism about the development of new stent designs and their adoption. This is further discussed in the section below.

6. Discussion – strategies and stakeholders to move through technology adjacency from medical devices to combination products

In the paragraphs above we have described the evolution of technologies used to restore blood vessel function and the specific case of Boston Scientific Corporation to argue that scientific convergence enables the development of new solutions with improved functionality when compared to the current standard of care. In addition, our evidence shows that science-based firms use convergence of distinct technological domains strategically to identify and bridge adjacent markets and this facilitates the creation of new market spaces. Bridging these adjacent markets constitute an opportunity for firm growth and expansion.
We now deepen our analysis and discuss our results from the perspective of different strategies and potential stakeholders, looking at main challenges for technology adjacency to be effective. Our motivation is to clarify potential limiting factors affecting this concept and its impact on corporate change and growth.

Our analysis gains from, and is influenced by the fact that the last decade has been particularly stimulating in helping us understand the need for improvement over the usually assumptions about economic development, and, instead, putting more emphasis on what else determines how consumers behave: the inertia of habits, the importance of default options on decision making, and more generally, the effect of how information is presented, and why human behavior so often follows the pattern of fast thinking/pattern recognition rather than slow, effortful, analytical thinking (Kahneman, 2011).

Kahneman’s experiments with two groups of patients undergoing painful colonoscopies are very inspiring. Group A got the normal procedure, while Group B — without patients being told — had a few extra minutes of mild discomfort after the end of the examination. Since the prolonging of Group B’s colonoscopies meant that the procedure ended less painfully, the patients in this group retrospectively minded it less. As with colonoscopies, so too with many other actions and decisions in life. It is the remembering self that calls the shots, not the experiencing self.

Also in recent years, Thaler and Sunstein (2009) argue about the increasing relevance of nudge and the way individuals are influenced by their social interactions. Nudge takes our humanness as a given. Knowing how people think can help designing choice environments that make it easier for people to choose what is best, at both individual and corporate level.

Our analysis is also influenced by the evolving phenomenon of “user innovation” and the work of Eric von Hippel over the decades (Von Hippel, 1988; 2009). In particular, DeMonaco et al. (2006) has shown the major role of clinicians in the discovery of new, off-label uses for drug therapies, which has important regulatory and health policy implications. In this study, clinicians found new and off-label uses in 22 out of 29 drugs. Physicians are an important source of innovation.

Also, user innovation played a very relevant role in the development of new medical devices, in particular, surgery instruments. An example is the development of specific tools by surgeons to apply in radial keratotomy to correct myopia. The history of medical procedures to shape the human cornea started as early as 1885 as described by Choi et al. [2002]. A diamond knife was used to perform incisions that penetrate in corneal stroma. Svyatoslav Fyodorov performed the procedure to removed glass from the eye of one of his patients. A boy with eyeglasses fell off
his bicycle and his glasses shattered on impact. To remove the glass particles lodging in his eyes, Fyodorov made numerous radial incisions that extended from the pupil to the periphery of the cornea. After the glass was removed and the cornea healed, Fyodorov found that the patient’s eyesight had improved significantly. Not only Fyodorov had supporting evidence for the advance of a new procedure to correct myopia, but he also helped developing new medical equipment to it—such as the diamond knife [Tannebaum, 1995]. Fyodorov made several contributions in the late 1970s in instrumentation and also pioneered the earliest formulas to improve its reproducibility [Stulting et al., 2000].

It is in this context that the following paragraphs discuss our results under two main headings. First, we address the evidence provided by our work in terms of collaborative patterns facilitating technology adjacency. Second, we discuss the key role of potential stakeholders affecting technology adjacency, as well as their potential limiting factors affecting corporate change and growth.

Collaborative patterns facilitating technology adjacency

Our analysis shows that the evolution of technologies to restore blood vessel function was enabled by the development of complementary technologies and convergence of multiple areas of knowledge. Cardiopulmonary bypass was enabled by the development of oxygenators and pumps that allowed extracorporeal circulation to perform open-heart surgery. Oxygenators and pumps for extracorporeal circulation are the complementary technologies that enabled the progression to cardiopulmonary bypass. Angioplasty, in turn, was enabled by the development of medical devices, such as steerable catheter and the arterial wall-expanding device. Later, the development of stents extended the applicability of angioplasty into a wider spectrum of patients. Stent function was then enhanced with drug delivery in order to overcome some of the limitations of the stent platform, such as the formation of in-stent restenosis.

We argue that the emergence of drug-device combination products was only possible due to the integration of ideas and skills from the drug space, an adjacent market, into the medical device. It required the convergence of knowledge from life sciences and engineering, including new understanding of the pharmacological and physiological behavior of drugs once inside the body, as well as improved understanding of device’s degradation and structural behavior.

The development of stents combined with cell therapies, or biological stents, bridged medical devices with another adjacent market - biologics. The former combine the mechanical action of the medical device with the regeneration potential of the stem cells outside the vessel to avoid in-stent restenosis and neointima thickness. The latter present biologics capable of promoting
faster vessel healing and avoiding neointima thickness. Again, the emergence of biologic-device combination products is calling for the convergence of ideas and skills from the life sciences, such as cell culture and physiological behavior, and engineering.

Engineered blood vessels result from the convergence of material science, engineering, and biology. This includes skills to develop biodegradable scaffolds and new understanding of cell culture, vascularization, and the physiological behavior of the vessels once inside the body. Engineered blood vessels will be a complementary technology to perform cardiopulmonary bypass for severe cases of stenosis, overcoming the limitation of using the patient own vessels. At the same time, engineered blood vessels result from a path of evolution from pure mechanical, to fully integrated biological solutions. They are likely to compete directly with drug-device or biological-device stents used in angioplasty.

It should be noted that the evolution of technologies to restore blood vessel function was progressive through the development and introduction of intermediary solutions that bridged the pure medical device with the drug-device and biologic-device. Nevertheless, the concept of “dominant design” is relevant here, suggesting that different designs emerge with the evolving pattern of new technologies. Tushman and Murmann (1998) refer to an “era of ferment” as characterized by experimentation and failure that results in the emergence of a superior design.

For example, bare-metal stents and drug-eluting stents are clearly two dominant designs. Different intermediary designs emerged in the process of their development, including stents with heparin. Also, biodegradable stents were introduced as technologically closer to biological stents and engineered blood vessels. This is because the development of what can be recognized a posteriori as intermediary technological solutions facilitates de transition from the pure medical device to the adjacent markets of drugs and biologics.

According to Cohen and Levinthal (1990), using outside knowledge is largely a function of the prior related knowledge, such as scientific and technological prior knowledge in a given area. Therefore, the development of intermediary solutions that make the transition between technologies from a single market, to the intersection of adjacent markets, are very relevant. It is the “prior-related” knowledge that confers the capacity to recognized value in the “new” knowledge.

Once new solutions are introduced at the intersection of two adjacent markets, firms seem to expand within the intersection space, capitalizing on the market adjacency as a new opportunity for growth. Boston Scientific Corporation constitutes an example of how firms were able to create new market spaces and capitalize on technologies from adjacent markets to penetrate on new spaces. Boston Scientific bridged the medical device with the drug and biologic market
spaces, including features from both market spaces in developing drug-eluting stents and exploring alternative biologic-based products, such as stem cells and gene therapy. The impact the new drug-device products had on Boston Scientific growth can be observed in their revenue stream when the drug-eluting stents were released to the European and the U.S. markets in 2003 and 2004, respectively.

Using current terminology in the specialized literature about the role of collaborative patterns in corporate strategy, “open innovation” has shown to be very relevant in Boston Scientific’s strategy. Licensing and acquiring technologies and firms helped BSC to add complementary technological features to their products. The acquisition strategy followed typically in two steps: first, a co-development agreement; second, the acquisition if the technology demonstrated to be crucial for the firm pipeline. Drug-eluting stents are an illustrative example. The drug is typically licensed from a firm, supplied by a different one, while another collaborator developed the polymer (Couto et al., 2011). In addition, the same firm that licensed the drug also advised drug delivery. Following Chesbrough (2003), BSC’s use of external ideas and paths to market complemented internal ideas and facilitated advancing to new technologies. Following Rosenberg (1982), Pavitt (1998), and Brusoni et al. (2001), BSC adopted an interactive process, cross-disciplinary, and inter-organizational to develop new solutions to restore blood vessel function. In addition, BSC’s absorptive capacity for exploiting external knowledge (e.g., Cohen and Levinthal, 1990) has allowed the firm to improve their innovative capacities and establish a dominant position in the drug-eluting stent market (Couto et al., 2011).

Stakeholder engagement towards technology adjacency

We now turn to the analysis of potential stakeholders in the process of technology adjacency, looking at potential limiting factors affecting corporate change and growth.

Our evidence follows that of DeMonaco et al. (2006), suggesting that physicians played an important role in the conception, design, and adoption of the technologies and procedures to restore blood vessels function. In particular to perform angioplasty, they provided important feedback in the development of instruments and medical devices used afterwards in surgery.

Eric von Hippel (1988) extensively documents the role of surgeons and other physicians to improve and develop specialized instruments and medical devices. Our analysis has shown their key role in developing new technologies to restore blood vessel function. For example, Andreas Gruentzig worked together with firms, such as Boston Scientific, in enhancing instruments and medical devices to use in Angioplasty. Also, we documented the interaction of
an interventional cardiologist with Boston Scientific in developing the concept of drug-eluting stents. But our analysis also shows that, besides a key “source of innovation”, physicians have a crucial role in the adoption of these technologies and in building-up “trust relationships” with potential end users. The evidence is provided by the impact of the SCAAR and COURAGE clinical reports on the adoption of drug-eluting stents by physicians, as documented earlier in this paper.

Researchers and the research community per si, are another key stakeholder in developing new medical solutions to improve blood vessel function. Integrating drugs and biologics with a medical device required new ideas and skills that only research could provide. Cohen and Levinthal (1990) describe the role of employees, notably researchers in a firm, in the absorptive capacity of an organization, which is a function of individual capacities. Our analysis shows that BSC’s researchers had a crucial role, both as providers of new knowledge to the firm and as integrators of ideas and skills within the organization to make effective the transition from medical devices to drug-device combination products. The evidence documented illustrates well how BSC acquired the absorptive capacity to expand to adjacent markets. For example, the agreement with Medinol included training manufacturing and marketing and sales personnel, two skills that Boston Scientific did not have for these types of new products. Later, the firm leveraged these skills to further penetrate into the drug-eluting stents space.

In general the interaction between people and researchers associated with different start-ups and established firms shaped the ecosystem that facilitated the development of new technologies to restore blood vessel function. Medinol, Angiotech Pharmaceuticals, and Natural Pharmaceuticals are some of the few examples documented in our analysis. Their collaboration enabled the acquisition of ideas and skills that allowed the evolution of BSC from pure medical devices and its penetration into adjacent markets, such as drugs and biologics.

Regulators are also another key stakeholder in the definition of the regulatory pathway associated with the processes describes in this paper. Their analysis is beyond the context of this paper, but we note that regulators are key in the process of introducing new standards of care that combine features from two adjacent markets. As extensively discussed by Couto et al. (2011), the primary function of the drug-device combination product is defined largely by the pioneering firm and their interaction with regulators. The decision process will then influence the nature of established firms that will take from the pioneer the process of continue with the process of technological development, marketing, and distribution. The uncertainty underlying the definition of the regulatory decision may be a key factor hindering product development in new areas, such as the intersection of technologies from adjacent spaces. However, we have
shown that this is partially mitigated once the first firm has obtained clinical approval for the first of a class of combination products.

7. Summary

We have identified technology adjacency as a corporate strategy at the intersection of business development and technological innovation, with relevant impact in science-based firms. Firms developing drug-device combination products disrupted specific areas of clinical application, intersecting adjacent markets and creating new market spaces to operate.

Our evidence suggests that the development of technologies that intersect adjacent markets strongly depends on collaborative strategies, as well as on the level of stakeholder engagement, particularly of physicians. This is supported by detailed case studies associated with technologies to restore blood vessels function and Boston Scientific Corporation evolution in addressing these markets. In the case of drug-eluting stents, medical devices firms added features from the drug space into their stents, creating the drug-device combination space for cardiovascular application and expanding their market potential.

Technology adjacency is emerging in association with the increasingly relevant convergence of technologies in the life sciences, the physical sciences and engineering, requiring new expertise and the advanced training of people able to develop the combination of technologies arising from distinct areas of knowledge. Our analysis also suggests that science-based firms will tend to use technology adjacency strategically to access new markets and penetrate into areas beyond their original area of expertise in a process of continuous adaptation to market opportunities.

For example, our analysis suggest that engaging diversified stakeholders will certainly play a very important role in enabling the development of new areas of combination products, such as engineered organs or tissues. Their success depends on the crucial role physicians play in adopting the technology and build a process of societal trust with patients. For example, firms developing combination products have comparative market advantages in out-looking tissue engineering product development. They understand how to integrate distinct single-regulated components into a final product and are prepared for clinical approval. Another important consideration is that medical devices firms developing combination products have a well-established reputation among surgeons and interventional physicians that will ultimately play a decisive role in adopting engineered organs or tissues. In particular, technology adjacency is
very relevant for the potential commercialization of tissue engineering by firms at the intersection of medical devices and biologics.

Acknowledgements
We acknowledge the support of the Portuguese Science and Technology Foundation, FCT, through the MIT-Portugal Program. Daniela S. Couto was awarded with an FCT fellowship, under the MIT-Portugal Doctoral Program on Bioengineering (ref SFRH / BD / 39444 / 2007). The Luso-American Foundation, FLAD, also supported part of her work and visit to MIT through a scholarship (Ref. FLAD C&T Links 2010, N. 266 / 2010).

Manuel Heitor is in a sabbatical leave at the University of Harvard for the academic year 2011-12 under a FLAD fellowship.

References


KB Clark, WB Chew, T Fujimoto, J Meyer (1987), JSTOR. Product development in the world auto industry


W. M. Cohen and D. A. Levinthal (1990), "Absorptive capacity: A new perspective on


Chapter 8:

Teodoro Trindade and Edgar Fernandes, “Improving burner design through the spectrum analysis of flames chemiluminiscence”.
Improving burner design through the spectrum analysis of flames chemiluminescence

Teodoro P. Trindade* and Edgar C. Fernandes**

* ISEL – Instituto Superior de Engenharia de Lisboa, Lisbon, 1959-007, Portugal
** IST – Instituto Superior Técnico, Lisbon, 1049-001, Portugal
E-mail: teodoro@deq.isel.ip.pt

Keywords: Flame color characterization, Gas premixed combustion, Chemiluminescence.

ABSTRACT
The applicability of image data information to estimate the spatial distribution of equivalence ratio at the flame front on gas combustion systems are presented and discussed. The objective is to devise a procedure capable of converting a single RGB image data, obtained by conventional CCD cameras, into a reliable sensing on local combustion state for practical atmospheric flames of premixed methane and propane gases. Flame image information need to be processed and filtered in order to extract representative data whose physical interpretation is usually not straightforward, depending on the transfer function of the instrumentation system. To overcome this issue, is necessary for subsequent analysis to have an experimental reference database of flame images, obtained at controlled camera and combustion conditions. The image-sensing signal used in numerical method must exhibit adequate sensitivity to changes in combustion state; otherwise it may entail some information loss. Selecting multiple signals at the same image location enlarges data validation but also increases the computational processing time. Using two camera types, the results obtained over a wide range of practical methane/propane flame conditions (0.8 ≤ φ ≤ 1.4) are thought to support the applicability of the proposed method for flame diagnostics.

1. INTRODUCTION

The design of next generation burners for low NOx emission relies on the strategy of “a collection of small burning devices” to keep a high power output with independent device control, towards a more flexible and controllable energy system. As an example, in Figure 1 it is represented a typical multi perforated burner (heater systems) based on a metal and ceramic plates. In this case segments of these matrix can be feed with different A/F ratios and different power. The biggest challenge in the study and implementation of the miniaturized burning devices lies essentially in the fact that the length scales are all of the same order of magnitude (size of the burner-hole / flame), which enhances the sensitivity of the flame stabilization processes in construction details of the burner, and the fact that the relation S / V increases implying a greater loss of energy from the flame. Together, these two important aspects controls the thermodiffusive effects, chemical kinetics, and hence the flame stability and pollutant emissions. It is in this context that the importance of the development of non-intrusive diagnostic techniques is even more important and in this case the technique studied explores the estimation of the mixing ratio between fuel and air in a flame. The spatial detection of A/F mixture is critical since most of the new small burning devices (as to those presented in Figure 1) explores the local interaction between rich-lean flames to control the Nox emission.
Therefore, due to small size flames, the design of the burner and the analysis of the flame itself, to improve burning conditions, have to be made though the use of non-intrusive diagnostic techniques, as for example the chemiluminescence. Following the schematics of Figure 2 the next paragraphs will address the a) traditional flame's chemiluminescence technique focusing on a new spectrum signal interpretation paper in annex), and b) a procedure to analyze the spatial distribution of equivalence ratio of a flame based on a digital colored image.

2. FLAMES CHEMILUMINESCENCE THROUGH SPECTROSCOPIC ANALYSIS

The radiation emitted by premixed flames of hydrocarbon-based fuels exhibits discrete bands in the visible electromagnetic spectrum (Figure 3) [1]. These bands correspond to the spontaneous light emissions of excited intermediate radicals, formed along the kinetic combustion mechanism, when returning to a lower energy level, such as their ground state. This phenomenon, known as chemiluminescence, is closely related to the combustion state and may be interpreted as a "signature" of a particular burning condition. Relevant combustion properties such as flame temperature, heat release rate and pollutant formation, typically depends on the burning conditions reflecting the proximity of the gas mixture to its

Figure. 1 – Flame pictures in the steady state for metallic (a) and ceramic burners (b), respectively. Air mass flow $m_{\text{air}}=0.18\,\text{kg/hour}$, $\phi=1.15$ and power output $P=0.16\,\text{W}$ (propane-air mixtures)

Figure. 2 – Diagram showing the two approaches used towards flames chemiluminescence analysis: a spectroscopic technique and a digital image sensor acquisition.

In both analysis the burner experimental setup used consists of a Bunsen-type burner with an outlet diameter of 20 mm was used to generate premixed laminar methane/air and propane/air flames, as calibration references. Uniform gas velocity profiles of the combustion flow at the nozzle outlet were ensured by a higher ratio of contraction areas that equals 25. Fuel and air flows to the burner were measured during the experiments by precision mass flow controllers. High purity methane and propane (≥ 99.95%) supplied in gas cylinders were used as fuels and compressed atmospheric air as oxidizing. Experimental combustion conditions using CH$_4$ and C$_3$H$_8$ were tested at flame power of 0.75 to 1.75 kW, in a range of equivalence ratios from 0.80 to 1.40. The precision of $\phi$ value in the reference flames where estimate to be in the range of ± 0.01.
stoichiometric proportion. A widely used quantification of the mixture stoichiometry is given by the equivalence ratio \( \phi \) denoted as \( \phi \) and generically defined by Eq. (1).

\[
\phi = \frac{(\text{Air} \to \text{Fuel Ratio})_{\text{actual}}}{(\text{Air} \to \text{Fuel Ratio})_{\text{stoichiometric}}} \tag{1}
\]

A value of \( \phi = 1 \) corresponds to stoichiometric conditions, at which the reactants molar proportion of fuel to air are well balanced in such that theoretically all combustible species could be completely burned with no oxygen remaining in the products. Values of \( \phi > 1 \) represent fuel-rich conditions (excess fuel) and \( \phi < 1 \) fuel-lean condition (excess air).

![Figure 3: Overall emission spectrums of nearly stoichiometric methane/air (solid) and propane/air (dashed) laminar premixed flames (FP = 0.70 kW).](image)

Chemiluminescent flame emissions in the near-UV/visible region of the spectrum mainly corresponds to the population distribution of radicals OH* (band head at 310 nm), CH* (band head at 430 nm) and \( C^*_2 \) (Swan system centered on 473 and 515 nm) who’s responsible for the blue-green coloration of premixed flames. Herein the symbol * indicates a combined electronic/vibrational/rotational excited level above the ground energy state of the correspondent chemical specie. Varying combustion conditions the response of the kinetic oxidation mechanism alters the radical distribution producing changes in the chemiluminescent intensities and thus in the coloration exhibit by flame. These intensities are evidences of the conditions at which the combustion mechanism take place. Variations on fuel/air proportion, fuel type, temperature, pressure, air humidity, etc., influence the kinetic combustion mechanism.

**Experimental Technique**

The overall light emitted by the whole flame was collected directly through a 200 μm core diameter quartz optical fibber (P200-2-UV), with a nominal aperture of 25°. A water cooled probe, with protective quartz window is mounted at one end of the fibber and axially aligned with the burner centre. The probe was placed at 180 mm distance from the burner exit, in order to ensure the collection of the overall radiation emitted by flame, i.e. the intensity measurements are truly measures of the total chemiluminescence emitted by the flame. A spectrometer (Ocean Optics, Inc., HR4000) was used to process the incoming light signal over near-UV/visible/near-IR regions (200 to 1100 nm). Inside spectrometer, the collected radiation is separated by a 300 grooves/mm holographic grating and quantified with a linear 3648-element CCD array (Toshiba TCD1304AP), which produces an overall optical resolution around 4 nm. The data acquisition system has an A/D, 1 MHz, eight-channel board that is connected to a personal computer through an USB interface.

This configuration, with proper adjustment, allows the flame radiation sampling using integration times of about 500 ms for all flame operating conditions. Each spectrum collected, correspondent to 3648 data points, were obtained as an average of 10 consecutive intensity readings of 500 ms of integration.
time each. For each tested conditions a minimum of 50 spectrums was used to produce the average reading spectrum. The data correspondent to the flame emitted radiation was obtained by the difference (at each wavelength) between the average reading spectrum and the average environmental radiation (without flame).

The errors in current measurements arise from two separate sources, either due to the error in the gas line affecting the calculated flame operating conditions or due to perturbations in the optical collection system, affecting signal intensity. The uncertainty of flame burning conditions is mainly influenced by the fuel and air measurement accuracy. Applying the error propagation skim it is possible to indicate an accuracy lower than ±0.03 for global flame equivalence ratio (less than ±2%). The uncertainties associated to the collected radiation at relevant wavelengths (310, 430 and 515 nm) are always less than ±2.5% of signal intensity. This accuracy level was obtained using reference data as a result of a large number of collection points (over 500 spectrums).

Results

Under normal atmospheric conditions the equivalence ratio is the main parameter that influences significantly the flame emission spectrum of a particular gas fuel mixture. Figure 4 evidences this behavior, comparing the relative emission peak intensities of methane and propane flames (taken from the spectras shown in Figure 3) with reactant equivalence ratio. In this figure the peak-intensity values were scaled from 0 to 1, dividing each measured chemiluminescence over the range of ϕ by its correspondent maximum value. For CH4/air flames, the I(OH⁺) is maximal at near stoichiometric conditions, I(CH⁺) at around ϕ=1.2 and I(C2⁺) at almost ϕ=1.4. For C3H8/air flames the emission peak intensities becomes strongest at slightly higher ϕ values than for CH4/air flames, almost ϕ=1.1 for I(OH⁺) and 1.4 for I(C2⁺). The behavior of I(CH⁺) is identical for both fuel types. The high similarity in Figure 4 profiles indicates an identical influence of ϕ in the relative variation of each chemiluminescent signal produced in methane and propane flames. The gradually higher deviations in the profiles of CH4 and C3H8 flames at lower emission intensity values may be partially attributed to the different CO2 background contribution for the two fuel types.

As a consequence of the flame emission profiles like the ones presented in Figure 4, it has been developed non-intrusive methodologies based on processing flame radiation in order to detect the premixture equivalence ratio. These methods involves the use of a spectrometer that collect radiation through an optical fiber, separates light into different wavelengths and produces the spectral intensity distribution. The flame radiation can be collected globally or locally by the optical fiber. Globally means that the light from the overall flame is combined into the same
spectrum locally corresponds to radiation in a line-of-sign of a small flame area. Parsing flame global spectrum means that an average result is produced, losing information about possible variations in the spatial light distribution. On the other hand, a local measurement allows mapping the flame zones but requires a high experimental and time demanding effort.

For both, global and local spectroscopic measurements, the ratios of specific intensities produce excellent indications about changes in flame burning conditions. These radiation ratios usually associated with two chemiluminescent radical emissions, such as \( I(\text{OH}^*)/I(\text{CH}^*) \), \( I(\text{CH}^*)/I(\text{C}_2^*) \) and \( I(\text{C}_2^*)/I(\text{OH}^*) \), has been used for online monitoring of premixed combustors, mainly to evaluate reactant equivalence ratios \([28, 11, 26, 9, 7]\). Figure 5 presents experimental data of radical emission intensity ratios in \( \phi \) detection, for propane/air and methane/air premixed flames.

These types of numerical relations constitutes the base of a real time control systems enabling a continuous flow rate adjustment of the premixture gases to the burner, to meet a predefined set point of equivalence ratio.

These spectroscopic techniques have already been explored to monitor and control a large number of relevant flame parameters \([3]\) such as temperature distribution \([4]\), location of reaction fronts \([5]\) magnitude of local heat release rates \([6]\) or equivalence ratio \([7]\). The practical relevance of these light related parameters forced the development of novel monitoring techniques \([8]\), in particular involving little demanding instrumentation requirements. In principle, flame images collected using conventional 2D CCD (charge coupled device) cameras covers the data needed requirements in a significant range of combustion conditions \([9]\). As the RGB photo-sensors respond to visible wavelength and also part of the near infrared, a color flame image is therefore a combined product correspondent to a broadband radiation of local emissions. Hydrocarbon flame images combines chemiluminescent radiation mainly due to \( \text{CH}^* \), \( \text{C}_2^* \) and \( \text{CO}^* \) emitters plus black-body emission from soot particles \([10]\). The presence of soot on flames has a strong and markedly different dependency on equivalence ratio, degree of premixness and fuel type. In general, emissions from gas flames at \( \phi > 1.4 \) are dominated by diffusive effects where the black-body radiation is very intense having small chemiluminescent contributions, which restricts the method applicability. Another limitation arises at lean flame conditions where visible chemiluminescent emissions from \( \text{CH}^* \) and \( \text{C}_2^* \) tends to decrease remaining only the contribution of \( \text{CO}^* \) broad-band, whose intensity is also lowered monotonically with flame temperature. Depending on the vision instrumentation sensitivity to smaller wavelengths, the method loses efficiency on \( \phi \) estimate at flame premixture approaching \( \phi = 0.8 \). However, a method has been devised to reinterpret the entire spectrum signal based on Gaussian Mixtures approach (see paper on Annex I) to overcome the loss of sensitivity of the technique, as lean conditions are approached. This method assumes that each chemiluminescent radical present in the reaction zone, including the CO2 specie, emits light according to a Gaussian profile. Therefore each spectrum can be represented as a Gaussian summation of multiple emitters. In general, the chemiluminescence technique, based on spectral analysis of flame light emission, offers a good estimation of burning equivalence ratio by either looking at a flame point of grabbing the global flame light emission.
3. FLAMES CHEMILUMINESCENCE THROUGH DIGITAL IMAGE ANALYSIS

A different approach is discussed here to evaluate the extent to which a color digital image (recorded with a CCD conventional camera) can be used to assess the spatial distribution of equivalence ratio.

In the overall process of local equivalence ratio estimate based on RGB data processing, identification of image relevant and usable data is thought to be the main challenge. These image features do not have an intrinsic physical meaning because the transfer function of the overall instrumentation system, relating flame emission characteristics and image intensities, is unknown, serving only as relative indicators of a particular combustion state. Therefore, for practical purpose any attempt to exploit a flame image data involves a previous calibration task using flames at reference conditions for each specific setup of instrumentation hardware. Having a flame image database at different $\phi$ enables a comparison between reference signal combinations, designed here as numerical descriptors ($D$), and the correspondent value in a tested flame image to estimate the unknown burning conditions.

Figure 5: Normalized chemiluminescent intensity ratios as a function of reactant equivalence ratio of (a) CH$_4$/air and (b) C$_3$H$_8$/air flames. Lines represents data fits who’s equations are presented at the top of the graphs.

Figure 6: Stoichiometric methane/air premixed flame (1.0 kW): (a) raw color image (Jai camera at exposure time: 1000 ms); (b) gray-scale image and region of interest (ROI: ----) for color processing methodology.
Figure 7: Probability distribution of pixel intensity at the flame front ROI on reference image (stoichiometric CH$_4$/air). Jai camera data at exposure time of 1 s: (a) gray-scale intensities; (b) green channel intensity.

Recording devices

In the present study, two imaging systems were used. One representing a conventional equipment, consists on a Reflex Nikon D80 digital CCD color camera (24 bit, 3872×2592 pixels), equipped with a 35:135 mm lens. An alternative image system is a RGB color 3-CCD area scan camera (JAI CV-M9GE) having a resolution of 1024×768 active pixels per color (10 bit). The optics used was a UV/VIS 105 mm CoastalOpt SLR lens, having light transmission efficiencies higher than 85% in the near-UV/Vis range (250-650 nm).

Bayer mosaic color cameras, like the Nikon D80, uses an array pattern of color filters and an interpolation algorithm to estimate the RGB values of each pixel. This results in an averaging of color values to maintain image resolution, producing smoother transitions between adjacent pixels. The overlap of spectral responses for the red (R), green (G) and blue (B) filters also contributes to the uncertainty of some color patterns in this type of equipment. On the other hand, the 3-CCD camera technology uses dichroic prism optics to split incoming light into three separate imagers based on spectral wavelength. Each color channel has the R, G and B raw intensity value at full spectral resolution, where no signal interpolation is required.

Additionally, the steep spectral response curves resulting from the dichroic prism coating reduce crosstalk between channels, being expected an enlargement of the pixel dynamic range and less color contamination.

Figure 8: Stoichiometric methane (*) and propane (o) premixed flames image data: (points) experimental, (line) model. Calibration profiles of G:B/R descriptor value on average gray-scale intensity using: (a) Nikon camera; (b) Jai camera.

Results

It is well known that the color exhibit by an image is, among several other parameters, a function of the camera shutter-speed. Variations in flame power produce the same effect of camera shutter-speed, as they alter intensity and thus the signal distribution between R, G and B image channels. Thus, an essential step on digital flame image post-processing must be a normalization procedure to account for intensity variations in the
source object. A possible evaluation of pixel intensity can be made forming a weighted sum of the $R$, $G$ and $B$ components ($I$), which corresponds to a gray-scale conversion, Eq. (2).

$$I = r_0 R + g_0 G + b_0 B \quad (2)$$

The coefficients $c_0 = (r_0, g_0, b_0)$ on Eq. (2) can assume any type of weighted proportion, although the values used in the present study was $c_0 = (0.2989, 0.5870, 0.1140)$, which represents the average human perception of colors [11], being substantially more sensitive to green and least sensitive to blue.

![Diagram of reference image collection and processing procedure]

Figure 9: Structure of reference image collection and post-processing procedure by Matlab routines for extracting the flame calibration transfer function.

In order to evaluate the intensity/color relations, reference experimental flame images at different combustion states ($0.80 \leq \phi \leq 1.40$) was experimentally obtained using several camera exposure time, ranging from 2 s to 1 ms. Premixed CH$_4$/air and C$_3$H$_8$/air Bunsen-type flames (0.75-1.75 kW), was used. All flame images acquired are transferred and post-processed by code routines (MATLAB version 7.10.0 R2010a, The MathWorks Inc., 2010) on a personal computer in order to extract characteristic signal data. Selecting a region of interest (ROI) on each of the reference images, covering the flame front area, (b), the intensity (gray-scale, $I$) and color signal numerical descriptors $D$ are computed on every inner pixel. Being the descriptors defined by arbitrary mathematical combinations of $R$ and/or $G$ and/or $B$ values. Typical results of signals density distributions are presented on Figure 7, for gray-scale and $G$ descriptor intensities. Ideally, reference image data should assume a single value, although the experimental dispersion exhibits average standard deviation which can be as high as $\pm 8\%$. Processing images at different level of intensity, it is possible to model the dependency between $I$ and each descriptor data. Figure 8 represents experimental calibration data (points) and numerical model (lines) dependency of descriptor $D = G/B/R$ on $I$ for premixed propane and methane/air flames at $\phi = 1$.

Extending the procedure to several flame conditions, it is possible to produce of a reference frame relating each descriptor value ($D$), the average intensity ($I$) and the flame state condition ($\phi$), which only depends on fuel type. A diagram that summarizes the image collection and image color processing methodology is presented on Figure 9.

Several descriptors, made by arrangements of $R$, $G$ and $B$ levels, were evaluated by its characteristic profiles against $I$, in order to establish his adequacy on $\phi$ estimate. Figure 10 represents the Nikon camera calibration transfer function of descriptor $D = B/G$, for methane/air (Figure 10a) and propane/air (Figure 10b) premixed flames.
Dark images (low intensity values) and saturated images (higher intensities) are of little interest as the descriptor values exhibits small differences among combustion states. In general, considering Nikon camera data, the gray-scale intensities lower than 20 and higher than 225 have a limited practical use in the detection method. For a given descriptor, the best intensity range on $\phi$ detection depends on their local derivative, i.e., their rate of change with respect to $\phi$ (Figure 11). A specific descriptor $D$ has characteristic regions of interest defining areas having larger values of $\delta D/\delta \phi$, correspondent to higher sensitivities on $\phi$ detection. Considering the descriptor $B/G$ the best detection zone using Nikon camera, roughly corresponds to pixels intensity in the range $25 \leq I \leq 100$, to detect flame equivalence ratios between 1.05 and 1.25 (Figure 11). These regions depend not only on the descriptor type used but also on the sensitivity response of the imaging equipment. Higher signal descriptors amplitudes are obtained by Jai camera when compared with data from Nikon camera. This effect is observed on Figure 12 where a substantially lower and noisy descriptor signal is obtained using the single CCD Nikon camera.

Figure 10: Nikon camera transfer function diagrams of $D = B/G$: (a) CH$_4$/air, and (b) C$_3$H$_8$/air flames.

Figure 11: Nikon camera contour maps of $D = B/G$ rate of change ($\delta D/\delta \phi$) for: (a) CH$_4$/air, and (b) C$_3$H$_8$/air flames.

Figure 12: Stoichiometric premixed CH$_4$/air calibration flame data of $B/G$ descriptor collected using Jai (Δ) and Nikon (□) cameras.
In a flame image, like the one presented on Figure 13a, the pixel intensities are not constant having regions of poor detection while others have reasonable intensity. To overcome this problem of anisotropy, the equivalence ratio estimation method should apply over the same image several descriptors, preferably having complementary sensitivity regions.

The image color processing methodology numerically applies over all pixel images the calibration transfer function of each descriptor. The results are several spatial equivalence ratio distributions, one per each of the \( N_D \) descriptors. A weighted sum of all computed distributions produces an average map of \( \phi \) and represents the model estimates. The weighted method used is a local descriptor rate of change, normalized over all descriptors (Eq.3). This procedure enables an increase of relevance in the predictions made by locally most sensitive descriptors.

\[
\phi_{xy} = \frac{\sum_{i=1}^{N_D} \left( \frac{\partial D_i}{\partial \phi} \right)_{I_{xy}} \phi_{D_{xy}}}{\sum_{i=1}^{N_D} \left( \frac{\partial D_i}{\partial \phi} \right)_{I_{xy}} I_{xy}}
\] (3)

The numerical image processing technique described before was tested under rich flame conditions on images of a domestic boiler burner (Figure 12a). Their estimation results produce a consistent distribution of \( \phi \) along flame front (Figure 13d). Using this method as a diagnostic tool, that can contribute to increase the overall combustion efficiency through improvements in the burner design.

The work presented explore a way to characterize premixed gas flames using embedded information on digital flame images of atmospheric premixed methane/air and propane/air, obtained by conventional CCD cameras, as a diagnostic tool for practical gas combustion applications. The possibility of using RGB color model as a combustion sensor relay on the observation that CH* and C*\(^2\) chemiluminescence emissions are linked particularly to the average values of the \( B \) and \( G \) color image channels, at rates dependent on fuel type. The main advantages of using a conventional camera as a flame sensor rely on cost effective, readily available equipment and a possibility to interface with a computer producing a real time 2D distribution with high spatial resolution.

An image database of CH\(_4\) and C\(_3\)H\(_8\) flames was experimentally obtained under strictly controlled conditions, using two different equipment types: a Nikon camera whose RGB images are obtained by a single CCD and Bayer mosaic interpolation scheme, and a 3-CCD Jai camera raw image. A numeric post-processing algorithm, written in MATLAB\(^\circledR\) and a reference flame image calibration data, was used to compute a set of numerical parameters describing the dependency between the descriptor value \( D \), the gray-scale intensity \( I \) and the flame equivalence ratio \( \phi \), for both gas fuel type tested (CH\(_4\))

![Figure 13: Methane/air flame of a domestic boiler burner (\( N_{Reyndels} = 40 \): Nikon RGB image v.s. spatial distribution of flame equivalence ratio estimates using 3 descriptors: \( B/G, G/B/R \) and \( R/(R+G+B) \).](image-url)
The application of multiple descriptors to a flame image data enables the estimate of local equivalence ratio distribution along the flame front, roughly in a range of $\phi$ between 0.8 and 1.4.

REFERENCES

Analysis of Flames Chemiluminescence -
Reinterpretation of Spectrum Signal

F. Ferro\textsuperscript{1,3}, J. Sanches\textsuperscript{2,3} and E. C. Fernandes\textsuperscript{1,3}

\textsuperscript{1} Laboratory of ThermoFluids, Combustion and Energy Systems - IN\textsuperscript{+}
\textsuperscript{2} Institute for Systems and Robotics - ISR
\textsuperscript{3} Instituto Superior Técnico - Technical University of Lisbon

Instituto Superior Técnico
Avenida Rovisco Pais 1049-001 Lisboa PORTUGAL
Tel: 00351218417732

Correspondent author: E. C. Fernandes\textsuperscript{1} (edgar.fernandes@ist.utl.pt).

\textbf{Abstract}

In a flame light spectrum from premixed laminar hydrocarbon-air flames the peaks, and respective chemical species, that show the highest intensity are the \( \text{OH}^* \) at 308 nm, the \( \text{CH}^* \) at 431 nm and the \( \text{C}_2^* \) at 515 nm. Add to those strongest chemiluminescent emissions there is also a continuous broadband, known as background broadband, which one was firstly attributed to the \( \text{CO}_2^* \) chemiluminescence. Besides the contribution of the \( \text{CO}_2^* \) chemiluminescence to the background broadband, there are several other interesting and important chemical species, namely \( \text{CN}^* \), \( \text{NO}^* \), \( \text{NH}^* \), \( \text{CH}_2\text{O}^* \), \( \text{HCO}^* \) that can only be observed depending on the spectrometer sensitivity, spacial resolution and bandwidth. A mathematical model based on the Gaussian mixtures approach is proposed, taking into account that the behaviour of individual chemiluminescent specie in the flame spectra can be approximated to a Gaussian profile, to estimate the majority part of the chemiluminescent emissions from species involved in the flames spectrum. Add to this, it was applied a deconvolution algorithm (based on the Convolution Theorem)
to each estimated chemiluminescent emission in order to estimate the real flame spectrum. Those mathematical models were applied to premixed laminar propane \((0.8 \leq \phi \leq 1.25)\) and methane \((0.85 \leq \phi \leq 1.25)\) air flames. It was shown that the background broadband arises from the contribution of the \(CO_2^{*}\) chemiluminescence together with a strong contribution from the minor chemiluminescence species, such as \(CN^*, NO^*, NH^*, CH_2O^*\) and \(HCO^*\). A further application of such methods had shown to be a consistent way to remove the background broadband and to know the exact behaviour from \(OH^*, CH^*\) and \(C_2^*\) chemiluminescent emissions, which ones are used as indicators of, for example, heat release rate and equivalence ratio.

**Keywords:** Chemiluminescence, Flame Spectroscopy, Equivalence Ratio, Gaussian mixtures, Chemical Species.

### 1. Introduction

The light spectrum from a hydrocarbon premixed flame, in the near UV-Visible range, is interpreted as composed by a long pedestal, known as background broadband, with distinct peaks over it. The peaks, and respective chemical species, that shown the highest intensity are the \(OH^*\) at 308 nm, the \(CH^*\) at 431 nm and the \(C_2^*\) at 515 nm (the asterisk indicates a radical or excited molecule). The background broadband is normally attributed the chemiluminiscent emission of \(CO_2^*\) ([2], [9] and [22]). Based on this spectrum interpretation several studies has been performed with flames chemiluminescence signals. The chemiluminescence intensity from \(OH^*, CH^*, C_2^*\) and \(CO_2^*\) in hydrocarbon premixed flames has been used as a good indicator of chemical reaction rate, heat release rate, and equivalence ratio ([1]-[21]). In most of those studies ([12]-[19]) the \(OH^*, CH^*\) and \(C_2^*\) ”true” intensities were obtained by simply subtracting the background broadband intensity, at desired wavelength, estimated from the spectrum baseline defined by a polynomial curve fit between two arbitrary points to the left and right of the peak wavelength. This practical approach is acceptable if the major peak intensity is abnormally high when compared with they near background broadband. In lean flames, for example, the use of \(C_2^*\) is limited because the peak intensity is of the order of background broadband. In rich flames the use of \(OH^*\) is limited for the same reason. These emitters can still be used in rich or lean conditions but the uncertainty becomes considerable high.

This background broadband signal is entirely attributed to the \(CO_2^*\)
chemiluminescence emission ([2], [9] and [22]) and is considered as a *continuum band* because of the considerable number of known emitters in the $CO_2^*$ family band, their wavelength close proximity and wide range of head bands [15]. Hence, $CO_2^*$ global family band emission is taken as an indissociable group over a broad wavelength interval (300 – 600 nm) ([9] and [22]). By Child et al [23] it was stated that the shape of the $CO_2^*$ radiation curve in the post combustion zone is the same for $CO$, $CH_4$ or $C_4H_{10}$ with air mixtures and also independent of fuel-air ratio, pressure and temperature. The use of this broadband emission as a heat release markers is gaining more attention, after the work of for example Samaniego et al [2] and Najm et al. [4]. The choice of the wavelength associated with the maximum peak of this broadband varies according to different authors. For example, the maximum emissions were selected at 370 nm in [24], 412.8 nm in [9], 430 nm in [2], 450 nm in [22] and 470 nm in [8]. Recently ****

In this context, the use of chemiluminescent techniques to study/control lean flames and/or low power flames, requires a better and more systematic approach to the estimation of the true intensity emissions of $OH^*$, $CH^*$, $C_2^*$ peaks and of $CO_2^*$. The objective of this work is schematically drawn in 1, i.e one wants to decompose the measured flame spectrum into a summation of the contribution of all emitters in the family of $OH^*$, $CH^*$ and $C_2^*$ bands [1], the *continuum* $CO_2^*$ emission broadband not considered a priori as the background broadband and other minor chemiluminiscent species emitters that are generally neglected as for example, among others, $CN^*$, $NO^*$, $CH_2O^*$, $HCO^*$ and $NH^*$ (see Gaydon [1]). Then, the emission signal of each radical contributor is deconvolve of the spectrometer Optical Transfer Function, $OTF$, to generate an estimative of the original flame signal emission. This original flame spectrum is then used to explore the ratios between, for example, $OH/CH$ and equivalence ratio comparing with published results in literature.

2. Experimental Apparatus

2.1. Experimental Configuration

The premixed flame was stabilized above a convergent cylindrical nozzle (Bunsen burner type), with an exit diameter of 10 mm ($D_{exit} = 10 \text{mm}$), 18 mm ($D_{exit} = 18 \text{mm}$) and 20 mm ($D_{exit} = 20 \text{mm}$), as shown in Figure 2. The fuel used was propane and methane and the experiments were conducted at atmospheric pressure with a flame power $P = 0.37 kW$, Reynolds number
752 < Re < 930 and (0.85 ≤ φ ≤ 1.25) for propane, and P = 0.65kW, 754 < Re < 1070 and (0.85 ≤ φ ≤ 1.25) for methane at the burner with \( D_{exit} = 10\text{mm} \). For the burner with \( D_{exit} = 18\text{mm} \) the experimental conditions were for propane with \( P = 0.5kW \) and \( P = 0.7kW \), Reynolds number 857 < Re < 1201 and (0.8 ≤ φ ≤ 1.25). Finally, for the burner with \( D_{exit} = 20\text{mm} \) the experimental conditions were for methane with \( P = 0.7kW \) and \( P = 0.8kW \), Reynolds number 957 < Re < 1205 and (0.85 ≤ φ ≤ 1). Both air and fuel flow rates were controlled by calibrated flowmeters with a reading error less than 5%. All experimental conditions are listed in table 1.

2.2. Measurement Systems

The complete system for spectral light analysis, either from flame or laser beam (for calibration purposes), consists of a fiber optics (OceanOptics – QP400 – 2UV/BX) with 400 ± 8μm diameter, 0.22 ± 0.02 numerical aperture and an acceptance angle of 24.8 in air [28]. After the light passes trough the fiber optics it follows to the spectrometer HR4000 from Ocean Optics that is connected to a computer (Figure 2). The spectrometer optical scheme (see details in [28]) follows an asymmetric crossed Czerny-Turner mount with a high-sensitivity linear CCD from Toshiba (TCD1304AP), having a readout rate of 1 MHz and shutter mode. This system is microcontroller-controlled and thus all operating parameters are implemented through software interfacing to the unit via USB. The electrical performance consists on a 14 bits, 5MHz A/D converter with variable integration times from 1 s to 60 s. The HR4000 spectrometer calibration in the intensity domain was performed by using the CCD camera sensitivity response. The HR4000 spectrometer optical resolution was estimated and is in the order of the 4 nm and the stray light is < 0.05 at 600 nm and < 0.1 at 435 nm. The readout noise in a single dark spectrum is 6 counts RMS and 20 counts peak-to-peak.

In order to capture, always, the radiation from total flame surface, the fiber optics was placed at a distance above the burner top of 180 ± 1mm as it can be observed in Figure 2. In all experimental acquisitions it was used an integration time of 2500 ms to maximize the SNR before CCD pixel saturation.

For wavelength calibration purposes and determination of Optical Transfer Function of the spectrometer, the reference light source was a laser (SP2010–05 from SpectraPhysics). The laser light passes through a pinhole and get in a dark chamber in which strike into the target (the distance between the
laser and the pinhole was 1.2 ± 0.05 m, and, in the dark chamber, the distance between the pinhole and the target was 0.4 ± 0.05 m. The fiber optic was placed at 0.1 ± 0.005 m from the target (3). For this analysis, it was acquired spectrum from the same laser light at different integration times, 500, 1000, 1500 and 2000 ms.

3. Spectrometer Optical Transfer Function - OTF

The flame spectrum that came from spectrometer is affected by the optical system and spectrometer resolution, technical similar to the case of an image blurring. In this context it is applied a deblurring algorithm to each measured chemiluminescent emission (after Gaussian mixture process) to estimate the real flame spectrum, i.e. the real spectrum emission from the flame. There are several deblurring algorithms [32], but here it is used the common Convolution Theorem. Hence, denoting $f_k(x)$ and $s_k(x)$ as the measured and real contribution of the $k$ chemical specie to the flame spectrum, the relation between them could be formulated as presented in Equation (1).

$$f_k(x) = s_k(x) \ast OTF(x)$$ (1)

In Equation (1) $OTF(x)$ is the Optical Transfer Function from the spectrometer and $\ast$ is the convolution operator.

The Convolution Theorem states that the Fourier Transform of a convolution is the product of the Fourier Transforms [33], as it is presented in Equation (2).

$$FFT(f_k(x)) = FFT(s_k(x)).FFT(OTF(x))$$ (2)

The real spectrum emission from the $k$ chemiluminescent specie, $s_k$ can be obtained as presented in Equation (3).

$$s_k(x) = iFFT(\frac{FFT(f_k(x))}{FFT(OTF(x))})$$ (3)

In Equation (3) $iFFT$ denotes the inverse Fast Fourier Transform.

To use the equation (3) it is necessary to know the spectrometer Optical Transfer Function ($OTF$). The spectrometer $OTF$ is also obtained by the Convolution Theorem using 2 where $s_k(x)$ and $f_k(x)$ are now, respectively, the theoretical experimental (measured) laser light emissions for the $k$ central wavelength. The laser used was a SpectraPhysics SP2010 – 05,
which one present emissions at the wavelengths 475.5nm, 487nm, 495.5nm and 513.6nm with a theoretical laser standard deviation of 0.0022nm (data provided by the manufacturer). Each of these light signals were acquired by the spectrometer with integration times 500, 1000, 1500 and 2000ms. In Figure 4 it is shown the spectrum for the spectrometer integration time of 2000ms and also the theoretical laser spectrum that should be observed with this spectrometer. In Figure 5 it is shown the obtained spectrum for the different spectrometer integration times. As it can be observed in Figures 5 and 4, and explained in [34] the laser light emission for each wavelength has a Gaussian profile and the spectrometer spread the signal by the OTF. It is important to notice that in this study there is not performed any deconvolution to the maximum peak emissions.

In Figure 6 it is presented the OTF obtained for the each of four wavelength peaks, quite similar and independent of spectrometer integration time, and an average OTF to be used throughout. A further confirmation of the average OTF applicability was made by applying it to the measured laser spectrum (integration time of 2000ms) with results presented in Figure 7. In this Figure ones can observe plotted the theoretical, measured and deconvolved laser light spectrum. The standard deviation values between the deconvolved and the theoretical signals is of the order of 0.00029, which represents a reasonable value and provides confidence on the OTF function and method.

4. Mathematical Formulation

According to Bohr’s theory, energy is emitted when a transition occurs between an energetic level to another, in any atom or molecule. The intensity of a spectral line at a given frequency can be calculated from the number of atoms or molecules in the initial state and from the fraction of atoms or molecules in the initial state undergoing transitions to the lower state per second. However, when the light spectrum emitted by a particular atom or molecule is examined in detail, it is found that the emitted light spectrum is over a finite bandwidth. This effect is called line broadening [26]. In flames, the spectral line broadening is due to three main reasons. The first is due to the limited lifetime of individual atoms or molecules when they are in excited quantum states which one is called natural broadening. The second reason is due to collisions between particles that limits the time interval available for the particle radiate without disturbance. This is called collisional or pres-
sure broadening once the frequency of collisions is proportional to pressure. Up to now, the effect of natural and collisional/pressure collisions in the line emission broadening, can be mathematically represented by a Lorentzian function. This model is a function of Lorentzian frequency (that depends on the time period that a particle is allowed to radiate), central emission frequency due to natural broadening and pressure/temperature because of collisional/pressure effect. The third reason for the spectral line broadening in flames is the random motion of atoms or molecules. This random motion, coupled with the Doppler effect leads to the Doppler broadening, mathematically represented by a Gaussian profile [26]. When it is considered the natural or collisional/pressure and Doppler broadening simultaneously, the convolution of these two profiles leads to the Voigt profile [26]. For chemiluminescent species present in flames at low pressure (e.g. atmospheric pressure) the Voigt profile can be replaced by a more simple Gaussian profile (Gaydon and Wolfhard [31]) because the Doppler frequency is much higher than the Lorentzian frequency (Pagnini and Mainardi [30]). Therefore, in practice, the spectral emission of a single specie, resulting from the contribution of a huge number of molecules, is not a single line but a Gaussian distribution given by the central limit theorem [?]. This effect, called line broadening [26], is the theoretical justification for the observation model adopted in this work based on a mixture of Gaussians.

4.1. Gaussian Mixture

Let us consider the vector of $N$ observations, $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$, containing the spectral intensities at wavelengths $\mathbf{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, obtained from a low-resolution spectrometer. The observed spectrum is affected by measurement errors, electronic noise and external disturbances that can be jointly described by using the Additive White Gaussian Noise (AWGN) paradigm, leading to the following observation model

$$\mathbf{y} = \mathbf{x} + \mathbf{\eta}$$

(4)

where $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ is the unknown discrete spectrum without noise to be estimated and $\mathbf{\eta} = \{\eta_1, \eta_2, \ldots, \eta_N\}$ is the noise vector containing $N$ independent and identically distributed (i.i.d) zero mean normal distributed random variables with variance $\sigma^2$, $p(\mathbf{\eta}) \sim \mathcal{N}(0, \sigma^2)$. The noiseless spectrum $\mathbf{x}$ contains the contributions of all species present in the flame and therefore it can be modeled as a sum of $M$ Gaussian
functions, each one, describing the broadening spectrum of each molecule

\[ x_i = \sum_{k=1}^{M} \omega_k \phi_k(\lambda_i) \]  

(5)

where \( \lambda_i \) is wavelength associated with the \( i^{th} \) sample on the original noisy spectrum, \( y \). \( \omega = \{\omega_1, ..., \omega_k, ..., \omega_M\} \) is the vector of weights where \( \omega_k \) is the unknown weight associated with the \( k^{th} \) specie present in the flame and \( \phi_k(\lambda) \) is the corresponding normalized, \( \int \phi_k(\lambda) d\lambda = 1 \), broadening spectrum,

\[ \phi_k(\lambda) = e^{-\frac{(\lambda - \tau_k)^2}{2\sigma_k^2}} \]  

(6)

which is a Gaussian function centered at the wavelength \( \tau_k \) characterized by the spread parameter \( \sigma_k \).

The 3M parameters \( \omega = \{\omega_k\}, \tau = \{\tau_k\} \) and \( \sigma = \{\sigma_k\} \) are unknown variables that must be estimated from the original data, \( y \).

This inverse problem is highly ill-posed, mainly, because there is a huge number of different sets of parameters \( \{\omega, \tau, \sigma\} \) that lead to the same solution \( x \). Therefore regularization is needed by introducing in the estimation process a priori knowledge about the parameters in order to guide the solution toward realistic solutions from a physical point of view. The a priori information is obtained from the literature and it is mainly related with the expected central frequency associated with each specie, \( \tau_k \), and the relative magnitude of the different species, \( c_{kr} = \omega_k/\omega_r \).

4.2. Estimation

The parameter estimation problem can be formulated as the following optimization task

\[ \hat{\theta} = \arg \min_{\theta} E(y, \theta) \]  

(7)

where \( \theta = \{\omega, \tau, \sigma\} \) and

\[ E(y, \theta) = \frac{E_y(y, \theta)}{\text{Data Fidelity term}} + \alpha_\omega E_\omega(\omega) \]  

(8)

is an energy function composed by two terms; the first, called data fidelity term, pushes the solution toward the data, and the other, called prior terms,
introduce a priori knowledge about the $\omega$ parameters in order to guide the solution. The prior hyperparameter $\alpha_\omega$ is used to tune the strength of the prior by imposing more or less constraint in estimation of the parameters. Small values of $\alpha$ impose less constraint in the estimation of the parameters which make the estimated values more dependent on the observations and can lead to large error due the noise corrupting the data. In the limit, if $\alpha_\omega = 0$, the solution depends totally on the data and the problem, as referred before, is ill-posed. Large values of $\alpha_\omega$, on the contrary, make the estimated values less dependent on the noisy observations but can lead to biased solutions. If $\alpha \to \infty$ the solution depends only on the prior and converges to the expected value of the prior distribution. The value used in this work for the parameters $\alpha_\omega$ is obtained by linear search method.

The data fidelity term, in the assumption of an AWGN model for the observations, is

$$E_y(y, \theta) = -\log \prod_i p(\eta_i)$$

and from (4), $\eta_i = y_i - x_i$, and (5) it is

$$E_y(y, \theta) = -\frac{1}{2\sigma_y^2} \sum_{i=1}^{N} \left( \sum_{k=1}^{M} \omega_k \phi_k(\lambda_i) - y_i \right)^2 + C$$

where $C$ is a constant. By using matrix notation, (10) may be written as follows

$$E_y(y, \theta) = \frac{1}{2\sigma_y^2} (\Phi \mathbf{w} - y)^T (\Phi \mathbf{w} - y)$$

where

$$\Phi = \begin{pmatrix} \phi_1(\lambda_1) & \phi_2(\lambda_1) & \phi_3(\lambda_1) & \ldots & \phi_M(\lambda_1) \\ \phi_1(\lambda_2) & \phi_2(\lambda_2) & \phi_3(\lambda_2) & \ldots & \phi_M(\lambda_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi_1(\lambda_N) & \phi_2(\lambda_N) & \phi_3(\lambda_N) & \ldots & \phi_M(\lambda_N) \end{pmatrix}$$

is a $N \times M$ matrix.

The prior terms are assumed to be the following logarithmic Gaussian distributions

$$E_\omega(\omega) = \frac{1}{2\xi_\omega^2} \sum_k (\omega_k - \omega_{0k})^2 + C = \frac{1}{2\xi_\pi} (\omega - \omega_0)^T (\omega - \omega_0) + C$$
where $\xi_\omega$ models the uncertainty associated with the parameters $\omega$ (the uncertainty is assume to be the same for all parameters $\omega_k$).

The expected central wavelengths, $\tau_0$ and corresponding width, $\sigma_0$, of the broadening Gaussian spectrum emitted by each specie are obtained form the literature [1, 27, 25].

The overall energy function is therefore the following

$$E(y, \theta) = (\Phi w - y)^T (\Phi w - y) + \alpha_\omega (\omega - \omega_0)^T (\omega - \omega_0)$$

where $\alpha_\omega = \frac{\sigma_\omega^2}{\xi_\omega}$ is obtained by linear search method.

### 4.3. Optimization

The parameter estimation is performed by finding the stationary point of the energy function (14) with respect the parameters

$$\nabla_\theta E(y, \theta) = 0$$

according the following strategy

$$\tau^t = \arg \min_{\tau} E(y, \tau, \sigma^{t-1}, \omega^{t-1})$$

$$\sigma^t = \arg \min_{\sigma} E(y, \tau^t, \sigma, \omega^{t-1})$$

$$\omega^t = \arg \min_{\omega} E(y, \tau^t, \sigma^t, \omega)$$

The minimization of (14) with respect $\tau$ and $\sigma$ is performed by the gradient descendent iterative method [?] and the minimization with respect $\omega$ is performed non iteratively by using the close-form solution of the equation $\nabla_\omega E(y, \tau, \sigma, \omega) = 0$. The minimization procedure is performed in the following three steps

$$\tau^t = \tau^{t-1} - d_\tau \nabla_\tau E(y, \tau^{t-1}, \sigma^{t-1}, \omega^{t-1})$$

$$\sigma^t = \sigma^{t-1} - d_\sigma \nabla_\sigma E(y, \tau^t, \sigma, \omega^{t-1})$$

$$\omega = (\Phi^T \Phi + \alpha_\omega I)^{-1} (\Phi^T y + \alpha_\omega \omega_0)$$

where $^t$ denotes iteration t, $I$ the identity matrix, $d_\tau$ and $d_\sigma$ are updating steps found by linear search method. $\Phi(\tau, \sigma)$ in equation (19) is computed with the updated estimates of the parameters $\tau$ and $\sigma$, $\tau^t$ and $\sigma^t$. 
4.4. Method Implementation and Results

From up to now, it was shown that the flame spectra can be interpreted as a sum of spectrum from all the species involved in the flames chemiluminescence, which could be expressed according to Equation (5). In this work our intend is to estimate the parameters $\omega_k, \tau_k$ and $\sigma_k$ that minimizes the error $\eta$ on Equation (4) and it was hown that the parameters estimation can be performed by Equations 19, 20 and 21.

The species selected to perform the Gaussian Mixing procedure are presented in table 2 with the respective vibrational quantum number and associated wavelength and includes all the emissions listed in [1], the major chemiluminescent emissions in flames due to $OH^*$, $CH^*$ and $C_2^*$ together with other important vibrational transitions from $OH^*$ family band as, for example, $OH^*(2,2)$, $OH^*(3,3)$ and $OH^*(4,4)$ referred on [1] and [27], and from $CH^*$, $C_2^*$, $CN^*$, $NO^*$ and $NH^*$ ([1] and [25]). The contribution of $CO_2^*$ specie is a special case because the considerable number of emitters in the $CO_2^*$ family band, the close proximity between them and wide range of head bands [15] justifies the continuum and major part of the background broadband. Hence, $CO_2^*$ global family band emission will be taken here as an indissociable group represented by $CO_2^*$.

Before the parameters estimation it is needed a pre-estimation process, which was performed as the follows:

- Considering a fix $CO_2^*$ family band emission with wavelength and intensity values taken on each spectrum at the region of 450nm (as it was proposed by Gaydon [1] and Child e Wohl [23]). The fix standard deviation was calculate for a Gaussian function center near 450nm with a quasi zero values closest to 270nm or 570nm.

- After to fixing the $CO_2^*$ family band emission, it was proceeding to the parameters estimation from all other individual electronic and/or vibrational transitions. On those estimation it was need an initial parameter values, as can be observed on Equation 19. The considered values were:

  1. The starting wavelength was taken from Table 2
  2. The standard deviation was initialized by Delta Dirac Functions.
  3. The intensity values were initialized with points taked directly from the spectrum but with an additional restriction that the ratio
between intensities of family band members must obey to values given by ([1], [27] and Straw [25]), known as a priori knowledge. As an example, from [1] the ratio between $OH - Q_2$ branch (307.8) and $OH - R_2$ branch (306.36) branch is 0.5.

On this pre-estimation process it is important to notice that the parameters estimation was performed exactly according to the order of Equations 19, 20 and 21, i.e., it was firstly estimated the $\sigma$ and $\tau$ values and hence those ones were placed on $\Phi$ (described on Equation 12) presented on Equation 21. After the pre-estimation, the process is then free to adjust the final species wavelength, standard deviation and intensity to all chemiluminescent species, including the special case of the continuum $CO_2^*$ family band. A stop criteria used for the $t$ iteration was such that, $||\tau^t - \tau^{t-1}||$, $||\sigma^t - \sigma^{t-1}||$ and $||\omega^t - \omega^{t-1}||$ should be less or equal to 0.001.

A typical result of this mathematical procedure is presented in Figures 9 for a premixed laminar propane-air (a) and methane-air (b) flame with $\phi = 1$. In these figures ones can identify the measured spectrum, each specie and the $CO_2$ Gaussian emissions and the summation of all contribution. The results are remarkable since the Gaussian mixture fits the measured spectrum. The fitting quality is measured here by a global RMS given by $\eta$ of 4.

The overall RMS is defined as $RMS = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2 / N}$, where $x_i$ is the signal after processing this method and $y_i$ the observed flame signal. This procedure was extended to all experimental conditions studied in this paper (table 1) and the overall $RMS$ is presented as a function of equivalence ratio, for each fuel, in Figure 10(a). As it can be observed in Figure 10(a) the mean $RMS$ value is 6, and relatively constant along the equivalence ratio.

In addition the fitting quality was extended to analyse the deviation in estimated wavelength and estimated intensity. In Figure 10(b) it is presented the wavelength percentage error that was determined by $\epsilon_\lambda = \frac{||\lambda_{theoretical} - \lambda_{estimated}||}{||\lambda_{theoretical}||} \times 100$ in order to compare the theoretical wavelength values ([1] [27] and [25]) with the estimated ones. As it can be observed in Figure 10(b) the mean value is of the order of 0.2% and is relatively constant along the equivalence ratio.

The estimated and theoretical intensity values ([1], [27] and [25]) was also compared but the values were reduced to the range between 1 and 10. In 10(c) it is presented the percent error in the intensity domain as a function of equivalence ratio, with values of the order of 2%. The error is quantitatively small and again is relatively constant along the equivalence ratio.
All together, the results suggest the robustness of the Gaussian mixture modelling for propane and methane air flames. The estimation process had a reasonable concordance between the theoretical and estimated parameters and also from experimental and estimated flames spectrum.

5. Estimation of 'Real' Flame Spectrum

The deconvolution of the estimated flames spectrum (after Gaussian mixture decomposition) in order to get an estimation of the real flames spectrum is based on Equation 3, since $OTF$ was already determined. In this inverse process all specie emissions were deconvolved except the $CO_2^*$ broadband. Note that our main interest is in the strongest emission of main species ($OH^*, CH^*$ and $C_2^*$) to be used in practical applications. In Figures 11 it is presented the deconvolve flame spectrum for a premixed laminar propane-air (a) and methane-air (b) flame for $\phi = 1$ respectively, unless a multiplicative value. Observing those Figures it is important to notice that the profiles from the various specie chemiluminescent emissions are almost like a linear single peak emission. Exploring in detail the 308nm region in the premixed laminar propane-air flame, as presented in Figure 12, it is observable that the deconvolve flame spectra shows an almost discrete contributions from the $OH^*(0,0)$ and also the $OH^*(1,1)$ with also the various rotational transitions associated. It is presented the $OH^*(0,0)R2$, $R1$, $Q1$ and $Q2$ branches as also the $OH^*(1,1)R2$ and $Q2$ branches.It is also important to notice that the results are quite similar to those ones obtained with a high resolution spectrometer at [13] and [21] for methane and propane air flames respectively, where it is presented $R$ and $Q$ branches from the $OH^*(0,0)$ and $OH^*(1,1)$.

This can be also observed for the $CH^*(0,0)$, $C_2^*(1,0)$ and $C_2^*(0,0)$ as in Figures 13 (a), (b) and (c) respectively, for the premixed laminar propane air flame. In Figure 13(a) it is shown the $CH^*(0,0)R2$ peak, the chemiluminescent emission associated to the $CH_2O^*$ at 424.28nm and also the $CN^*(0,1)$ band and the respectives $CN^*(1,2)$ and $CN^*(2,3)$. In Figure 13(b) it is observable the $C_2^*(1,0)$ band with the other vibrational transitions (2, 1), (3, 2), (4, 3), (5, 4) and (6, 5). In Figure 13(c) it is seen the $C_2^*(0,0)$ band with the other vibrational transitions (1, 1), (2, 2), (3, 3) and (4, 4) and it is possible to observe a quite similarity between those bands and that ones acquired by [13].
The knowledge of the chemiluminescent intensities from $OH^*$, $CH^*$, $C_2^*$ in hydrocarbon premixed flames has several applications, namely, the equivalence ratio estimation. From several studies, the ratios from $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ chemiluminescence were used to estimate the local equivalence ratio [11]-[21]. In the major part of those studies ([12]-[19]) it was used a quadratic fit in order to extract the influence of the background broadband in the band of interest ($OH^*$, $CH^*$ or $C_2^*$). In the following one goes to find a correlation between the chemiluminescent ratios from $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ and the equivalence ratio from laminar premixed propane and methane air flames. This correlation will be performed in three distinct methods as follows:

- **Method I**: to considering the intensity of a single emitter after deconvolution proposed in this work, i.e. from $OH(0,0)R1^*$, $CH(0,0)R1^*$ and $C_2(0,0)^*$. Note that this intensity is free of any co-lateral contribution and from background broadband (see Figure 14).

- **Method II**: to considered the difference between the maximum peak emission and the background broadband determined by a quadratic fit interpolation, as suggested by other authors ([12]-[19]). Note that this procedure is taken over the measured spectrum (see Figure 14).

- **Method III**: to consider just the maximum emission peak observed in the measured spectrum without any correction (see Figure 14).

### 6.1. Propane Air Laminar Premixed Flame

The acquired data from laboratory experiments for laminar premixed propane air flames were for a flame power range from 0.37 to 0.7 kW, Reynolds number $752 < Re < 1201$, and equivalence ratio range ($0.8 \leq \phi \leq 1.25$). In the following one goes examine the intensity ratio dependence on the equivalence ratio.

In Figures 15 are shown the evolution of the $OH^*/CH^*$ (a), $C_2^*/OH^*$ (b) and $C_2^*/CH^*$ (c) ratios over the equivalence ratio, respectively, for the three methods. In this Figures 15 it is also shown the respective linear fit (the associated error is not shown, once this one is the same order of length of the various data points).
Observing Figures 15 (a) and (b), it is seen that the $OH^*/CH^*$ and $C_2^*/OH^*$ had an analogous behaviour against the equivalence ratio in the various methods. In the case of the $OH^*/CH^*$ ratio it is notice that their behaviour by Method I is unless a constant multiplicative value of the Method II and III. If ones look at the $C_2^*/OH^*$ it is observable here that the values from the various Methods are very closest, however, Method I presents an higher slope, which implies an improvement on the equivalence ratio estimation. In Figure 15 (c) it is shown the dependence of the $C_2^*/CH^*$ ratio on the equivalence ratio. It is observable that the $C_2^*/CH^*$ ratio by Method I is a good indicator of the equivalence ratio, once by Method II and Method III, for lean flames this value tends to be almost constant. In [21] it is shown the various intensity ratios against the equivalence ratio at the flame front and also the $OH^*/CH^*$ at the Anchor point for a laminar premixed propane air flame. The experiments performed in this paper were for flame conditions quite similar to ours, namely, Reynolds number range, burner type (12 mm), equivalence ratio range and flow velocity. It was studied a point located at the reaction zone and the Anchor point of the flame, and those were acquired by a Cassegrain optic coupled to a high resolution spectrometer [21]. In this paper, it was subtracted the influence of the background broadband to the maximum $OH^*$, $CH^*$ and $C_2^*$ intensities, and this broadband was solely attributed to the $CO_2^*$. In Figure 16 it is presented the comparison between the intensity ratios along the equivalence ratio by the proposed method and the results obtained by [21]. In Figure 16 (a) it is observable that the $OH^*/CH^*$ ratios had different behaviours along the equivalence ratio in Method I, at the Anchor Point and in a point of the flame in the reaction zone by [21]. The main difference can be due to two factor, one, to the fact that in our experiments it is acquired all the flame area, and consequently it is obtained the spectrum of an integrated image, the global flame area, on the contrary, in [21] it is obtained the spectrum of just one flame point, which one has a greater $OH^*$ concentration compared with the other two radicals $CH^*$ and $C_2^*$. The second factor is due to the fact that in Method I it is obtained the $OH^*(0,0)$ major chemiluminescent emission and in [21] this one has the influence of the other $OH^*(0,0)$ radicals (which makes the major emission much higher than it is in reality). It is also seen that by method I it is obtained an higher line slope, which means that the spectrum acquired from the global flame area presents an improvement on the equivalence ratio estimation for laminar premixed propane air flames. In Figure 16 (b) it is seen that by [21] it wasn’t possible to obtain an exact relationship between the $C_2^*/OH^*$
and $C_2^*/CH^*$ intensity ratios and equivalence ratio, but by Method I it is possible to estimate the equivalence ratio by those intensity ratios. This can be due what was explained before and also to the possibility to obtain the $C_2^*$ intensity for lean flames, which gives an higher line slope.

6.2. Methane Air Laminar Premixed Flame

For laminar premixed methane air flames, the experimental conditions were for a flame power range from 0.65 to 0.8 kW, Reynolds number $754 < Re < 1204$, and equivalence ratio range ($0.85 \leq \phi \leq 1.25$). In Figures 17 are shown the evolution of the $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ ratios over the equivalence ratio, respectively.

Observing Figure 17 (a) it is seen that the $OH^*/CH^*$ ratio variation along the equivalence ratio is quite similar in both three methods, and this characteristic is also observed on the $C_2^*/OH^*$ (Figure 17 (c)). On the other hand, it is observed that Method I presents an higher line slope than the other Methods, resulting that this process of flame spectrum analysis improve the equivalence ratio estimation based on the $OH^*/CH^*$ and $C_2^*/OH^*$ ratios. In Figure 17 (c) it is reinforced the importance of this method on the equivalence ratio estimation by the $C_2^*/CH^*$ ratios, i.e., next to lean and/or stoichiometric conditions by Method II and III, the $C_2^*/CH^*$ ratios tends to a constant value and by the presented Method it is presented a major slope on the line of the $C_2^*/CH^*$ intensity ratios against equivalence ratio. In [7] it was measured the chemiluminescent emissions from $OH^*$, $CH^*$ and $C_2^*$ and also the continuous emissions from $CO_2^*$ in natural-gas fuelled premixed counterflow flames from a reaction zone volume with a Cassegrain optics coupled to a spectrometer for $0.7 \leq /\phi \leq 1.3$. It was shown that those chemiluminescent intensities are a good indicator of heat release rate, except the emissions from $C_2^*$ and also that the $OH^*/CH^*$ intensity ratio is a good indicator of equivalence ratio but the $C_2^*/CH^*$ wasn’t once it presents a non monotonic dependence on equivalence ratio (for lean flames it tends to be constant). In [13] it was measured spatially and spectrally resolved chemiluminescence in the reaction zone of a laminar premixed methane/air flame ($0.9 \leq /\phi \leq 1.5$) by using a Cassegrain system and an high resolution spectrometer. It was shown strong correlations between the peak intensity ratios of $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ on the reaction zone to the equivalence ratio. Also with a Cassegrain optical system and a spectrometer, in [18] it was obtained a correlation between the $OH^*/CH^*$ and $C_2^*/OH^*$ intensity
ratios to the equivalence ratio on the reaction zone of a partially premixed methane swirling flame \((0.8 \leq \phi \leq 1.4)\).

In Figure 18 (a) (b) and (c) it is compared the various intensity ratios along the equivalence ratio obtained by the proposed method and those obtained by the latest mentioned authors, but, of course, it is important to notice that those results are extrapolated to our experimental conditions. Observing 18 (a) at a first analysis, it is seen that the \(OH^*/CH^*\) intensity ratio along the equivalence ratio by Method I has an higher slope, and this can be explained as it was for propane, it is only considered the major emission from the \(OH^*(0,0)\) and there is no contribution from the collateral neighbours (namely the other \(OH^*(0,0)\) radicals).

In Figure 18 (b) the \(C_2^*/OH^*\) intensity ratio along the equivalence ratio it is seen that the line slope obtained by Method I is higher than the one presented by [13] and [18], which means that it is really important if ones taken into account all the band emission (even with no influence of the background broadband), or if ones just taken into account the major emission from the chemiluminescent specie. This fact is really observed on the evolution of the \(C_2^*/CH^*\) in [7] and [13] along the equivalence ratio. Comparing both authors it is seen that this evolution is almost the same unless a multiplicative constant value, and for lean conditions this tends to a constant value. On the contrary, by Method I it is observable that the \(C_2^*/CH^*\) can be a good indicator of the equivalence ratio.

The method proposed on this paper reinforce what was performed before, the \(OH^*, CH^*\) and \(C_2\) background broadband removed by a quadratic fit interpolation or with any correction, it also give ones another prospects in order to understand which are the major contributions to the flame spectrum and with this one it is possible to use the \(C2^*/CH^*\) ratio for the equivalence ratio estimation, once in lean flames it doesn’t tends to an almost constant value.

7. Summary and Conclusions

The behaviour of individual chemiluminescent specie in the flame spectra, at atmospheric pressure conditions, can be approximated to a Gaussian profile [31]. In the present paper it was developed a mathematical model in order to estimate the majority part of the chemiluminescent species involved in the flames spectrum obtained from a low resolution spectrometer. This mathematical model was based on the assumption that the final spectrum can
be given by the contribution of all individual chemiluminescence emission, which has a Gaussian profile. Therefore, this model can be defined as a Gaussian mixture. This mathematical model was verified for premixed laminar hydrocarbon air flames, namely for, premixed laminar propane and methane air flames. The estimation of individual chemiluminescence specie was performed by incorporating the uncertainty about its central wavelength, intensity and spreading. It was estimated the spectra from the major (attributed to $OH^*$, $CH^*$, $C_2^*$ and $CO_2^*$ and minor (attributed to $NO^*$, $NH^*$, $HCO^*$, $CH_2O^*$ and $CN^*$) chemiluminescence emissions in the flame spectrum. It was observed that the estimated wavelengths and intensities from the major (from $OH^*$, $CH^*$, $C_2^*$) and minor (from $NO^*$, $NH^*$, $HCO^*$, $CH_2O^*$ and $CN^*$) chemiluminescence emissions are in agreement with the theoretical values [1]. It was verified that the mathematical approximation to the flame spectrum by this method represents an uncertainty very small. This is real important once with this method it is possible to estimate the spectral behaviour of all the molecules involved in the flame spectrum. It was also show, from the Convolution Theorem the possibility to know the real flames behaviour by simple optical means.

One application of such method was on the analysis of the amount of signal that is superimposed in one chemiluminescent emission by the adjacent neighbours. In fact it was shown that it is possible to discriminate all individual chemiluminescent contributions as it was as an artificially improve of the spectrometer spatial resolution. In the spectrum background analysis, i.e., from the knowledge of the $CO_2^*$ and minor chemiluminescent species emissions it was possible to know the exact behaviour of the background broadband and how this one affects the major emissions, of course those ones becomes completely different from those proposed by other authors ([12]-[19]).

It was also estimated the equivalence ratio by the $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ intensity ratios for the two fuel type. This was performed in order to compare the three methods, namely if it was considered the main emitter from $OH^*$, $CH^*$ and $C_2^*$, i.e., the peaks attributed to the transitions $OH(0, 0)R1^*$, $CH(0, 0)^*$ and $C_2(0, 0)^*$, if it was performed a background broadband correction by the background broadband correction as suggested by other authors or, finally, if it wasn’t any kind of correction. In the analysis of laminar premixed propane and methane air flames the estimation of the equivalence ratio by using all methods seems to be reasonable, and this method presents a novelty in the use of the $C_2^*/CH^*$ ratios on the equivalence.
ratios estimation, once at lean and/or stoichiometric conditions it shown a better linearity.

8. Acknowledgements

Special gratitude to Professor Teodoro Trindade for the availability, and for the help to advance forward during this work. Acknowledgements also to the funding which made this research possible. The author F. Ferro and this work has been sponsored by ”FCT“, under grant PTDC/EME-MFE/68830/2006.

References


unstable combustion in lean premixed combustor", AIAA-2001-0787 , 39

and turbulent propane-air and ethylene-air flames", National Advisory

measuring primary zone equivalence ratio in gas turbine combustors",
SPIE Conference on Advanced Sensors and Monitors for Process Indus-
tries and Environment, Boston, Massachusetts, November 1998, SPIE

[11] Roby, R. J., Reaney, J. E., and Johnsson, E. L., "Detection of Tem-
perature and Equivalence Ratio in Turbulent Premixed Flames Using
Chemiluminescence", Proceedings of the 1998 Int. Joint Power Genera-

spectra measurement in high-pressure premixed laminar flame", AIAA-
2002-0191, 40 Aerospace Science Meeting and Exhibit, Reno, Nevada,

of OH*, CH*, AND C2 Chemiluminescence in the Reaction Zone of
Laminar Methane/Air Premixed Flames", Proceedings of the Combustion
Institute, Volume 28, 2000/pp. 1757-1764

[14] Y. Ikeda, J. Kojima and H. Hashimoto, "Local chemiluminescence spec-
tra measurements in a high-pressure laminar methane/air premixed
1495-1501

and J.M. Seitzman, "Chemiluminescence Based Sensors for turbine
Engines", AIAA-2003-4490 , 39 Joint Propulsion Conference & Exhibit,
Huntsville, Alabama, July 2003.

B.W. Roming and J.M. Seitzman, "Optical equivalence ratio sensors for


[28] HR4000 and HR4000 CG-UV-NIR Series High-resolution Fiber Optic Spectrometers - Installation and Operation Manual - Ocean Optics


9. Figures and Tables
Figure 1: Global scheme defining the work layout objective - an estimative of flame 'true' spectrum emission

Figure 2: Experimental setup
Figure 3: Laser analysis setup

Figure 4: Experimental laser spectrum for the spectrometer integration time 2000ms and respective theoretical spectrum
Figure 5: Spectrum from the Laser (SP2010 – 05 from SpectraPhysics) acquired at different spectrometer integration times, namely 500, 1000, 1500 and 2000 ms

Figure 6: Spectrometer OTF obtained for each wavelength emission and respective mean spectrometer OTF
Figure 7: Laser spectrum after the signal deconvolution, spectrometer laser spectrum and theoretical laser spectrum.

Figure 8: Description of the parameters used in the first step for the \( CO_2^* \) initialization, notice that \( \sigma_{CO_2} = \frac{(x_{CO_2} - x_0)^2}{2ln(\frac{y_{CO_2}}{y_0})} \), \( \omega_{CO_2} = y_{CO_2} \) and \( x_{CO_2} \approx 450\text{nm} \) (light spectrum from a laminar premixed propane air flame \( \phi = 1 \) and \( P = 3.7\text{kW} \)).
Figure 9: Measured and estimated flame spectra with the real $CO_2^*$ broadband and the individual Gaussians for a premixed laminar (a)- Propane-air flame with $\phi = 1$, $P = 0.37KW$ and $d = 10mm$ and (b)- Methane-air flame with $\phi = 1$, $P = 0.65KW$ and $d = 10mm$. 
Figure 10: (a)- Overall RMS for each fuel condition and respective equivalence ratio values, the overall RMS is defined as $RMS = \sqrt{\frac{\sum_{i=1}^{N} (f(x_i) - y_i)^2}{N}}$ where the $f(x_i)$ is the estimated signal at a certain spectra point $x_i$, $y_i$ is the measured signal and $N$ is the total number of spectra points. (b)- Percent Error in the wavelength domain determined for all the experimental conditions, it is important to notice that the Percent Error is defined as $\epsilon_{\lambda} = \frac{||\Lambda_{\text{Theoretical}} - \Lambda_{\text{Estimated}}||}{||\Lambda_{\text{Theoretical}}||} \times 100$, where $\Lambda_{\text{Estimated}}$ are the estimated wavelength values and $\Lambda_{\text{Theoretical}}$ are the theoretical wavelength values presented in [1] [27] and [25]. (c)- Percent Error in the intensity domain determined for all the experimental conditions
Figure 11: Estimated flame spectra, deconvolve spectrum and respective individual deconvolve Gaussian Functions for a premixed laminar (a)- Propane-air flame with \( \phi = 1 \), \( P = 0.37KW \) and \( d = 10mm \) and (b)- Methane-air flame with \( \phi = 1 \), \( P = 0.65KW \) and \( d = 10mm \).
Figure 12: Estimated flame spectra, deconvolve spectrum and respective individual deconvolve Gaussian Functions of the $OH^*$ band from a premixed laminar propane-air flame with $\phi = 1$, $P = 0.37KW$ and $d = 10mm$. 
Figure 13: Estimated flame spectra, deconvolve spectrum and respective individual deconvolve Gaussian Functions of the (a) $CH^*(0,0)$ band, (b) $C_2^*(1,0)$ band and (c) $C_2^*(0,0)$ band from a premixed laminar propane-air flame with $\phi = 1$, $P = 0.37KW$ and $d = 10mm$. 

31
Figure 14: Used methods in the peaks study and respective designations (this \(OH^*\) band spectrum is from a premixed laminar propane-air flame with \(\phi = 1\), \(P = 0.37KW\) and \(d = 10mm\)).
Figure 15: Exponential dependence of the $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ ratios on the equivalence ratio for a laminar premixed propane air flame.
Figure 16: Comparison between the (a) $OH^*/CH^*$, (b) $C_2^*/OH^*$ and (c) $C_2^*/CH^*$ ratios against the equivalence ratio for a laminar premixed propane air flame obtained by Method I and by Ikeda et al. [21]
Figure 17: Exponential dependence of the $OH^*/CH^*$, $C_2^*/OH^*$ and $C_2^*/CH^*$ ratios on the equivalence ratio for a laminar premixed methane air flame.
Figure 18: Comparison between the $OH^*/CH^*$, $C_2*/OH^*$ and $C_2*/CH^*$ ratios against the equivalence ratio for a laminar premixed propane air flame obtained by Method I and by Hardalupas et al [7], Kojima et al [13] and Cheng et al [18]
Table 1: Experimental conditions for laminar premixed Propane and Methane air flames

<table>
<thead>
<tr>
<th></th>
<th>$U_{eq}(0.5, \text{mm})$</th>
<th>Flame Power (0.05, kW)</th>
<th>$\phi$</th>
<th>$Re_{1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Propane Air Flames</strong></td>
<td>0.5</td>
<td>0.35</td>
<td>957</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.35</td>
<td>910</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.9</td>
<td>577</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1034</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>1</td>
<td>921</td>
<td>1.05</td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td>810</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.2</td>
<td>780</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>752</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Methane Air Flames</strong></td>
<td>0.5</td>
<td>0.85</td>
<td>802</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.9</td>
<td>1054</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>957</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.95</td>
<td>2205</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.85</td>
<td>1070</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>1014</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.95</td>
<td>966</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>821</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.05</td>
<td>882</td>
<td>1.1</td>
<td>845</td>
</tr>
<tr>
<td></td>
<td>1.15</td>
<td>812</td>
<td>1.2</td>
<td>792</td>
</tr>
<tr>
<td></td>
<td>1.25</td>
<td>754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical Species</td>
<td>l (nm)</td>
<td>Chemical Species</td>
<td>l (nm)</td>
<td>Chemical Species</td>
</tr>
<tr>
<td>------------------</td>
<td>--------</td>
<td>------------------</td>
<td>--------</td>
<td>------------------</td>
</tr>
<tr>
<td>CO2</td>
<td>450</td>
<td>CH2</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>HCO system</td>
<td>278.15</td>
<td>NH2</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>HCB system</td>
<td>278.84</td>
<td>NH3</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>HCO system</td>
<td>279.71</td>
<td>NH3</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>OH(1) R2</td>
<td>287.13</td>
<td>HCO.H2</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>OH(1) R1</td>
<td>287.13</td>
<td>HCO.H2</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>OH(1) Q1</td>
<td>287.13</td>
<td>HCO.H2</td>
<td>341.7</td>
<td>CH3</td>
</tr>
<tr>
<td>N0(0,5) R2</td>
<td>284.88</td>
<td>HCO.B system</td>
<td>356.52</td>
<td>C2</td>
</tr>
<tr>
<td>HCO system</td>
<td>284.88</td>
<td>HCO.B system</td>
<td>356.52</td>
<td>C2</td>
</tr>
<tr>
<td>NH(2,1) R2</td>
<td>287.53</td>
<td>CN(1,0)</td>
<td>356.52</td>
<td>C2</td>
</tr>
<tr>
<td>NH(2,1) Q1</td>
<td>287.53</td>
<td>CN(1,0)</td>
<td>356.52</td>
<td>C2</td>
</tr>
<tr>
<td>HCO system</td>
<td>294.92</td>
<td>HCO.B system</td>
<td>383.56</td>
<td>C2</td>
</tr>
<tr>
<td>N0(0,0) R2</td>
<td>299.76</td>
<td>HCO.B system</td>
<td>389.77</td>
<td>C2</td>
</tr>
<tr>
<td>HCO.B system</td>
<td>380.15</td>
<td>C2</td>
<td>370.53</td>
<td>C2</td>
</tr>
<tr>
<td>NH(0,0) R1</td>
<td>380.38</td>
<td>HCO.B system</td>
<td>373.35</td>
<td>C2</td>
</tr>
<tr>
<td>HCO system</td>
<td>381.48</td>
<td>C2</td>
<td>376.72</td>
<td>C2</td>
</tr>
<tr>
<td>NH(0,0) R2</td>
<td>398.36</td>
<td>HCO.B system</td>
<td>380.27</td>
<td>C2</td>
</tr>
<tr>
<td>NH(0,0) R1</td>
<td>398.73</td>
<td>HCO.B system</td>
<td>383.40</td>
<td>C2</td>
</tr>
<tr>
<td>NH(0,0) Q2</td>
<td>307.8</td>
<td>C2</td>
<td>396.55</td>
<td>C2</td>
</tr>
<tr>
<td>NH(0,0) Q1</td>
<td>306.04</td>
<td>C2</td>
<td>397.4</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R2</td>
<td>312.17</td>
<td>C2</td>
<td>398.34</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R1</td>
<td>512.34</td>
<td>C2</td>
<td>398.34</td>
<td>C2</td>
</tr>
<tr>
<td>NH(2,2) R2</td>
<td>318.6</td>
<td>C2</td>
<td>396.30</td>
<td>C2</td>
</tr>
<tr>
<td>NH(2,2) R1</td>
<td>326.45</td>
<td>C2</td>
<td>402.63</td>
<td>C2</td>
</tr>
<tr>
<td>NH(2,2) Q2</td>
<td>326.45</td>
<td>C2</td>
<td>405.52</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R2</td>
<td>325.11</td>
<td>C2</td>
<td>406.22</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R1</td>
<td>337.37</td>
<td>C2</td>
<td>415.24</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R2</td>
<td>337.47</td>
<td>C2</td>
<td>415.24</td>
<td>C2</td>
</tr>
<tr>
<td>NH(1,1) R1</td>
<td>341.15</td>
<td>C2</td>
<td>416.78</td>
<td>C2</td>
</tr>
</tbody>
</table>

Table 2: Chemical species and respective vibrational quantum number and associated wavelength
\[(\text{OH}/\text{CH}) = y_0 + Ae^{-\frac{t}{T}}\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.979</td>
<td>0.015</td>
<td>7.112</td>
<td>0.379</td>
</tr>
<tr>
<td>Method II</td>
<td>0.934</td>
<td>0.218</td>
<td>7.583</td>
<td>0.394</td>
</tr>
<tr>
<td>Method III</td>
<td>0.949</td>
<td>0.101</td>
<td>8.066</td>
<td>0.4058</td>
</tr>
<tr>
<td>Ikeda [21]</td>
<td>0.902</td>
<td>0</td>
<td>6.803</td>
<td>0.640</td>
</tr>
</tbody>
</table>

\[(C_2/OH) = y_0 + Ae^{-\frac{t}{T}}\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.999</td>
<td>0.115</td>
<td>0.0009</td>
<td>0.149</td>
</tr>
<tr>
<td>Method II</td>
<td>0.995</td>
<td>0.163</td>
<td>0.0004</td>
<td>0.152</td>
</tr>
<tr>
<td>Method III</td>
<td>0.998</td>
<td>0.088</td>
<td>0.0008</td>
<td>0.158</td>
</tr>
<tr>
<td>Ikeda [21]</td>
<td>0.934</td>
<td>0</td>
<td>0.0005</td>
<td>0.163</td>
</tr>
</tbody>
</table>

\[(C_2/CH) = y_0 + Ae^{-\frac{t}{T}}\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.999</td>
<td>-0.031</td>
<td>0.0043</td>
<td>0.225</td>
</tr>
<tr>
<td>Method II</td>
<td>0.984</td>
<td>0.196</td>
<td>0.0015</td>
<td>0.203</td>
</tr>
<tr>
<td>Method III</td>
<td>0.994</td>
<td>0.167</td>
<td>0.0015</td>
<td>0.195</td>
</tr>
<tr>
<td>Ikeda [21]</td>
<td>0.967</td>
<td>0</td>
<td>0.0032</td>
<td>0.222</td>
</tr>
</tbody>
</table>

Table 3: Equations and respective constants which relates the intensity ratio with equivalence ratio for the various Methods and for those obtained by [21], where \(R^2\) is the coefficient of determination, for the laminar premixed propane air flames.
\[(\text{OH}^*/\text{CH}^*) = y_0 + Ae^{-\phi t}\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.991</td>
<td>0.227</td>
<td>105.77</td>
<td>0.222</td>
</tr>
<tr>
<td>Method II</td>
<td>0.987</td>
<td>-0.311</td>
<td>27.518</td>
<td>0.392</td>
</tr>
<tr>
<td>Method III</td>
<td>0.987</td>
<td>-0.285</td>
<td>27.902</td>
<td>0.388</td>
</tr>
<tr>
<td>Hardalupas et al [7]</td>
<td>0.964</td>
<td>0.363</td>
<td>6.843</td>
<td>0.497</td>
</tr>
<tr>
<td>Kojima et al [13]</td>
<td>0.995</td>
<td>0.526</td>
<td>779.016</td>
<td>0.159</td>
</tr>
<tr>
<td>Cheng et al [18]</td>
<td>0.993</td>
<td>0</td>
<td>28.546</td>
<td>0.347</td>
</tr>
</tbody>
</table>

\[(C_2^*/\text{OH}^*) = y_0 + Ae^\phi t\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.997</td>
<td>0.009</td>
<td>0.00002</td>
<td>0.111</td>
</tr>
<tr>
<td>Method II</td>
<td>0.978</td>
<td>0.139</td>
<td>0.00002</td>
<td>0.112</td>
</tr>
<tr>
<td>Method III</td>
<td>0.964</td>
<td>0.101</td>
<td>0.00007</td>
<td>0.1266</td>
</tr>
<tr>
<td>Kojima et al [13]</td>
<td>0.784</td>
<td>-0.048</td>
<td>0.00012</td>
<td>0.136</td>
</tr>
<tr>
<td>Cheng et al [18]</td>
<td>0.988</td>
<td>0</td>
<td>0.0045</td>
<td>0.2416</td>
</tr>
</tbody>
</table>

\[(C_2^*/\text{CH}^*) = y_0 + Ae^{\phi t}\]

<table>
<thead>
<tr>
<th>Methods &amp; Authors/Constant</th>
<th>(R^2)</th>
<th>(y_0)</th>
<th>(A)</th>
<th>(t)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method I</td>
<td>0.999</td>
<td>0.078</td>
<td>0.00022</td>
<td>0.149</td>
</tr>
<tr>
<td>Method II</td>
<td>0.997</td>
<td>0.450</td>
<td>2.294 \times 10^{-6}</td>
<td>0.0996</td>
</tr>
<tr>
<td>Method III</td>
<td>0.982</td>
<td>0.440</td>
<td>6.401 \times 10^{-6}</td>
<td>0.108</td>
</tr>
<tr>
<td>Hardalupas et al [7]</td>
<td>0.958</td>
<td>4.189</td>
<td>2.368 \times 10^{-6}</td>
<td>0.105</td>
</tr>
<tr>
<td>Kojima et al [13]</td>
<td>0.783</td>
<td>0.165</td>
<td>0.00006</td>
<td>0.135</td>
</tr>
</tbody>
</table>

Table 4: Equations and respective constants which relate the intensity ratio with equivalence ratio for the various Methods and for those obtained by Hardalupas et al [7], Kojima et al [13] and Cheng et al [18], where \(R^2\) is the coefficient of determination, for the laminar premixed methane air flames.
Chapter 9: 
Emanuel Teodori, Ana Moita and Antonio Luis Moreira, “Evaluation of pool boiling heat transfer over micro-structured surfaces by combining high-speed visualization and PIV measurements”.
Evaluation of pool boiling heat transfer over micro-structured surfaces by combining high-speed visualization and PIV measurements

E. Teodori, A. S. Moita and A. L. N. Moreira

IN+ - Instituto Superior Técnico, TU Lisbon, Lisbon, Portugal

ABSTRACT
The present work introduces an alternative approach which combines heat transfer measurements, high speed visualization and PIV to infer on the effect of surface micro-structuring in the various heat transfer parcels. The PIV provides particularly interesting information on the bubbles vertical velocity, allowing to infer how the micro-structures affect the bulk induced flow. The micro-patterns are composed by arrays of cavities with fixed shape and depth, only varying the distance between cavities, \( S \).

The results confirm the importance of enhancing the pool boiling performance, by promoting the increase of the bubble frequency and active nucleation sites density with micro-patterned surfaces. The role of the patterns is different, depending on the properties of the fluids. Hence, for liquids with very high latent heat of evaporation and surface tension, the parcel of heat transfer related to the phase change is very important. Based on the force balance on the departing bubbles, it is evident that in this case the micro-patterns will mainly affect the bubble coalescence, which may lead to steep deterioration of the heat transfer coefficient. On the other hand, for liquids with lower thermal properties (and lower surface tension) the micro-patterns play an important role in increasing and stabilizing the vertical velocity of the bubbles, thus favoring a very efficient boiling performance, with high boiling frequencies and large active nucleation sites density. This role of the micro-patterned surfaces is well captured by the representation proposed here, which relates the heat transfer coefficient with the distance between micro-patterns \( S \), made dimensionless by the characteristic dimension \( (L_c) = (\sigma_l/(\rho_l-\rho_v))^{1/2} \). This representation actually shows the best performance of the HFE7000 in comparison to that of ethanol, when the micro-patterned surfaces are used, despite the lower thermal properties of HFE7000. This trend was confirmed experimentally in our measurements.

INTRODUCTION
Most of the studies addressing enhanced surfaces, often modified by changing surface topography, make use of a trial and error approach, to improve the heat transfer coefficient and increase the Critical Heat Flux (e.g. [1]). Although valuable outcomes were obtained from such studies, this may not be the most effective approach given the lack of universality of the empirical correlations devised by this approach, as recently shown by McHale and Garimella [2]. While surface wettability mainly plays a role on the onset of boiling and on keeping the nucleation stable (through contact angle hysteresis), so that it affects CHF [3] surface topography promotes the increase of active nucleation sites density, but also affects significantly the interaction mechanisms (e.g. [4]). In [4] one has explored in detail the effect of surface topography on bubble dynamics and suggested a relation linking the distance between micro-cavities and the heat transfer coefficient. Basically, one can determine an optimum distance which maximizes the heat transfer coefficient \( h \) by triggering bubble growth. Further decreasing that distance, bubble coalescence near the surface will be excessively strong causing a steep deterioration of \( h \). The work performed in [4] produced guiding results, which should be confirmed for a wider range of refrigerants. Also, given the considerably high latent heat of evaporation of the liquids used in [4], the relations devised are mostly related to the heat transfer parcel associated to the latent heat. However, there are other two parcels involved in the pool boiling heat transfer, namely the natural convection and the bulk convection (induced by bubble growth and motion) [5] whose relative importance is not clearly accessed.

In line with this, the present work proceeds with the analysis suggested in [4] over a wider range of liquids to infer on the relative importance of the various pool boiling heat transfer parcels. The experimental approach combines heat transfer measurements, high speed visualization and PIV to infer on the effect of surface micro-structuring in the various heat transfer parcels.

MATERIALS AND METHODOLOGY
Pool boiling is investigated for various liquids, namely the dielectric fluid HFE 7000, ethanol and water. These fluids were selected since they have a gradual decrease in the values of surface tension and of the relevant thermal properties such as the thermal conductivity, the heat capacity and the latent heat of evaporation, as shown in Table 1.
Table 1 Thermophysical properties of the liquids used in the present study, taken at saturation, at 1.013x10^5 Pa.

<table>
<thead>
<tr>
<th>Property</th>
<th>Ethanol</th>
<th>Water</th>
<th>HFE7000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{sat}$ [°C]</td>
<td>78.4</td>
<td>100</td>
<td>34</td>
</tr>
<tr>
<td>$\rho_l$ [kg/m$^3$]</td>
<td>736.4</td>
<td>957.8</td>
<td>1374.7</td>
</tr>
<tr>
<td>$\rho_v$ [kg/m$^3$]</td>
<td>1.647</td>
<td>0.5956</td>
<td>4.01</td>
</tr>
<tr>
<td>$\mu_l$ [mN m/s$^2$]</td>
<td>0.448</td>
<td>0.279</td>
<td>0.3437</td>
</tr>
<tr>
<td>$C_{pl}$ [J/kgK]</td>
<td>3185</td>
<td>4217</td>
<td>1352.5</td>
</tr>
<tr>
<td>$k_l$ [W/mK]</td>
<td>0.165</td>
<td>0.68</td>
<td>0.07</td>
</tr>
<tr>
<td>$h_{fg}$ [kJ/kg]</td>
<td>849.9</td>
<td>2257</td>
<td>142</td>
</tr>
<tr>
<td>$\sigma_{lv}$ [N/m]$\times$10$^3$</td>
<td>17</td>
<td>58</td>
<td>1.2</td>
</tr>
</tbody>
</table>

Heat flux and heat transfer coefficients are determined for the various liquid/surface pairs. Afterwards, they are related to the bubble dynamics. This characterization is made by combining high-speed visualization with PIV measurements.

The micro-cavities are squares with fixed size length $a=52$ µm and fixed depth $h_R=20$ µm. The distance between the centers of the cavities $S$ is mainly our optimization variable, ranging between 300 µm$<S<$1200 µm. The parameters characterizing the micro-patterns are schematically defined in Figure 1, together with a photo of a sample.

![Figure 1](image.png)

**Figure 1** a) Identification of the main parameters quantifying the micro-patterns. b) Sample of a pattern with square cavities.

Heat flux and heat transfer coefficients are determined for the various liquid/surface pairs. Afterwards, they are related to the bubble dynamics. This characterization is made by combining high-speed visualization with PIV measurements.

The patterns are custom made from silicon wafers combining wet etching with plasma etching and the roughness profiles measured using a profile meter with a precision of ±100 Angstroms. Table 2 depicts the main topographical characteristics of the surfaces used in this study. The Table includes the average values of the static contact angles, which were measured as described in [6]. The contact angles obtained with ethanol and HFE7000 in contact with all the surfaces are close to zero.

Table 2 Main range of the topographical characteristics of the micro-patterned surfaces. $\theta$ is the average static contact angle measured with water at room temperature. $\theta\approx0^\circ$ for all the surfaces in contact with ethanol and HFE7000.

<table>
<thead>
<tr>
<th>Material</th>
<th>Reference</th>
<th>$a$ [µm]</th>
<th>$h_R$ [µm]</th>
<th>$S$ [µm]</th>
<th>$\theta$ [$^\circ$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon Wafer</td>
<td>Smooth</td>
<td>$\varnothing$ 0</td>
<td>$\varnothing$ 0</td>
<td>$\varnothing$ 0</td>
<td>86.0</td>
</tr>
<tr>
<td>C1</td>
<td>52</td>
<td>20</td>
<td>304</td>
<td>90.0</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>52</td>
<td>20</td>
<td>400</td>
<td>91.5</td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>52</td>
<td>20</td>
<td>464</td>
<td>71.5</td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>52</td>
<td>20</td>
<td>626</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>52</td>
<td>20</td>
<td>700</td>
<td>95.0</td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>52</td>
<td>20</td>
<td>800</td>
<td>60.5</td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>52</td>
<td>20</td>
<td>1200</td>
<td>66.3</td>
<td></td>
</tr>
</tbody>
</table>
HEAT TRANSFER MEASUREMENTS

The boiling curves are presented for each liquid and each heating surface by varying the imposed heat flux in steps of 15W/cm². Each curve is obtained from the average of seven experiments. The liquid is degassed before each experiment by maintaining it in the pool at 20°C above the saturation temperature and the experimental procedure is started, for each heat flux, only after the system has reached thermal equilibrium, i.e. when the temperature oscillation is smaller than ±0.5 °C.

Experiments are conducted to obtain average boiling curves by both increasing and decreasing the heat flux, to infer on hysteresis effects, as also pointed by Mohamed and Bostanci [7]. The temperature measurements have an uncertainty of ±1°C. The relative error associated with the determination of the heat transfer is 5%.

IMAGE ANALYSIS OF BUBBLE DYNAMICS

Following an approach similar to that presented in many works reported in the literature, the parameters selected here to characterize bubble nucleation are the bubble departure diameter, the bubble departure frequency and the active nucleation sites density. This characterization is based on high-speed visualization and image post-processing. The images are recorded with a frame rate of 2200fps. For the optical configuration used here, the spatial resolution is 9.346µm/pixel.

The bubble departure diameter is measured for each test condition from 300 to 1060 frames. For each image a mean value is averaged from 5-16 measurements for every nucleation site that is identified in the frame.

At higher heat fluxes, the various interaction mechanisms, which will be discussed in the following section, may alter significantly the value of the departure diameter, especially when horizontal coalescence occurs. Therefore, in those cases, the measured diameters are a mean value taken from the averaged diameters, which are evaluated after the occurrence of such events close to the wall.

The error associated to the measurements of the bubble departure diameter is ±9.346 µm.

The bubble departure frequency is estimated by determining the time elapsed between apparent departure events, which are counted for a defined interval of time. The departure frequency is assessed, for each test condition, for at least five nucleation sites, which are evaluated based on extensive image post-processing of 300 to 1060 frames. The final value of the bubble departure frequency is the average between the frequencies of each nucleation site. The uncertainty associated to these measurements is ±1 fps.

Finally, the evaluation of the active nucleation sites density must be done by visual inspection of the frames, which introduces an uncertainty associated to the subjective criterion of the observer. To lessen this uncertainty, at least ten frames are chosen, at different times during the single experiment. The final values of the active nucleation site density are an average of the ten evaluated values.

PIV MEASUREMENTS

Several studies in the literature confirm the potential of using PIV to measure bubble velocity inside a flowing fluid, as for example reported by [8] and [9]. However, the results obtained from this technique are very sensitive to the characteristics of the flow and to the parameters used during the visualization and the post-processing of the images (e.g. [9]). In the present work, the main focus is to measure the velocity of the bubbles, so seeding was not used, but instead the bubbles are followed, as suggested by [9]. Bubbles diameter is in the range of 500-800 µm, measured by image post-processing. These dimensions and the low characteristic velocities of the bubbles (1-10 cm/s) require a careful analysis of all the parameters which have to be selected in the PIV configuration. The PIV system uses a CCD camera Kodak Megaplus, Model 1.0, with an image resolution of 1018x1008 pixel². The bubbles are illuminated via a dual Nd:YAG Litron laser. The PIV set-up is shown in Figure 2, together with and the coordinate system considered in the measurements.

The time delay between laser pulses is varied (1<∆t<8ms) depending on the imposed heat flux: the time between pulses is smaller for higher imposed heat fluxes. Furthermore, the interrogation area and the overlap are also varied for the various imposed heat flux conditions, in an optimization process, to assure that the chosen values are adequate to obtain accurate measurements. Hence, the selected interrogation area was varied between 16 and 64 pixels (1pixel/58µm) to assure that at least five bubbles are inside. An overlap of 50% is chosen by analyzing two consecutive frames and evaluating the average displacement of the bubbles. The most appropriate approach for this kind of flow is using a recursive cross correlation or the average correlation algorithms (e.g. [9]). In the present work, after analyzing extensively both approaches, the cross correlation was considered to be the most appropriate. The measurements performed using PIV are compared with extensive image post-processing, within quite good agreement. The PIV data were processed with the software Flow manager 4.2.
RESULTS AND DISCUSSION

A detailed analysis of bubble dynamics concerning the quantitative evaluation of the bubble departure diameter and frequency as well as nucleation sites density is performed for water, ethanol and HFE7000 over the same surfaces studied here, in a previous work [10]. Based on a force balance to the bubble at detachment it is clear that the largest bubbles are formed during the boiling of the fluid with the largest surface tension (water). The departure frequency is similar for most of the surfaces, for the three liquids, but there is a significant reduction of the frequency of the water bubbles, for the surface with the smallest surface tension, which is associated to an excessively promoted coalescence.

The relation between coalescence and the distance between the cavities is easily confirmed in Figure 3: the coalescence factor $D_b/D>> 1$ (being $D_b$ the averaged bubble diameter and $D$ the diameter as the bubble exits the cavity, i.e. with no coalescence) for water and varies significantly with $S$, thus evidencing the strong coalescence effects observed at water boiling (Figure 3a). This behaviour contrasts with that of well wetting fluids, with low surface tension, such as ethanol and HFE7000 (Figures 3b and 3c, respectively).

This analysis is consistent with the boiling curves and heat transfer coefficients shown in Figure 4: although ethanol and HFE 7000 presented a more homogenous and vigorous boiling with limited interaction mechanisms (and particularly coalescence), this was not enough for these fluids to reach the overall higher heat transfer coefficients of water. Nevertheless, the strong horizontal coalescence phenomena characterizing water boiling on surface C1 ($S=304 \mu m$), leads to a steep deterioration of the heat transfer coefficient, so that the heat transfer of water and ethanol are quite similar for the aforementioned surface. Hence, the coalescence effects can be empirically related to the distance $S$ since this distance between cavities is directly acting on the coalescence mechanisms which occur close to the surface and so should be well related to the force balance describing the bubble detachment, as proposed by Fritz [11] and followed by many other researchers to scale bubble departure diameters: $Lc=(\sigma_v/(\rho_v-\rho))/2$. This empirical relation, earlier suggested in [4] is extended here for ethanol and HFE7100 pool boiling in Figure 5.
Figure 4 Boiling curves and heat transfer coefficients over micro-patterned surfaces of a) water, b) ethanol, c) HFE7000.
The representation in Figure 5, already allows to account for the relative improvement of the performances using the micro-patterned surfaces since, on one hand compares the enhancement of the heat transfer coefficient (given by the three heat transfer parcels: natural convection, evaporation and induced bulk convection) using each micro-textured surface in relation to that of the smooth surface. On the other hand, the distance \( S \) is related to the characteristic bubble departure dimensions. This relation is important as it defines the critical distance \( S \) up to which the bubbles with a certain diameter, depending on the balance between surface tension and buoyancy forces \( (L_c = (\sigma/lv/(\rho_l - \rho_v))^{1/2}) \) will coalesce. Such relation must still be refined, but one may already identify a maximum for the water boiling, which is related to the maximum \( S \) above which the horizontal coalescence will conduce to the declination of the heat transfer coefficient (as identified in the boiling curves in Figure 5).

The absolute values of \( h \), being the result of the sum of the 3 parcels of the heat transfer [5] are dominated by the largest values of the thermal properties of the water, which presents the highest values of \( h \), despite its worse performance as \( S \) decreases. However, the plot in Figure 5 clearly indicates that although the properties of ethanol, namely the latent heat of evaporation are significantly larger than those of HFE7000, the relative enhancement in the heat transfer obtained with the micro-patterned surfaces is higher than that obtained for ethanol, so the curve for HFE7000 is actually above that of ethanol. This is probably due to significant improvement of the micro-structure on bubble dynamics, which is affecting one of the parcels of the total heat transfer. Given the low thermal properties of HFE7100, the most probable effect of \( S \) should be on the induced bulk convection. To infer on this, a detailed study of bubble dynamics was performed, based on PIV measurements.

**FLOW DESCRIPTION: ANALYSIS OF THE AVERAGE VERTICAL BUBBLE VELOCITY**

The effect of the surface patterning in the parcel of the induced convection was investigated evaluating the average vertical bubble velocity (average of the velocity profile for a fixed value of H/D), along the vertical dimensionless distance H/D, where H is the vertical distance from the top face of the surface in (mm) and D is the bubble departure diameter (also in mm), for different heating conditions and different micro-patterns. Naturally that this effect is relevant close to the surface, but given the well known restrictions of PIV measurements performed very close to the surface (e.g. Raffel et al. [12]) the assessment of bubbles’ velocity must be performed at various distances H/D, in order to understand bubbles’ motion. The results are presented in Figure 6, for ethanol and HFE7000, boiling over micro-patterned surfaces with decreasing \( S \).

Overall, the bubbles are ejected with larger average velocity, as the imposed heat flux is higher, although this trend is more evident for ethanol. The bubbles then decelerate due to the braking effect forced by the zero velocity at the top of the pool. This is quite more evident in ethanol pool boiling with is related to its physico-chemical properties and to a less efficient boiling process, as aforementioned. Indeed a vigorous boiling activity already pointed out for HFE7000 can be identified in the analysis of the bubbles vertical velocity profiles. There is a strong oscillation along the vertical velocity around the mean value for HFE7000, which is speculated to be due to the very vigorous boiling. Similarly to what is observed for ethanol, surfaces with closer cavities (C2 S=400\( \mu \)m, and C5 S=700\( \mu \)m) present more uniform and stable profile when compared to those with sparser cavities (C6 S=800\( \mu \)m, and C7 S=1200\( \mu \)m). In fact it is now clearer that the stronger interaction phenomena still occur at the surface with closer cavities, but they are observed at a higher distance from the wall for HFE7000, when compared to ethanol and to water. Hence, the coalescence will still affect directly the heat transfer of ethanol by producing vapor bubbles which block the fluid circulation near the surface, while for HFE7000 it contributes for the formation of a denser bubble plume, acting as stabilization factor when one speaks in terms of vertical velocities.
For surfaces with sparser cavities, the oscillations of the vertical velocity are more evident than those observed for ethanol, due to the higher number of active nucleation sites and the higher value of bubble departure frequency. However, the vertical velocity is overall much higher for HFE7000, so at the end, the induced bulk convection is more efficient for this liquid, when compared with ethanol.

**Figure 6** Vertical velocity profile for a) Ethanol and b) HFE 7000 over micro-patterned surfaces: C2 S = 400 µm, C6 S = 800 µm and C7 S = 1200 µm.

The much higher bubble vertical velocities measured for HFE7000 when compared to that of ethanol are quite evident in Figure 7. The overall value oh h is yet larger for ethanol, as it results from the sum of the 3 heat transfer parcels.
Figure 7 Characteristic velocity vs wall superheat for ethanol and HFE 7000. a) Surface C2 (S=400 µm) b) Surface (C5 S=700 µm) c) Surface (C6 S=800 µm) d) Surface (C7 S=1200 µm).
This analysis is consistent if the parcel of the induced bulk convection is more important, relatively to the others, which is true, if one considers the approach of other researchers such as Kandlikar [3] or Gerardi et al. [13,14]. Indeed, considering that the induced bulk convection is strongly related to the so-called quenching heat flux, which is proportional to $f_bN_{SD}$ (being $f_b$ the bubble frequency $N_{SD}$ the active nucleation sites density), it is clear the importance of augmenting these two parameters, as achieved in the HFE7000 with the micro-patterned surfaces. This parcel of bulk induced convection is actually estimated in our study to be dominating, which is in agreement with the observations of Gerardi et al. [13,14]. The approach followed here allows to obtain experimentally all the data that is required to check on the universality of the formulations devised by many of these authors (e.g. [3, 13, 14]). However, these formulations are rather empirical so that a more theoretical approach is preferred, possibly based on the classical relation $Nu=Re^{a}Pr^{b}$, which should be compared to some empirical approaches in the near future.

**SUMMARY**

The present work introduces an alternative approach which combines heat transfer measurements, high speed visualization and PIV to infer on the effect of surface micro-structuring in the various heat transfer parcels. The results confirm the importance of enhancing the pool boiling performance, by promoting the increase of the bubble frequency and active nucleation sites density with micro-patterned surfaces. The role of the patterns is different, depending on the properties of the fluids. Hence, for liquids with very high latent heat of evaporation and surface tension, the parcel of heat transfer related to the phase change is very important. Based on the force balance on the departing bubbles, it is evident that in this case the micro-patterns will mainly affect the bubble coalescence, which may lead to steep deterioration of the heat transfer coefficient. On the other hand, for liquids with lower thermal properties (and lower surface tension) the micro-patterns play an important role in increasing and stabilizing the vertical velocity of the bubbles, thus favoring a very efficient boiling performance, with high boiling frequencies and high active nucleation sites density. This role of the micro-patterned surfaces is well captured by the representation proposed here, which relates the heat transfer coefficient with the distance between micro-patterns $S$, made dimensionless by the characteristic dimension ($L_c=(\sigma v_l/(\rho_l-\rho_v))^{1/2}$). This representation actually shows the best performance of the HFE7000 in comparison to that of ethanol, when the micro-patterned surfaces are used, despite the lower thermal properties of HFE7000. This trend was confirmed experimentally in our measurements.

**ACKNOWLEDGMENTS**

The authors are grateful to Fundação para a Ciência e a Tecnologia (FCT) for partially financing the research under the framework of project PTDC/EME-MFE/109933/2009 and for supporting A.S. Moita with a Fellowship (Ref.:SFRH/BPD/63788/2009).

**REFERENCES**


Chapter 10:

Hugo Horta, “Global and prominent higher education: internationalization, competitiveness and the role of the state”.
This article provides a characterization of the internationalization of “global” European universities and discusses the role of the State in promoting greater internationalization and competitiveness levels of prominent national universities. The analysis supports previous arguments stating that global ranking of universities is strongly based on research, but reveals that the internationalization of research universities’ student population is also arranged to enhance research capacity. This finding is further reinforced by a positive association between the internationalization of the academic staff and the internationalization of the student population in one of those universities, being this association particularly strong with the doctoral student population. Finally, based on the analysis of two prominent national universities with different global competitiveness levels, we discuss the role of the State as a central supporter of these universities internationalization and global competitiveness arguing that public funding and support is critical if countries want to have national prominent universities competing at global level.

KEYWORDS:

World class universities; Prominent national universities; Internationalization academic staff; Internationalization student population; Role of the State; Resources

1. Introduction

In the last decades of the 20th century, the theme of internationalization in higher education has gained increasing relevance and attention by researchers and policy makers alike. For higher education institutions, and more concretely, prominent national universities oriented to research activities, the changing environment has meant a new challenge: having activities increasingly performed cooperatively and competitively at global level (either in dynamics of internationalization or globalization; see Teichler 2004). For most of these institutions, this has entailed managing conflicting interests derived from increasingly performing activities at local, regional, national and global levels (Marginson and Rhoades 2002). It has also entailed that, in the global competitive environment, their often undisputed dominant national positioning meant less, sometimes much less than before (Marginson 2006). This is bluntly shown by the somewhat polemic, but generally accepted, university global ranking systems (Jobbins 2005).

In the current knowledge economy and globalization mix, these rankings have added pressure on university leaders to compete and cooperate internationally. Internationalization of higher education and higher education institutions, thus, is high on the strategic agendas.

Notwithstanding the political and policy discourses on the need to foster international connections and competition, the internationalization strategy in research oriented universities in several countries is still marred by blurriness and lack of resources (Marginson and Sawir 2006). In this article, we discuss critical aspects related to the internationalization of research oriented national universities. Firstly, we acknowledge the argument forwarded by Marginson that “global competition is vectored by research capacity” (2006: 1) and that research capability represents the core of the positional goods that rank universities globally.

Thereafter, based on the analysis of several major European research universities, we provide evidence that the universities placed in the top international university ranking positions in Europe share a common student structure. This structure relies heavily on graduate student population and on its strong internationalization. We further find that the internationalization of the student population at graduate level is associated to the internationalization of the academic staff. Finally, two prominent European national technical universities are analyzed—one ‘ranked’ as a global research university and other as a national research university struggling to internationalize its activities and campus—to highlight the role of the State as an enabler of these universities’ internationalization dynamics.

2. Globalization, internationalization, rankings and world class universities: research capability is the key

The evolution of globalization and of the knowledge society has led to systemic and institutional changes in higher education systems, and has required universities to adapt their character and functions to meet complex societal demands and expectations (Mok and Welch 2003). Supra-national institutions, such as the OECD, have guided national policies towards higher education leading to major systemic and structural changes in higher education systems and institutions (Vaira 2004). These include changes in managerial attitudes and cultures
(Deem and Brehony 2005), the changing role of the State, from a position of almost full control to one of steering at a distance (Amaral and Magalhães 2001) which has resulted in increased institutional autonomy, but also in the increment of accountability exercises (Horta and Vaccaro 2008), and the promotion of schemes of performance based-funding and institutional competition (Jongbloed 2006).

Most universities worldwide have needed to change and become more entrepreneurial (e.g.: Slaugther and Leslie 1997), and this entrepreneurial attitude has led universities to extend the scope of their activities outside the national borders. As the activities of higher education institutions become more developed in international (in terms of cooperation) and global (in terms of competition) frameworks (see Teichler 2004), the more these institutions face new challenges. Universities that dominate in their national education systems now see themselves struggle to improve the quality of their academic activities when compared with their international peers.

If an analogy can be made to the world of football, these universities are no longer playing exclusively with peer institutions in national higher education leagues, where they enjoy a certain degree of supremacy based on their reputation, history, teaching quality or national research relevance. Instead, although they continue to play there, they have also started playing in the “Champion’s League”, in the “World higher education Champions League”. There, they may also develop activities in cooperation, but are in effective competition with other universities that have far greater resources, positional goods, and are equally or better integrated in international research and teaching networks of excellence. The universities that have established themselves earlier in global higher education have the competitive advantage. This is so because they belong to countries with dominant scientific systems, have more resources or acquired international reputation and experience by developing activities at the international level earlier. They have created a brand, and these are the universities that tend to be regarded as the per excellence “world class universities” (Shattock 2003).

Harvard, MIT, Yale or Cambridge tend to represent the image of the “world class universities” that most universities in the world aspire to be. (These are for higher education, the Real Madrid’s, Liverpool’s, Inter Milan’s and other major teams of the football champion’s league.) The position of these “world class universities” in the international arena is legitimized by worldwide university league tables that assess mainly performance characteristics associated to research activities, but which nonetheless, fuel the competitive enthusiasm among universities at global level (e.g.: Dill and Soo 2005). Also, as in football, all countries want to have a top team competing internationally, or in the case of higher education, a “world class university”. In this regard, Altbach (2003) states “Everyone wants a world class university. No country feels it can do without one”. The same author adds “The problem is that no one knows what a world-class university is...”. It may be true that there is no agreement on a definitive concept of “world Class University”, but the research university model is in everyone’s mind

---

when the idea of “world class universities” is mentioned. The struggle that most European and Asian universities undergo to improve their positions in the research focused international rankings (Deem et al. 2008), and the willingness of most of these universities to have their name or academic activities associated with a better ranked university (Marginson and Sawir 2006), also supports this assumption.

When looking at the top 100 or 500 worldwide university ranking table’s one feature is obvious: all universities are recognized worldwide as “research universities”. Research capacity, output and quality define global competition in higher education (Marginson 2006). Their research capability itself is fueled by student selectivity processes that ensure that these universities recruit the most promising students. These universities, unlike others, do not have difficulties in attracting students, as these are driven by the positional goods that the courses taken at those universities are able to offer them in the labor and academic markets. The fact that these universities are based in countries with strong economies further enhances these countries’ role as international student attraction magnets, thus, contributing to one-way flows of international students between developing and developed countries (Marginson 2006).

But what type of international students are these top research universities attracting? Table 1 shows relevant characteristics of the top 10 European universities in the academic world rankings. Although these rankings have severe methodological flaws recognized by others (e.g.: van Raan 2005) and the authors themselves (Liu and Cheng, 2005), they present a good outline of what the “university world hierarchy” is (Jobbins 2005). The table focuses on European universities instead of the world universities in order to decrease the overwhelming domination of the native English language speaking universities in the table, especially the doctoral universities in the US that dominate the international competition for students (Marginson 2006).

<table>
<thead>
<tr>
<th>European rank Position</th>
<th>World rank position</th>
<th>University</th>
<th>Country</th>
<th>% International faculty</th>
<th>% International undergraduate students (vs total undergraduate students)</th>
<th>% International graduate students (vs total graduate students)</th>
<th>% International Students (vs total students)</th>
<th>Ratio international graduate students/international undergraduate students</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>Univ Cambridge</td>
<td>UK</td>
<td>39</td>
<td>14</td>
<td>51</td>
<td>27</td>
<td>1.82</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>Univ Oxford</td>
<td>UK</td>
<td>36</td>
<td>11</td>
<td>57</td>
<td>27</td>
<td>2.66</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>Imperial College</td>
<td>UK</td>
<td>39</td>
<td>34</td>
<td>49</td>
<td>40</td>
<td>0.77</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>Univ College London</td>
<td>UK</td>
<td>29</td>
<td>23</td>
<td>42</td>
<td>29</td>
<td>0.92</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>ETH Zurich</td>
<td>Switzerland</td>
<td>53</td>
<td>13</td>
<td>35</td>
<td>23</td>
<td>2.38</td>
</tr>
<tr>
<td>6</td>
<td>39</td>
<td>Univ Paris 06</td>
<td>France</td>
<td>3</td>
<td>17</td>
<td>66</td>
<td>27</td>
<td>1.00</td>
</tr>
<tr>
<td>7</td>
<td>42</td>
<td>Univ Utrecht</td>
<td>Netherlands</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>46</td>
<td>Univ Copenhagen</td>
<td>Denmark</td>
<td>17</td>
<td>11</td>
<td>16</td>
<td>13</td>
<td>1.2</td>
</tr>
<tr>
<td>9</td>
<td>48</td>
<td>Univ Manchester</td>
<td>UK</td>
<td>30</td>
<td>15</td>
<td>47</td>
<td>23</td>
<td>1.01</td>
</tr>
<tr>
<td>10</td>
<td>52</td>
<td>Univ Paris 11</td>
<td>France</td>
<td>10</td>
<td>10</td>
<td>27</td>
<td>17</td>
<td>2.12</td>
</tr>
</tbody>
</table>
Previous analyses have shown that the student structure of these universities is organized according to their primary mission: research (e.g.: Horta 2008). Therefore, the concern of these universities is to guarantee their research capability by focusing their teaching activities at graduate level, and particularly emphasizing doctoral programs and research. As shown on Table 1, these universities tend to do the same regarding the enrollment of international students and recruitment of academic staff. On average, the universities in the table have 24% of their student population composed of international students, ranging from a maximum of 40% in the Imperial College in the United Kingdom to a minimum of 13% in the University of Copenhagen in Denmark.

However, when the student population is disaggregated by level of education, a more revealing picture is shown. At these universities, only 16% of the undergraduate student population consists of international students while at graduate level this percentage increases to 41%. This indicates that the internationalization of the student population at the top European research universities is focused on graduate education. The average ratio between international graduate and undergraduate students also indicates that these universities have 36% more international students at graduate level than international students at undergraduate level. The number of international graduate students is more than twofold when compared with the number of undergraduate international students at least in 3 universities (Oxford, ETH Zurich and Paris 11), more than 50% in the University of Cambridge, and only in two universities there were more international undergraduate students than international graduate students (Imperial College and University College London).

This in turn implies that the internationalization strategies of these universities follows the same rationale as that of the student structure. That is, the international population is mostly concentrated at educational levels that support the research activities of these universities, thus fostering their scientific performance and institutional reputation both nationally and internationally (as suggested by Marginson 2006). Also, because the current analysis is set in the European context, it is likely that the rationales for the internationalization of the student population at these universities are deemed to be different according to the educational levels. For example, student exchange programs, such as the SOCRATES program or similar ones, tend to be focused on undergraduate education under the scope of an Europeanisation process, which is characterized by the rationale that cooperation and mobility aim to reinforce cultural and social ties in the European space (Teichler 2004). The rationale of broadening cultural and social horizons of the students at a campus is – like the SOCRATES/ERASMUS programs - more related to the internationalization of students at undergraduate level than
supporting research activities which is more associated with the internationalization of the graduate student population.

Another characteristic of the universities shown in Table 1 is the high rate of internationalization of the academic staff. On average 27% of the academic staff at the top European research universities consists of international faculty members. This not only suggests that these universities are deeply integrated in international teaching and research networks, but also that English is widely used as the lingua franca. It is in English that these universities are mostly offering graduate level courses, articles are presented and discussed at conferences, knowledge is read in the international major journals; it is also in English that colleagues communicate either by means of ICT technologies or in person (Dedoussis 2007). This is quite an important aspect of a research university. The importance of English is particularly stressed in the scientific world where the prestige (and impact) of publishing a scholarly article in the English language in an international or national journal or book surpasses the scholarly publication in non-English journals or books (Marginson and Rhoades 2002).

University Paris 6 and University Paris 11 represent special cases mirroring particularities that can be more associated to the French higher education system than to the universities themselves. Unlike their international peers, both universities have a reduced percentage of international faculty; in contrast, they have a substantial percentage of international students which is similar to their international peers. The reduced percentage of international faculty in these two universities, when compared with their counterparts in Table 1, is explained by two factors that led to urges for reform of academic careers at French universities (see Musselin, 2005). Firstly, the recruitment of faculty and upward career mobility in French universities involve extremely bureaucratic and complex processes, making them unattractive for international faculty, even if tenured positions are offered from the start (Musselin, 2006). Secondly, the French academic labor market still favors “local” careers, while at the same time it is observed a decreasing number of vacancies offered by French universities (Musselin, 2006), a declining number of international students being granted doctorates in France (Moguérou, 2002) and a diminishing number of international candidates applying to vacancies at French universities (Cytermann et al, 2003). These problems have been acknowledged as it has been increasingly perceived in French academia that there is a need for increased mobility and inclusion of international staff at French universities because “the presence of foreigners among university and research staff is becoming an indicator of dynamism and success” (Musselin, 2006: 125).

In this context, the strong internationalization of the student population of University Paris 6 and University Paris 11 can be explained by one critical factor besides the recognized academic and scholarly prestige of these universities. France has long had the power to attract international students through its globally-recognized scientific prominence (such as Germany; OBHE, 2007). This makes France, and its prominent universities, the main destinations for

---

2 As a matter of fact, it has been suggested that some elements of the French elites have been leaving the French system looking for more attractive academic positions abroad (see Postel-Vinay, 2002).
potential students of French speaking countries worldwide that want to enroll in tertiary education. Indeed, the great majority of the international students coming to France are from French speaking countries with which strong cultural and historical ties exist, such as Algeria or Senegal. The challenge is to widen the country of origin of the international students and explore new global markets, and there, French universities have faced difficulties (see OBHE, 2007).

The discussion so far underlines that the internationalization of the student population in the top European research universities is mostly focused at graduate level, and that the internationalization of the academic staff is significant. In this context, it is possible that a relationship between the internationalization of the academic staff and the student population exists.

3. Internationalization at a top research university in Europe: the ETH Zurich case

In order to explore the relationship between the internationalization of the academic staff and the internationalization of the student population, we analyze the Eidgenössische Technische Hochschule Zürich (ETH Zurich), or as it is internationally known, the Swiss Federal Institute of Technology. The choice for the ETH Zurich is based on various factors. It is a top European Research University based in a non English native speaking country; although Switzerland has a very advanced scientific system, it is not a ‘heavy weight’ like France or the UK (see Horta and Veloso 2007); it is a country with a medium-small sized population, therefore having a rather limited national student population. Its universities face a declining demand of national students that most European countries are now increasingly facing after the late 20th century student boom in higher education (Amaral and Magalhães 2008). But there is another reason for choosing ETH Zurich to explore the relationship between the internationalization of the academic staff and the internationalization of its student population: it is a relatively young higher education institution compared to most other European research universities. It provides, therefore, an interesting case study from which other ‘growing’ and ‘developing’ universities may learn.

ETHZ was founded in 1855 as a federal polytechnic institute, becoming a university in 1909. This transformation led to the growing importance of the research function, which is underlined by the fact that doctoral degrees were granted at the very early stages (Krull 2008). The internationalization of ETH Zurich at this time was based on the curricula, which combined national oriented specifications and interests with international curricular models and the recruitment of several international faculty members.

---


4 Information taken from: www.ethistory.ethz.ch [accessed several times between the 10th of July and the 1st of August]
Since its origins, the ETH Zurich has represented the pinnacle of academic research in Switzerland. In the 1960s and in the 1970s, the awareness that the research performed at the institute was lagging behind internationally lead to reforms, and the need for internationalizing the university was perceived as a priority. In the late decades of the 20th century, academic freedom became further related to entrepreneurial activities as funding started to be gradually allocated on the basis of achievement which enhanced the academic staff resourcefulness, proactivity and work quality standards. Furthermore, an awareness that international linkages and human resource attraction are critical in supporting the quality of the scholarly activities of the institute was reinforced. This is clearly perceived by the ETH Zurich mission statement: “the ETH Zurich can only compete with the world’s best by establishing international links, by recruiting its academic and research staff worldwide, and by remaining attractive to students from abroad” (as quoted by Dudler and Korosec 2005). This statement is clearly an indication of an international policy that combines the internationalization of the academic staff and the student population. Table 2 shows the numbers and percentage of professors that were recruited by all departments of ETH Zurich between 1990 and 2002 on what concerns their nationality and country of recruitment.

<table>
<thead>
<tr>
<th>Nationality</th>
<th>%</th>
<th>Country of recruitment</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Switzerland</td>
<td>155</td>
<td>Switzerland</td>
<td>170</td>
</tr>
<tr>
<td>Germany</td>
<td>90</td>
<td>USA</td>
<td>65</td>
</tr>
<tr>
<td>USA</td>
<td>22</td>
<td>Germany</td>
<td>57</td>
</tr>
<tr>
<td>Austria</td>
<td>16</td>
<td>Great Britain</td>
<td>15</td>
</tr>
<tr>
<td>Italy</td>
<td>12</td>
<td>France</td>
<td>10</td>
</tr>
<tr>
<td>France</td>
<td>11</td>
<td>Belgium</td>
<td>10</td>
</tr>
<tr>
<td>Great Britain</td>
<td>10</td>
<td>Austria</td>
<td>6</td>
</tr>
<tr>
<td>Belgium</td>
<td>10</td>
<td>Italy</td>
<td>5</td>
</tr>
<tr>
<td>Canada</td>
<td>5</td>
<td>Canada</td>
<td>5</td>
</tr>
<tr>
<td>Other European countries</td>
<td>13</td>
<td>Other European countries</td>
<td>5</td>
</tr>
<tr>
<td>Asian countries</td>
<td>7</td>
<td>Asian countries</td>
<td>3</td>
</tr>
<tr>
<td>African</td>
<td>2</td>
<td>Oceania</td>
<td>2</td>
</tr>
<tr>
<td>Central and South American</td>
<td>1</td>
<td>Central and South American</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>354</td>
<td>Total</td>
<td>354</td>
</tr>
</tbody>
</table>

Source: ETH Zurich, Annual Report

Table 2 - Professors taking office at ETH Zurich from 1990 to 2002, by nationality and country of recruitment

Although it is not surprising that the larger proportion of the recruited professors is Swiss nationals compared to other nationalities, it is rather unusual that the proportion of recruited Swiss professors (nationals) is lower than the recruitment proportion of international professors. This certainly accounts for the fact that ETH Zurich has 53% of its academic staff composed by international professors in 2007 (see Table 1). The range of the nationality of the recruited international faculty is rather broad. The substantial recruitment of German
nationals is explained not only by the sharing of the same language of the canton where ETH Zurich is located, but also by existing linkages between ETH Zurich and German universities that have a solid tradition in the hard and applied sciences of similar interest to the ETH Zurich (Krull, 2008). The latter reason seems to be far more important, since the percentage of Austrian professors hired (also German speaking natives) is much smaller than that of the German professors hired (4.5% Austrians to 25% Germans) with only about 2% of all faculty staff recruited in Austria and 16% of all faculty staff recruited in Germany.

The recruitment of scholars from many nationalities suggests that academic and scholarly quality and potential are critical in the recruitment process, whereas cultural and linguistic similarities are less so. Perhaps, the best indication of this is that ETH Zurich is recruiting scholars from scientific leading countries. ETH Zurich recruits about 42% of their staff in major scientific “superpowers” such as the United States, Germany, Great-Britain and France (Horta and Veloso 2007). This shows that ETH Zurich, as a European Research University, is able to be competitive in the world academic market in attracting the faculty members with promising and established academic and scientifc potential and capability, something that is critical to maintaining research proficiency and positional goods at global level (Marginson 2006).

However, the internationalization of the ETH Zurich is also reflected in the student population. ETH Zurich is a highly internationalized university in terms of its student population, which is focused at graduate level (35% of its graduate students are international students while only 13% of its undergraduate students are international students) in a similar fashion to its institutional peers in Table 1. One question, then, relates to whether internationalization of the academic staff is interrelated with the internationalization of the student population? In fact, the analysis shows that a higher internationalization of the faculty staff is indeed correlated to a growing internationalization of the student population (Figure 1a). However, it is when the relation between the internationalization of the faculty staff is done in relation to the student population disaggregated by educational levels that one can perceive where this relationship is strongest.

From the analysis of Figures 1b (undergraduate student population), 1c (master student population) and 1d (Doctoral student population), one can perceive that the internationalization of the academic staff at a research university is correlated with the internationalization of the student strata that produce scientific outputs and maintain the research capability of the university: the doctoral stratum. This is perceived by the very strong linear relationship between the internationalization of the faculty staff and the doctoral student population ($R^2=0.86$), as well as by the strong positive correlation among them (Pearson’s correlation coefficient: 0.926; $p < 0.01$). But an equally important result that sustains this argument is that the correlation between the internationalization of the academic staff and the undergraduate population is significant but negative (Pearson’s correlation coefficient: -0.617; $p < 0.05$).

This is not a surprising result for several reasons. At undergraduate level the overwhelming majority of the students are still by far national (see Teichler 2004). In the case of ETH Zurich, the German language still dominates as the teaching language at undergraduate level, while
English is considered the lingua franca at graduate level\textsuperscript{5}. Also, despite the greater internationalization of higher education and student flows, the great majority of undergraduates that conclude their studies still find work in their native national economy (Teichler and Jahr 2001). In this context, hiring international faculty to teach extensively at undergraduate level seems unnecessary particularly when universities are also under pressure to respond to local and national demands that are better understood by nationals (Ahola 2005).

\footnote{ETH Zurich information at Wikipedia \url{http://en.wikipedia.org/wiki/ETH_Zurich} [accessed several times between the 15\textsuperscript{th} of July and the 30\textsuperscript{th} of July]}
Note: Pearson’s correlation (coef: 0.667; sig 0.05); Source: ETHZ Admin. Services

Figure 1a – relationship between international to national faculty ratio and international to national student ratio, 1991-2003

Note: Pearson’s correlation (coef: 0.410; not sig); Source: ETHZ Admin. Services

Figure 1b – relationship between international to national faculty ratio and international to national undergraduate student ratio, 1991-2003

Note: Pearson’s correlation (coef: 0.926; sig 0.01); Source: ETHZ Admin. Services

Figure 1c – relationship between international to national faculty ratio and international to national master student ratio, 1991-2003

Note: Pearson’s correlation (coef: 0.926; sig 0.01); Source: ETHZ Admin. Services

Figure 1d – relationship between international to national faculty ratio and international to national graduate doctoral student ratio, 1991-2003
Finally, as at any research university, the focus of the strategic investment on human resources is mainly laid on the graduate school and on the research effort of the university as a way to sustain and improve the current research capability and the student selectivity that supports the reputational base that these universities hold worldwide (Marginson 2006). In this sense, the selectivity of academic staff to be hired is also critical. Hiring promising or established academic staff from universities from developed scientific countries brings alongside with them reputation, know-how, new perspectives, connections, and possible avenues to bring in international research funding (Lacetera et al. 2004). But the competition for bright minds is fierce (see Rosovsky 1990).

The competition for academic staff requires positional goods and financial resources, which are not available to the overwhelming majority of national research universities that are struggling to internationalize their campuses and activities, reform their structures, and compete globally with peer universities. The question then is how these universities can further internationalize their activities, and increasingly collaborate and compete with their international peers. Perhaps more to the point, it is important to consider whether these universities can do it ‘on their own’ or if the role of the State still matters.

4. The State as a “university internationalization engine”: two different perspectives

In a global world of higher education, most national governments want to have at least one university considered as an international research university competing and cooperating with other peer research universities globally. Governments promote competitive frameworks in higher education based mostly on research funding competition (e.g.: Jongbloed 2006) in order to prepare universities to compete both in the national and global level. The tendency to organize higher education systems according to mechanisms of government-induced managed competition known as quasi-markets have been occurring more frequently and across many countries (e.g.: Agasisti and Catalano, 2006). In some cases these quasi-markets have not had an effective implementation (see Teixeira et al. 2004). Yet even if the implementation of these quasi-markets has been effectively implemented, the problem seems to be that preparing national universities to adopt more flexible organizational structures, entrepreneurial and managerial attitudes and be more proactive through the creation of these State driven quasi-markets does not seem to be enough. Even in the latter cases, national prominent universities continue to lack resources to compete globally, particularly with the doctoral university sector in the US (Litjens 2005). In this context, we argue that the role of the State in the effort to internationalize and enhance the global competitiveness of some national universities should go beyond the creation of national quasi-markets.

In order to make our argument, the case of the Instituto Superior Técnico (IST) in Portugal is portrayed. IST is the school of engineering of the Technical University of Lisbon, but due to historical roots related with its institutional development, has a very high level of autonomy and can be perceived, for the sake of the analysis, as an independent higher education institution. The choice for this school is based on the similarities that this school has in terms of national prominence, longevity, and scientific and disciplinary focus with the ETH Zurich.
which will be used as a reference when relevant. However, IST is also chosen for its dissimilarities with ETH Zurich. The latter is a world research university while the former is a national research intensive higher education institution that is struggling to enhance its international activities and student population.

IST was founded in 1911 and was intended to be a prime technical knowledge base of innovation for the Portuguese economic system. Its original academic staff included German, Swiss and French nationals (Heitor et al. 2004). IST was integrated into the Technical University of Lisbon in 1930, and for most of the 20th century, had its activities constrained for most of the century by a dictatorial political regime, as well as by constant underfunding and erratic scientific and educational policies (Ruivo 1995). As in most universities in Portugal, IST is very dependent on State policies for science and higher education, and as such, efforts to internationalize its activities and campus are associated with the internationalization driven by the State (Heitor and Horta 2004). The most significant efforts to internationalize IST activities started when Portugal joined the European Union in 1986. Only then did Portugal begin to integrate and participate in international research organizations, such as the CERN, and start to become actively involved in the European Framework programs (Heitor and Horta 2004).

The internationalization effort of the Portuguese universities led by the State had another impulse. The evaluation assessment of university based research units implemented in the 1990s strongly emphasized in its evaluation process the importance to publish in international journals. This led to a change in the organizational behaviors and working attitudes at IST, particularly in the perception that there was a need to internationalize the academic activities and increase research collaboration (Horta 2008). In addition, as in other countries, associated policies increased the autonomy of universities, promoted managerial practices and entrepreneurial attitudes that fostered both inter-institutional cooperation and competition at national level (see Magalhães and Amaral 2007).

The entry of Portugal in international scientific institutions and the implementation of the national assessment of the university research units impacted the publication dynamics in international journals (Figure 2). Before these State led efforts, and taking the year of 1983 as a reference point, the number of IST authored scientific publications in international journals in mathematics, physics, electrical, mechanical and chemical engineering was less than 50. Moreover, only 9% of these were done in collaboration with colleagues based in international universities. In 1993, seven years after Portugal joined the European Union, the number of publications tripled, and 27% of these were published in collaboration with colleagues based at international universities. During both of these periods, single authored articles predominated (79% in 1983 and 56% in 1993). In 2003, not only did the production of articles more than triple compared to 1993 figures, but more articles were produced in collaboration with academics based in international universities (35%) than were single authored (34%). Also, the percentage of articles produced in collaboration with colleagues from other national universities increased from the 12% in previous decades to 25%, suggesting that not only was collaboration in research becoming more important, but also that IST academics had become more integrated into national and international networks.
That said, and in spite of the fast development of Portuguese universities in the last decades of the 20th century (Heitor and Horta 2004), the competitiveness of IST and of the majority of the Portuguese universities in the global market to bring in promising academic staff is difficult. A reason that could be a hampering factor in the attraction of international faculty is the low level of salaries practiced in Portugal. Departing from the exploratory work of Rumbley et al. (2008) we compared the salaries of German, French and Portuguese academics at entry-level and found insignificant salary differentials. In fact, if purchasing power parity is used in the calculations, the salary of Portuguese academics at entry-level when compared with their German and French counterparts can be considerably higher. Therefore, salaries are not a factor in lessening the attractiveness of Portuguese universities. Instead, the difficulties in attracting promising academic staff in the global market seem to be much more associated with overly bureaucratic recruitment processes, underlined by the low attractiveness of the Portuguese universities in terms of reputation and constrained by the scarcity of available resources to develop scholarly work (Horta 2008). The result is that unlike the global research universities, which tend to have at least one fourth of their academic staff composed by international academic staff, at IST only 2% of its academic staff is not Portuguese. At IST, not

---

6 Departing from the exploratory work and analysis of Rumbley et al (2008), we compared the academic salaries of German, French and Portuguese assistant professors. The results indicate that a German assistant professor would earn around 180 Euros more than a Portuguese Assistant professor, but a French Assistant professor would earn around 100 Euros less than a Portuguese assistant professor per month at entry-level. If Purchasing Power Parity was used to take into account standards of living then a Portuguese assistant professor would earn more than 800 Euros per month than a French assistant professor and around 550 Euros more per month than a German assistant professor. Data for the Portuguese assistant professor salary was obtained from official sources (DGAEP, 2008) and Purchasing Power Parity from the World bank website (as did Rumbley et al, 2008): http://web.worldbank.org/WEBSITE/EXTERNAL/DATASTATISTICS/ICPEXT/0,menuPK:1973757~pagePK:62002243~piPK:62002387~theSitePK:270065,00.html [Accessed on the 6th January 2009].
only is the majority of the recruited academic staff national, but the great majority also completed their doctoral degrees in Portugal (Table 3).

The characterization of the hiring process seen on Table 3, and the overwhelming recruitment of Professors that held their doctoral degree from Portuguese universities, results to a great extent from State led policies aimed to increase the qualifications of the academic staff that is still rather low in Portugal (GPEARI 2006). It is also the result of the expansion of doctoral programs in Portuguese universities, which has entailed that a substantial share of this hired academic staff are ‘inbreeds’ (the inbreeding rate at IST is very high; see Horta 2008). This may be an obstacle for the internationalization process at IST since inbreeds tend not be so connected and collaborate less with peers outside the university when compared with their non-inbred academic peers (Horta 2007). Nonetheless, the hiring of 22% of all hired academic staff who did their doctoral degrees in universities from countries with developed scientific systems is important because this academic staff is deemed to bring with them contacts and the ability to integrate themselves and others in international scientific and teaching networks (Holtermann 1996). This allows IST to respond to globalization by assuming an internationalization strategy as many European universities do where cooperation is favored to competition (van der Wende 2001), but it also reflects the institutional inability to compete globally.

<table>
<thead>
<tr>
<th>Country where the doctorate degree was earned</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal</td>
<td>74.6</td>
</tr>
<tr>
<td>United States</td>
<td>12.2</td>
</tr>
<tr>
<td>Great Britain</td>
<td>7.5</td>
</tr>
<tr>
<td>France</td>
<td>2.1</td>
</tr>
<tr>
<td>Switzerland</td>
<td>0.8</td>
</tr>
<tr>
<td>Other European countries</td>
<td>2.1</td>
</tr>
<tr>
<td>Central and South American</td>
<td>0.6</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3 - Professors taking office at IST Lisbon from 1990 to 2003, by the place where they earned their doctoral degree

Even if the international size and scope of mainly research oriented activities of IST increased strongly driven by State led policies, its student structure is still too focused on undergraduate education to enable IST to become a competitor in the international global arena (Horta 2008). Figure 3a shows that the internationalization of IST student structure is as well focused mostly at undergraduate level. In 2004, the internationalization of the student population at IST was three times larger at undergraduate level than at doctoral level. This contrasts with the situation of ETH Zurich, where from 2000 onwards the larger international student population was composed by doctoral students (Figure 3b). In 2003, there were more international doctoral students enrolled at ETH Zurich than national doctoral students (7% more and the number of international and national master students was basically the same; Figure 3d). At IST, the proportion of international doctoral students enrolled in comparison with national students was minimal (ratio of 0.16) and even lower at master’s level (ratio of 0.08) (Figure 3c).
However, the internationalization of the student population cannot be perceived only under a structural framework, but requires its size to be taken into account as well. In the context of this article, and the importance that research capability has on defining the “world classiness” of a university, the size of the doctoral student population is critical because they represent human resources that most support research activities. Doctoral students not only contribute to the research outputs of academics that they are working with, but are also likely to free the academics’ time from teaching towards research by assuming the role of teaching assistants (Austin, 2002). They also have the potential to create linkages between academics, facilitate the access to network resources, and bring in new ideas (Melin 2004). In this sense, the realities of IST and ETH Zurich are quite different. In 2004, IST had 83 international doctoral students enrolled in various doctoral programs while ETH Zurich had 1307 in 2003 (Figure 3a and Figure 3b).

Finally, and also unlike ETH Zurich which enrolls international students from all over the whole world, the internationalization of IST’s student population is strongly based on the incoming of students from Portuguese speaking countries. In 2004, of the whole international
undergraduate student population at IST, 80% were international students from Portuguese speaking countries such as Brazil, Angola or Mozambique. In fact, the majority of the international students at IST are from these countries as they represent 68% of the whole international student population. The sharing of the same language seems to be critical in the attraction of these students, although its importance as a central factor in attracting international students at graduate level wanes since at IST only 30% of the total graduate international students come from Portuguese speaking countries.

The structure of the student population and particularly the structure of the international student population imply that the research capability of these two universities (at least in terms of human resources) is quite different, and in turn, so is their positioning in the global higher education market in terms of their competition (see Marginson 2006). The lower international prestige and inadequate student structure to compete directly with the top world and European research universities of IST is evident. Despite the reinforcement of graduate education in the last decades of the 20th century, as well as the strong emphasis in further internationalizing master and doctoral programs (UTL 2005) the ability of IST, or the Technical University of Lisbon itself, to do it on its own is discouraging.

This is more related to a matter of lack of financial resources than a lack of will and institutional planning (UTL 2005). The total income of IST in 2004 neared the 98 million Euros mark, of which about 52 million Euros were provided by the State. In the same year, the total income of ETH Zurich amounted to more than 704 million Euros, of which about 579 million Euros was provided by the State (i.e.: Swiss Confederation Funding). Not only the total income obtained by a top research university such as ETH Zurich is seven times larger than the one of IST, but the contribution of the State is also more substantial (82% in the case of ETH Zurich compared to 52% at IST)\(^7\). The amount of income that ETH Zurich receives from the Swiss state is also eleven times larger than the amount of income that IST receives from the Portuguese government.

Two main insights can be withdrawn from the analysis of these numbers. The first is that with the current amount of income IST cannot aspire to improve its activities in a way that allow it to compete at the same level with ETH Zurich and other top European and US research universities. Second, and perhaps most important, for both cases the role of the State is or can be decisive.

In the case of ETH Zurich the role of the state is decisive because it provides large public funds that help to ensure the positioning of the university as a top research university in the world. It is not unreasonable to argue that without the support of the State, the quality, internationalization and reputation of the teaching and research work performed at ETH Zurich

---

would suffer. In the case of IST, although a substantial amount of resources are obtained from other sources than the State, they are clearly not enough to reach decisive levels of investment that allow the change in the student structure and its further internationalization. IST on its own with or without the current funding from the state cannot compete internationally with major European and US universities. The IST case provides a paradigmatic example to sustain the argument that if a country wants to have one or a few more world class universities, then the State has to invest in the “creation” of those universities. While doing this it is required to take into account the characteristics of existing research universities, but the process needs to remain faithful to the context of the higher education, scientific and economic systems of that country and the university requires to be aware of its “glonacal” levels of action (Ahola 2005). This indicates that internationalization processes and policies require more than funding, but that without funding there is hardly even a fair start to competitive ‘races’.

In Portugal, the state is once again acting as a driver of internationalization in the Portuguese higher education and scientific system. It has signed protocols with three US research universities – Carnegie Mellon University, Massachusetts Institute of Technology and University of Texas at Austin – as a means to promote the national scientific and technological capability, but the objective is also to facilitate the internationalization of the Portuguese universities. The aim of the program of international partnerships developed between the Portuguese government and the above mentioned US research universities is to establish advanced tertiary education and research networks that can facilitate the recruitment of academic staff and researchers. The program involves several Portuguese higher education institutions, but includes the participation of State Laboratories, industry and Technological development agencies. The contracts with the US universities have an initial duration of five years and an estimated budget of 141 million Euros. Although it is too soon to perceive the impact of this State driven internationalization plan, its sole implementation and enthusiastic adherence of the Portuguese universities to it underlines the critical role that the state still has in higher education. It also underlines that in spite of the changes in higher education in the last 50 years, the sustainability of a national “world class” university in the global arena cannot be achieved without the strong commitment of the state and public funding.

Conclusion

In this article, we show that the internationalization of national research oriented universities requires to be focused on graduate education and research activities, if they aim to compete in the global higher education market. The analysis of the major research universities in Europe shows that the internationalization of the campuses is dual: student population and academic

---

8 For an analysis of the role of the State in the internationalization of Swiss universities and the Swiss academic research see Horváth et al. 2000

staff. On what concerns the student population, one observes that the internationalization of the student population is focused at graduate level, mainly at doctoral level in order to support the universities research activities. Our analysis also shows that there is a relationship between the internationalization of the academic staff and the internationalization of the student population. The internationalization of the academic staff is strongly and positively associated with the internationalization of the doctoral student population. This indicates that the recruited academic staff is hired to perform research and teach at graduate level.

However, two facts are important in terms of accounting for these results:

1) one cannot assume that the internationalization of the academic staff enhances the internationalization of the student population. There is a relationship between them, but one cannot infer causality from that relationship since the methods of analysis were based on simple correlations. More information would be necessary to perform an analysis that would allow inferring causality and a better understanding of the nuances involved in the relationship between the internationalization of academic staff and student population;

2) the case of the ETH Zurich represents a case of a European research university where some lessons can be learned. Its internationalization and research capacity evolution is path dependent and cannot be withdrawn from its national and institutional contexts. Although ETH Zurich and IST share many characteristics, the future internationalization path of universities such as IST should take into account both the critical characteristics of the top research universities in the world, but also the characteristics of the S&T and higher education systems as well as the national and local social and economic context.

Finally, both for a “world class university” such as ETH Zurich and a prominent national research higher education institution such IST, the role of the State, through funding and higher education internationalizing policy initiatives seems to be critical. Although the role of the State in higher education has been changing gradually, it can still be an engine of internationalization of national higher education institutions, and only the State can provide the financial resources for a country to build or maintain a competitive “world class university” in the global arena.

Acknowledgements

I would like to thank Emma Uprichard for her valuable comments on the manuscript.

References


Chapter 11:
Manuel Heitor, Hugo Horta and Joana Mendonça, “Developing human capital and research capacity: science policies promoting brain gain”.
Developing human capital and research capacity: science policies promoting brain gain

Manuel Heitor¹, Hugo Horta² and Joana Mendonça³

A version of this chapter was accepted for publication in Technological Forecasting and Social Change

Abstract

Science policies emphasizing the advanced qualification of human resources, together with democratizing access to science and internationalizing the science base, are shown to help build the conditions needed to drive brain gain over time.

Adopting a dynamic approach and exploring a new set of data on Portugal for the period 1970-2010, this paper focuses on the analysis of flows of doctorates, with the ultimate goal of helping to promote the absorptive capacity that emerging regions and countries worldwide need to acquire to learn how to use science for economic development. It shows a notable process of brain gain by the end of this period and, above all, it provides a dynamic approach to the cumulative process of building knowledge-based societies. The results show the need to consider the co-evolution of brain gain, brain drain and brain circulation over time and space. In addition, they suggest the importance of certain major counter-intuitive policy instruments to facilitate the co-evolution of human capital formation and research capacity building. In the case of Portugal, these instruments have included a centralized program of research grants, research careers independent of traditional academic career tracks, and a diversified system of funding research units and institutions based on research assessments through international peer reviews.

² Center for Innovation, Technology and Policy Research, Instituto Superior Técnico, Technical University of Lisbon; hugo.horta@ist.ut.pt
³ Center for Innovation, Technology and Policy Research, Instituto Superior Técnico, Technical University of Lisbon; previously at the Observatory for Science and Technology; Statistical Office, Ministry of Science and Education, Portugal
1. Introduction

The main argument of this paper is the need to focus science policies in developing regions on the process of building advanced human capital, which requires stable public strategies over time, together with adaptable and resilient research institutions.

Critical mass is vital for the creation and dissemination of knowledge, and attaining that critical mass is of the utmost importance for both developed and developing countries, but is particularly relevant for emerging regions worldwide [1, 2, 3]. The key role for policy-makers and governments in regions where major investments in science and higher education are being made is to select priority actions and make appropriate decisions as to where and how to start the process.

Our evidence and lessons learned show that achieving this ultimate goal requires science policies and strategies aimed at advanced human capital in a context in which alliances and partnerships between research organizations and higher education institutions, as well as between them and corporations, are of particular importance. Nevertheless, the main public policy issue relates to the conditions for fostering brain gain. This has become a politically sensitive issue worldwide and a critical one in developing and emerging regions [4].

It is against this background that our ultimate goal is to help understand the dynamics of brain gain, brain drain and brain circulation over time and space in order to deepen our current knowledge of policies that can drive brain gain in emerging regions worldwide [5, 6]. This is important because brain gain depends on a cumulative process that is associated with efforts to build knowledge-based societies [7]. In this context, our research hypothesis underlines that aspects of time and space and processes associated with the co-evolution of human capital formation and institutional building help promote the absorptive capacity that regions and countries need to acquire in order to learn how to use science for economic development.

It is clear that public policies to attract and retain talented people have mostly been focused on favorable immigration strategies [8]. We acknowledge the importance of such policies, but focus our argument on the need to promote science policies based on building advanced human capital and the internationalization of the science base.

Two further issues drove our research work and should be emphasized, as follows. First, innovation must be considered together with competence building and advanced training of individual skills through the complex interactions between formal and informal qualifications. This requires broadening the social basis for knowledge activities, including higher education enrolment, as well as strengthening the top of the research system to foster knowledge production at the highest level. It is well known that the world’s most developed regions (such as Japan and the US) have high rates of researchers per 1000 workforce, and are striving to increase these rates even further [9].
Second, strengthening experimentation in international knowledge networks necessarily involves flows of people. It is organized cooperation among networks of knowledge workers, together with different arrays of users, that will help diffuse innovation in the design of products and services. But establishing these innovating communities requires the systematic development of routines of collaboration on the basis of formal education programs, sophisticated research projects, and a diversified and non-structured array of informal networking processes. This requires public policies, particularly science policies, to foster “brain circulation” between leading institutions worldwide.

As this point, we should remember that scientific progress is a source of development and that tertiary education institutions play a critical role in this process [10]. The investment of public resources in the context of rigorous international assessments leads to new knowledge, better advanced training of new human resources for society, and new ideas and processes, which increasingly result in innovation, modernization of institutions, improved quality of life, economic productivity and better employment opportunities [11].

Consequently, our goals require the renovation and expansion of the social basis for scientific and technological development. This calls for strong commitment not only from the scientific and technical professions and from public and private research organizations, but also from students and from the general population. The growing appropriation of scientific and technological culture by society was one of the central aspects of the analysis by Heitor [12] and here we explore that central principle, making use of a new set of data.

It should be noted that any analysis of this subject needs to be context-sensitive, taking into account changes in the mobility of talent and corresponding perceptions. The literature on the topic is itself in transformation, from the apparently one-way, inherently competitive and mutually exclusive mobility of the highly skilled (as understood as brain drain in both the internationalist and nationalistic views of Johnson [13] and Grubel and Scott [14]) to brain circulation [15]. In this framework, this paper is not intended to provide any type of recipe. Rather, it discusses lessons learned based on a historical analysis and under the assumption that science is contextualized [16]. As a result, we consider that knowledge diffusion processes are context-sensitive and should be oriented towards inclusive learning [17].

We base our analysis on a set of data on flows of doctorates in Portugal over the period 1970-2010. Portugal represents an interesting case study since continuous investment in science and technology in a small south-western European region has resulted in significant progress after some four decades of lagging behind [18]. In addition, a thorough legal reform of the Portuguese tertiary education system has been successfully completed [18, 19] and there has been a uniquely large increase in public investment in science and technology [20].

The rest of this paper is divided into five parts. The following section describes the new set of data used for our analysis of Portugal in the period 1970-2010, together with the research methodology. Section 3
presents data up to 2010, the evidence on the achievement of brain gain after four decades of investing in the science base. Section 4 analyzes the evolution over time of relatively unstable science policies over those four decades, and Section 5 discusses structural factors that determine research capacity. In both sections, we use the number of doctorates as the main proxy for our analysis. The paper concludes with our main summarizing remarks and lessons learned in terms of the need for science policies in many different regions of the world to focus on advanced human capital formation and research capacity.

2. Methodology and data

The data used in this paper were made available by the services of the Ministry of Science, Technology and Higher Education responsible for gathering and publishing statistics [21]. They include identification of the professional situation of all doctorate holders whose studies were awarded at, or recognized by, Portuguese universities over the last four decades, as well as those working in Portugal in 2009. This identification was based on analysis of microdata from various sources of statistical and administrative information, on the basis of which all doctorate holders and their corresponding professional situation in 2009 were identified.

These doctorate holders can be divided into three groups, namely: 1) PhDs whose doctorates were awarded by Portuguese universities; 2) PhDs who gained their doctorate abroad and requested recognition, equivalence or registration of their degree in Portugal; 3) PhDs who gained their doctorate abroad and did not officially register their degree in Portugal.

The identities of the PhDs in each of these groups and data on their mobility and professional situation came from different information sources, gathered directly or administratively, as follows:

i) the official annual R&D survey of all entities that carry out research and development (governmental, higher education, private not-for-profit institutions and companies), including information on all their individual researchers;

ii) the national register of ongoing doctoral theses, which is an annual survey of all higher education institutions that collects data on ongoing and completed doctoral theses, with information on the researcher, institution(s), field of science, supervisor(s) and abstract. This survey feeds annual information to the dataset of all PhDs whose degrees were gained at or recognized by Portuguese universities since 1970;

iii) the survey of careers of doctorate holders, which is a statistical operation that collects information on the categorization of PhDs, their academic training, professional situation, international mobility and scientific production;

iv) the biographical records of teachers in higher education, which collects annual information on the composition of the academic staff of all higher education establishments;
v) the employer-employee matched dataset collected annually by the Ministry of Labor and Social Security (quadros de pessoal), with information on firms, establishments and workforce, including information on individual employees’ education levels. Data on employees with PhDs was confirmed directly with the companies hiring these individuals;

vi) administrative employment registers in the public administration sector, which include the list of entities hiring PhDs, who were contacted individually to provide detailed information about each individual;

vii) administrative data from the Portuguese Science and Technology Foundation (FCT), which include biographical data on post-doctoral fellowships funded directly by FCT, post-doctoral fellowships and contracts of PhD researchers funded by scientific research projects supported by FCT, and PhD researchers contracted by Associate Laboratories and research units supported by FCT (specifically in the period 2007-2009);

viii) administrative information on teaching staff at non-tertiary education institutions, gathered through the services of the Ministry of Education.

After the collection of data from these sources, a new dataset was built with a total of 19,876 doctorate holders identified by name and including their professional situation, work location, institution and job description. Individuals not found through these sources were searched through detailed internet surveys, which recovered information on a small number of individuals. Even so, the professional situation and workplace of 654 PhDs in 2009 could not be determined, which corresponds to 3.3% of the total number of PhDs identified.

3. Results: accumulated data

Table 1 reports the professional situation of all the doctorate holders identified, which indicates a brain gain for Portugal in 2009. Out of a total of 19,876 PhD holders, only 669 were found to be working abroad (3.4%), while 1,836 foreign PhDs were working in Portugal. It should also be noted that 83% of these foreign PhDs are working in R&D-related activities [21]. In other words, the data on PhD holders in Portugal over the last 40 years show the increasing capacity of the country to attract talent and employ doctorates.

[INSERT TABLE 1 HERE]

In addition, the data show that the great majority of PhDs who gained their degree at Portuguese universities since 1970 were involved in R&D in Portugal (86%), mostly in higher education and research institutions, a total of 17,010 doctorates. Of these, 13,888 were working in the public and private higher
education sector, while 2,427 were researchers in the private sector, and 695 were working in R&D in state laboratories and other public entities. In addition, only 4.5% of the total were not working in R&D, and 3.3% were retired [21].

The majority of PhDs whose degrees were gained abroad and recognized by Portuguese universities between 1970 and 2008 are working in Portugal (76.3%), also mainly at higher education and research institutions. Of these, 313 were foreigners. Additionally, 1,523 foreign PhDs were identified as working in R&D in Portugal in 2009, although they did not apply for their degree to be recognized by a Portuguese university.

These results show that, in terms of flows of PhDs, a total of 1,836 PhDs working in Portugal in 2009 were identified as either foreign nationals or Portuguese nationals whose doctorate was gained exclusively abroad. It should be borne in mind that this value is probably an underestimate, since there may be other PhDs who concluded their studies abroad and are working in Portugal, but who have not registered their doctorate and were therefore not identified.

There has thus been a positive flow of PhDs into Portugal (brain gain), mainly consisting of active researchers working in higher education or private research institutions. In order to better understand these results and to contextualize them in terms of developments in the Portuguese situation, it should be noted that Portugal has overcome a gap in scientific and technological development that had plagued the country for decades, indeed centuries, and has exceeded the OECD average in terms of researchers per thousand workforce [18].

That small statistic -- going from 1.5 full-time researchers per thousand workforce in the late 1980s, to 3.5 in 2005, to 8.2 in 2010 -- tells an important story about how countries and regions can compete and level the playing field in the dynamic global economy. At the same time, the Portuguese tertiary education system has been reformed, the social basis for students recruitment has been enlarged and industry-science links have been strengthened, together with increases in business expenditure on R&D (which represented 0.8% of GDP in 2009, compared to 0.3% in 2005 and less than 0.2% until some ten years ago [20]). In this process of technological change, Portugal has seen the creation and nurturing of opportunities for research and advanced training through strategic partnerships with leading partners worldwide. These cover diverse areas, from deep-sea biotechnology in the North Atlantic to the internet of the future, and involve building further competencies in nano- and biosciences, as well as in engineering systems and advanced computing [22].

In general, an overview of the recent evolution of the Portuguese S&T system, and an explanation of the background of how Portugal achieved a brain gain of doctorates, can be gained from figures drawn from various reports produced by the statistics office of “ministry of science and technology”. We start with
the scientific output of Portuguese research institutions in all scientific fields\textsuperscript{4}, which reached 10,081 publications in 2009, from 6,450 publications in 2005, and just 3,792 in 2000. Also, that scientific output, as measured by the number of internationally refereed scientific publications in science and technology fields\textsuperscript{5}, reached 7,470 articles, letters, notes and reviews in 2009, up from 2,702 in 2000.

Analysis also shows that: i) internationally refereed Portuguese articles, letters, notes and reviews in the exact sciences, natural sciences, health sciences, agriculture and engineering have increased nearly 2.8 times since 2000; ii) this growth can also be observed in the number of publications by total population, which reached 703 articles per million in 2009, from 373 in 2004; iii) the number of articles by total population is now 77% of the EU-27 average, whereas it was only 51% in 2004, suggesting that Portugal’s science base in the fields of science, technology, engineering, and mathematics is becoming internationally competitive.

The above figures reveal the outcomes of public policies designed to foster investment in people and institutions. They reflect the growing number of researchers in Portugal, which reached 8.2 per thousand workforce in 2010 (about 46,000 full-time equivalent [FTE] researchers, nearly a quarter of them in the business sector). This figure is now above the EU and OECD averages, and is similar to (and in some cases even higher than) the levels of Austria, France, and the US.

However, the evolution is even more wide-ranging, as in recent years Portugal has had the highest percentage growth rate in Europe in terms of the total number of researchers (in FTE) per thousand workforce (about 95%), well above the European average (which grew by only 14% from 2003 to 2008), Spain (21%) and Ireland (13%). There has been an increase of 25% over the last two years in the number of researchers working in academic R&D centers (12,000 FTE doctorate researchers) and a doubling of the number of doctorate researchers since 2000.

The number of researchers (headcount) has also increased in all scientific areas since 2005. In 2008, 69% of all researchers in Portugal were performing research in the exact, natural, and agrarian sciences and engineering. However, despite accelerated growth in human resources for science and technology, and particularly of new PhDs, the growth rate in Portugal has remained relatively low compared with other European countries and US states. Even so, more doctorates are awarded per year in Portugal than in some comparable American states, such as Florida or New Jersey, having reached that position from a very weak starting point: in 1990, Portugal turned out a mere 0.68 new PhDs per ten thousand workforce [18].

\textsuperscript{4} As measured by the Science Citation Index Expanded (SCI), together with other databases such as the Social Science Citation Index (SSCI) and the Arts & Humanities Citation Index (AHCI) using the fractional counting method.

\textsuperscript{5} As measured by the SCI in relation to scientific publications in the exact, natural, and health sciences, as well as agriculture and engineering.
4. **Evolution over time - Learning from an evolutionary path**

This section makes use of historical data for the last four decades, in order to shed light on the net accumulated results presented above. First, we look at annual data in terms of human capital formation, using the number of doctorates as the main proxy for our analysis. Second, we identify four major periods of science policy and science and technology development in Portugal, which are briefly described in a way that emphasizes key issues associated with human capital formation and institutional capacity.

4.1 **Annual data: Portugal 1970-2010**

Figure 1 quantifies the number of doctorates awarded or recognized annually at Portuguese universities between 1970 and 2008. While this number was below 100 in 1970, it exceeded 1000 per year by 2003 and 1500 by 2008. Between 2000 and 2010, the number of doctorates grew by more than 74%.

Figure 1 also shows the increasing capacity to train PhDs at a national level by an increasing number of universities. Until the late 1970s only the four oldest universities – the universities of Coimbra, Lisbon and Porto and the Technical University of Lisbon – had the capacity to award doctorates, even though the universities created in the early 1970s were also entitled to do so. Forty years later, 50% of all the doctorates awarded in Portugal are obtained at the oldest universities, which is typical of the evolution of other higher education systems worldwide [23, 24]. This pattern alone shows the importance of time in the implementation of policies. Another important fact from this analysis is that the number of PhD degrees obtained abroad and registered in Portuguese universities continues to grow, even though its share vis-à-vis the total number of new PhDs is increasingly lower, another indicator of the system’s evolving capacity to train PhDs.

![INSERT FIGURE 1 HERE]

Figure 2 quantifies the ratio of PhDs obtained abroad to PhDs obtained in Portugal. While in the early 1980s the number of doctorates awarded abroad and recognized by Portuguese universities was similar to the number earned at Portuguese universities (about two hundred per year), in the last decade the capacity for awarding doctorates in Portugal has risen considerably, exceeding 1300 new doctorates in 2008. At the same time, over the period under analysis, the number of new doctorates per year obtained abroad and recognized by Portuguese universities has remained nearly constant.

This reveals two major points with policy implications. First, until the mid-1980s the Portuguese higher education system did not have the capacity to train doctorates in general, and there was a clear lack of
critical mass in many scientific areas. Second, mobility at doctoral level was treated as a strategic policy, to establish a science base in Portugal on the return of the doctorates, but also as a means to internationalize the Portuguese scientific and academic communities [25].

In the Portuguese case, in several universities disciplinary areas were created or fostered by Portuguese doctorates who had obtained their doctorates abroad [26]. Importantly, this mobility strategy was strongly supported by public policies that awarded a growing number of doctoral fellowships focused on the advanced training of human resources (see Figure 2), but also on fostering internationalization and integration of Portuguese science into global research networks [27]. This was complemented in the initial years by grants and fellowships provided by private foundations, which have been active since the 1950s [28]. Again, this is a pattern that is echoed in other countries, such as South Korea, where highly skilled returnees are playing a major role in South Korean higher education by developing research areas, teaching programs and international networks [29].

[INSERT FIGURE 2 HERE]

Regarding the annual numbers of new PhDs in relation to the country’s workforce, Portuguese universities awarded 2.7 new doctorates per ten thousand active workforce in 2008, whereas about twenty years ago, in 1990, Portugal produced only 0.68 new doctorates per ten thousand active workforce. In comparative terms, the state of Massachusetts (USA), Switzerland, Slovakia, Sweden and Finland produced more than twice as many doctorates per ten thousand active workforce in 2008. Particularly revealing is the number of new PhDs per ten thousand active workforce in the state of Massachusetts (7.6), and in Switzerland (6.8), which is practically three times the number in Portugal. Still, Portuguese universities granted more doctoral degrees in 2008 than some North American states, such as Florida and New Jersey, having reached this position starting from numbers that were very low until about ten years ago.

An obvious outcome of public investment in the advanced training of human resources in Portugal has been the growing qualification of academic staff at Portuguese higher education institutions. The percentage of academic staff holding a doctorate degree had reached 68% in public universities, 39% in private universities, and 19% in both public and private polytechnics by 2009, as opposed to 48% at public universities, 21% at private universities, and roughly 8% in both public and private polytechnics in 2001.

It should be noted that improvements in qualifications of academic staff have naturally been achieved in a context of “academic inbreeding” practices at the oldest universities, as measured broadly as the percentage of academic staff holding a doctorate from the same university. This practice is usually
referred to in the literature as harmful to research productivity and collaboration [30]. It is still endemic in Portuguese academia, as acknowledged in the latest OECD tertiary education review of the Portuguese higher education system, which considered Portuguese universities as “too academic and inward looking, resulting in a high degree of insularity and inbreeding” [19: 146].

Figure 3 presents evidence of the resilience of academic inbreeding with reference to the specific case of the Faculty of Engineering of the University of Porto, where academic inbreeding rates were high over a long period of time, remaining stable from the early 1990s up to 2010, ranging from 40% to 50% of the academic staff holding a doctorate, particularly at the assistant and associate professorship levels, suggesting a trend that will endure in future years.

This figure, however, is particularly relevant when considered together with Figure 2, in the sense that academic inbreeding levels start to increase when most PhDs begin to be awarded in Portugal. The issue for policy is to what extent universities are able to change academic recruitment practices that are strongly rooted in institutional cultures and to what extent public policies can influence this change for the years to come.

In addition, it should be noted that the data in Figure 3 do not include doctorate researchers independent of traditional academic careers. These doctorate holders have been critical for the development of Portuguese research capacity and the increasing scientific productivity of most research units and institutes, despite their temporary and insecure employment situation [31]. It is clear that they are liable to move from one institution to another and help to reduce levels of inbreeding in the oldest university schools [32]. This suggests a need for further policy instruments directed towards their integration into academic staff, as in a way to help renew the oldest institutions.

Returning to the qualification of academic staff, another phenomenon that needs to be taken into account is that average figures at a systemic level can hide different rates of evolution at the institutional level and in individual disciplinary fields. For example, Figure 4 quantifies the qualifications of the academic staff of the faculties of medicine, engineering and law in three different universities and suggests that there are distinct institutional and disciplinary paths. On average, engineering faculties have been the fastest in adopting the PhD track for their staff, while in faculties of medicine the qualifications of their academic staff have either stagnated or actually worsened. This is an indication of the importance of specific disciplinary cultures, norms and habits [33], although the Portuguese case is also influenced by different levels of engagement in research [34].
4.2 Patterns in development of science policy and science and technology in Portugal

Following the chronological framework set out in Heitor and Horta [18], as inspired by Ruivo [35], the development of the Portuguese S&T system since the early 1970s, and its relationship with higher education and Portuguese society in general, can be divided into four main periods, as described in Table 2. In short, the S&T system was only effectively established after independent international research assessments started in the mid-1990s. In this respect, the first decade of the 21st century consolidated and strengthened earlier policies associated with vigorous public investment in R&D, together with a strong drive towards internationalization.

It should be noted that before the early 1970s we could only speak of a residual science base, with minor incentives for R&D in a non-integrated system, as designed by a totalitarian political regime, the Estado Novo. This regime was structurally averse to scientific knowledge, in which state laboratories were the main centers of scientific activity and universities were blocked from scientific development. As a result, the analysis in this paper starts in the early 1970s.

1970-1985: Early attempts at growth, with 50% PhDs abroad and brain drain

The National Committee of Scientific and Technological Research (Junta Nacional de Investigação Científica e Tecnológica, JNICT) established in 1967, marked the beginning of science planning in Portugal. It was the outcome of several NATO studies during the early 1960s and was driven forward by the OECD project “Pilot Teams in Sciences and Technology” for Portugal, as commissioned by Minister Francisco Leite Pinto. This followed the Regional Mediterranean Project, which had focused on the conditions of education in Portugal and other southern European regions. But, at that time, overall expenditure on R&D was only about 0.25% of GDP, one of the lowest among European countries.

In the absence of incentives to promote the social and economic dimensions of the search for knowledge, a major education reform was launched in 1973 (the “Veiga Simão Reform”), which laid the foundation for the Portuguese science base. Grounded in an expanding higher education system, including the creation of new universities, this reform was shaped by a legal enactment recognizing the equivalence of doctoral degrees obtained abroad and restructuring the career path of academic staff. It should be noted that these legislative acts were passed more than ten years after Manuel Rocha's
statement to the first congress of engineering education that “the fundamental aim of the university is to teach and disseminate culture, and this function cannot be performed without research activities” [36].

These ideals were only to become reality in 1979 when the Statute on the careers of academic staff in universities took effect. This gave final and formal expression to the obligation on academic staff at university wholly and exclusively to undertake teaching and research. We hold this step to be decisive in establishing the science base in Portugal. It stipulated the necessary and basic conditions for R&D to be effectively performed in universities. As a pointer to the nascent dynamism thus unleashed, between 1967 and 1986 overall expenditure on R&D only rose from 0.25% to 0.36% of GDP. In 1982, across all subject areas, the country mustered 5,736 research personnel and 9,258 by 1986.

1986-1995: The late awakening of the science base, still with brain drain and inbreeding

Joining the European Union represented an important opportunity for Portuguese S&T development. The period from 1986 to 1989 saw science policies guided by a more complex model of technological change and a strengthening of international cooperation, including membership of international organizations such as the Conseil Européen pour la Recherche Nucléaire (CERN). Under the leadership of José Mariano Gago, JNICT’s President at the time, a “Mobilizing Program for Science and Technology” defined and implemented a series of S&T priorities and projects in specific areas. In the early 1990s, a number of new programs were implemented, supported by European Union structural funds, including a major initiative of individual research fellowships [25]. The CIÊNCIA Program, between 1990 and 1993, promoted advanced training and the construction of physical infrastructure. Under this program, 3,204 fellowships – roughly half at PhD level – were granted and brought about a considerable increase in the numbers of Portuguese research staff. Several of these fellowships supported doctoral degrees abroad (54% of the total PhD fellowships awarded).

This period is characterized by academic inbreeding in the oldest universities, while university-science relations played a significant role in building the knowledge base at research and higher education institutions. Still, by 1995, compared to other European countries, Portugal presented a relatively low R&D effort, and its outlay on R&D remained below 1% of GDP, the second lowest figure among OECD countries. The institutional rigidity of universities led to the emergence of new interface institutions to draw on European Union funds, to encourage the flexible transfer of technology and, above all, to hire research staff.

1996-2005: Struggling towards the European average, promoting brain circulation, still with inbreeding
In 1995, the incoming Government created the Ministry of Science and Technology, led by José Mariano Gago, a move that led to profound changes in public institutions associated with science and technology. The Portuguese S&T system was stimulated further by fundamental reforms in the assessment of R&D institutions. The new assessment system for R&D units, using international reviewers and linked to research funding, the publication of the methodologies employed and the results obtained, together with the right of appeal, guaranteed the independence and effectiveness of research evaluation. Investment in science increased significantly: for example, public funding for university-based research units rose from 7.5 million euros in 1995 to 25.5 million euros in 1999. In 1999, overall expenditure on R&D accounted for 0.76% of GDP, though this was still well below the European Union average of 1.74%.

The assessment exercise of 1999-2000 confirmed that successive assessments of S&T institutions since 1996 had injected a dynamic of change into the Portuguese research community, a change that had brought about a rapid increase in the numbers of young doctorates, doctoral students, and international connections, albeit still with strong patterns of inbreeding at leading institutions [37]. The steady increase in the number of doctorates, compared to European and international levels, was seen by the majority of assessment panels as decisive in gaining critical mass in scientific development. Even so, in 2000, the number of research staff plotted against the workforce still remained about half the European average –2.9 per thousand workforce compared to 4.9.

2006-2010: Fostering knowledge-integrated communities, moving towards brain gain

At the end of the first decade of the new millennium, science investment in Portugal took on a new lease of life, breaking with earlier pattern of relatively sluggish or fluctuating investments, and reaching unprecedented levels of development.

Above all, the build-up over two decades of public funding for the advanced training of human resources and the establishment of new scientific institutions began to bear fruit. Particularly clear was the impact on qualifications and on modernization of both higher education and business-based R&D, which increased considerably during these years. In 2006, the historical tipping point of 1% of GDP invested in R&D was finally reached. Three years later, in 2009, it was to attain 1.64%. Thus, Portugal overtook countries that historically had always invested more in R&D, among them Italy and Spain, with 1.26% and 1.39% respectively [38]. At the same time, the share of business in gross expenditure on R&D increased from about 20% to 50%, representing in absolute terms a threefold increase in business expenditure on R&D, which rose from about 400 million euros in 2005 to 1.3 billion euros in 2009.

The recent acceleration in the development of the Portuguese S&T system went hand in hand with its capacity to attract and train human resources and clearly bolstered the critical mass factor in academic institutions. It launched new provisions which brought researchers together across disciplines and built
up knowledge-integrated communities with an increasingly marked international outreach. Thus, in Portugal research staff, expressed as the number of researchers per thousand workforce, reached the OECD average. In 2009, this stood at 7.9 per thousand, a level similar to – even higher than in some cases – Spain, Ireland, Italy, Germany, the Netherlands, and the United Kingdom [38].

5. Analysis and discussion: Structural factors determining research capacity

This section discusses the results presented above in terms of a novel contribution to the design of science policies in emerging regions worldwide. The data used are the basis for a dynamic approach to the cumulative process of attempting to build knowledge-based societies, and show the need to consider the co-evolution of brain gain, brain drain and brain circulation over time and space.

In the case of Portugal, it took almost four decades to achieve reasonable levels of investment in science and technology and to overcome the country’s continuous lagging behind international standards. We argue that other regions worldwide can accelerate this process, if appropriate policy measures are systematically adopted to facilitate the co-evolution of human capital formation and institutional capacity building. In order to achieve these goals, the results presented above are discussed in the following paragraphs under four main headings, emphasizing major policy instruments that may be used in emerging regions worldwide to foster science and technology. We start by looking at levels of public investment, then move on to specific programs for human capital formation, including initiatives to promote scientific employment. Specific policies and instruments to foster research capacity through institutional building are also discussed. In addition, the need to promote science awareness in general is briefly addressed in terms of the challenges associated with the social construction of knowledge-based societies. We conclude our discussion with a brief identification of remaining challenges.

5.1 Levels of public investment in R&D and the maturity of a meritocratic culture

In the early 21st century, investment in science and technology in Portugal has reached – and gone beyond – the long-awaited moment when the amount set aside for R&D topped 1% of GDP. In 2009, gross expenditure on R&D attained 1.64%. Figure 5 shows that there were three main periods of investment growth, of different intensities (1986-1992, 1995-2002, and 2004-2011), separated by periods of stagnation or reduction (1980-1986, 1992-1995, and 2002-2004), while Table 3 briefly characterizes the focus of policy instruments throughout each of these periods.

In the first period of growth, gross expenditure on research and development (GERD) as a percentage of GDP took ten years to double, from 1982 to 1992. At that time, the main science policy instrument was based on doctoral fellowships, a large percentage of which were still spent abroad.

In the second period of growth, from 1995 to 2002, GERD as a percentage of GDP grew by about 50% and new policy instruments were introduced, including post-doctoral fellowships, external independent
assessments of research units with funding implications (based on three-year contracts), which was considered essential for the consolidation, integration and modernization of the Portuguese research system, and the creation and funding of large research laboratories with ten-year contracts (Associate Laboratories). From this time onwards, public policy maintained international research assessment exercises as a regular practice, assessing the output of scientific establishments and concentrating on achieving a critical mass of doctoral researchers across disciplines and institutions. Priority was consistently given to creating job opportunities in science to bring new blood into the science domain and to achieve critical mass. Much emphasis was placed on setting up international partnerships to foster scientific networks and industry-university partnerships, and on strengthening the bonds between graduate education and research.

After a period of stagnation, the third growth period, starting in 2005, saw the heaviest investment in science. By 2009, GERD as a percentage of GDP had doubled, mainly through existing policy instruments, although enlarged through additional public funding. The number of Associate Laboratories was increased, and new policy instruments were introduced. These included a major program to foster scientific employment through doctorate researchers on five-year contracts, the establishment of Research Chairs, and a program of international partnerships with prominent research universities and institutions worldwide.

As a result of the accumulation of public investment, the last few years have seen the emergence of three distinct but inter-related trends. First, there has been a remarkable increase in overall business expenditure on R&D (BERD) and in industry’s capacity to undertake research in collaboration with academic research centers. BERD rose from 425 million euros in 2005 to over 1.3 billion euros in 2009.

Second, the relative spread across the number of firms investing in R&D has grown considerably. The five most R&D-intensive firms account for only 30% of BERD, the top twenty for 59%, and the top hundred for some 80%. These figures suggest that Portuguese business R&D is not dependent on a few large companies. This is a good sign for the overall goal of raising and sustaining the business sector’s participation in the drive to increase the country’s technological intensity. Nevertheless, analysis also suggests that large companies need to increase their R&D investment significantly if science-based job opportunities in the business sector are to increase. In addition, further specialization in the skills required by emerging areas is equally necessary if business competitiveness is to improve.
Third, there has been a considerable increase in academia’s research capacity, with the number of PhD fellowships and post-doctoral research contracts more than doubling. Nevertheless, despite the impressive increase in investment in R&D in recent years, this still does not guarantee scientific maturity. Rather, given the development trajectory of Portugal’s science system, it is more appropriate to regard investment as a further step in the recovery from a late awakening and its slow – and often intermittent – progress along the path to maturity.

The recent positive trend in science investment is best understood by comparing it with other European countries, not only over the same period, but over a longer time frame as well. From this longitudinal perspective, two main results emerge. First, despite reaching the same levels of investment as Spain, Italy or Ireland, Portugal’s accumulated science investment over the past few decades is not even close to that of those countries. Building up the nation’s scientific development to a position similar to the above countries requires a far larger and sustained investment in science, at a rate faster than in those countries, and over a long period.

Second, despite the investments in S&T during the periods analyzed, Portugal is still far from the investment levels of other small and medium-sized European Union countries such as Belgium, Austria, Denmark or Finland. One indirect consequence of these two features is Portugal’s persistent “infantile status” in industry-science relationships and its present “immaturity” in both industry and academia in planning long-term joint ventures. This is affected by the structure of enterprises, as well as by the lack of large companies in sectors traditionally involved in advancing these ties in other industrialized countries.

5.2 Advanced training of human resources and human capital formation

Creating and enhancing critical mass is vital for the creation and dissemination of knowledge, and attaining that critical mass is of the utmost importance for both developed and developing countries, but is particularly essential for emerging regions worldwide. It is in this context that the main lesson to be drawn from the evolution of Portuguese research capacity towards a level characterized by a notable brain gain of advanced trained human resources is the importance of a major, long-term, publicly funded and centralized program of research fellowships. In addition, it should be noted that in recent decades it has been consistently based on independent national assessments of individual proposals, totally independently of any higher education institution.

Four main points are discussed below, namely: i) the required accumulation over time of individual fellowships at doctoral and post-doctoral levels; ii) the nature of increasingly open competitions for individual fellowships, including the need to attract foreigners; iii) the need to evolve from fellowships to research contracts, at least at post-doctoral level and on a temporary basis; and iv) the role of academic inbreeding in the oldest universities.
Figures 6 and 7 quantify the considerable public effort put into the advanced training of human resources and providing research-oriented careers, on a temporary basis, giving recent doctorates the opportunity to carry out research in a relatively independent fashion.

Nearly 14,500 doctoral fellowships and more than 4,500 post-doctoral fellowships were awarded through a centralized program of individual grants from 1995 to 2008. All of them included an option for a spell abroad doing research or studying, for varying periods ranging from three months to the full study period abroad. This favored the mobility of researchers and the growing internationalization of the Portuguese research staff and academia [25, 39]. Analysis clearly shows a gradual change in mobility patterns over time, with fewer fellowships spent fully abroad [32].

It should also be noted that, for the last fifteen years analyzed, foreigners were awarded more than 1,100 doctoral fellowships (8% of all doctoral fellowships awarded) and more than 1,500 post-doctoral fellowships (33% of all post-doctoral fellowships awarded) to do their doctorates at Portuguese universities or to engage in post-doctoral research. This has been a strategic asset in the international competition for talent and follows many other national strategies at a global level (see, for example, the Science Fellowship program in Japan, as discussed by [5]).

Our analysis also shows the importance of public policies that foster and create scientific employment through a research career path in universities. This has gone beyond traditional academic careers, in a way that facilitated the continuous recruitment of young researchers to work in university research units, following best practices internationally, but independent of internal university procedures. As a case study, investment in people through qualified human resources was particularly promoted in Portugal in the period 2007-2010 to support contractual arrangements for researchers through academic university research centers and laboratories.

Figure 8 presents the nationalities of the contracted doctorate researchers in higher education by early 2010, and shows the following main points. First, about 1,200 new PhD researchers with more than three years of post-doctoral experience were hired by 2009, of whom 41% were foreigners. Unlike the internationalization of the student population at Portuguese universities (which relies on nationals from Portuguese-speaking countries [39]), the internationalization of the research community shows a predominance of nationals from the European Union and elsewhere in the world, with nationals from
Portuguese-speaking countries accounting for only 4% of all the research contracts awarded. The hired PhD researchers were based at 264 research units in all areas of knowledge (around 43% in the natural and exact sciences and 24% in engineering and technology).

Second, this program stimulated major changes in the academic community and facilitated the renewal of teaching and research staff, although with less effect on faculty positions. The number of foreign academics in the Portuguese tertiary education system totaled 1,400 in 2009 (an increase of 26% since 2001). Additionally, the number of foreign researchers almost doubled, from nearly 1,900 in 2005 (6% of the total number of researchers) to about 3,800 in 2008 (7% of the total).

5.3 Institutional building promoting research capacity

In addition to the human resources component, our research clearly shows that the co-evolution of human capital formation and research capacity building is critical to promote the absorptive capacity that emerging regions and countries worldwide need to acquire in order to learn how to use science for economic development. In this context, a key policy instrument in the Portuguese case was a public program to fund research units by multi-annual contracts based on national research assessments, totally independent of internal university hierarchies. This is again a long-term process, requiring different institutional speeds and types of multi-annual contracts.

Table 4 quantifies the impact of the national research assessments in Portugal, showing a substantial growth in the number of R&D units, but particularly a doubling of the average number of PhDs in each unit. This helped achieve critical masses at the institutional level, with clear effects in terms of the number of articles in international peer-reviewed journals and increased science-related networking at national and international levels [40].

The growth in the number of PhDs in R&D units, many of them based at or associated with universities, occurred independently of traditional academic careers and of internal university procedures; it was based on recruitment following best international practices. Figure 9 reveals this by showing that the employment of doctorates in higher education has also followed this trend, as universities appear unable to absorb the new doctorates into their traditional academic staff.
This can be seen more clearly since 2003, when the number of new doctorates per year exceeded the 1,000 mark. Although in the short term this is sustainable – as long as there is funding available through post-doctoral fellowships and PhD research contract grants for a substantial share of the doctorates in the scientific system – in the medium and long term it could lead to a situation of brain-drain [41]. As discussed above, this requires policy instruments designed to foster scientific employment throughout the labor market, with emphasis on industry [see also 42, 43, 44, 45].

[INSERT FIGURE 9 HERE]

A landmark in terms of institutional building in the Portuguese research landscape was the establishment of Associate Laboratories in November 2000, through long-term contracts with the Portuguese Science Foundation (ten-year as opposed to the three-year contracts with typical research units). By 2001, 15 of these laboratories were active, bringing together 31 research units and more than 2,200 researchers, of whom 880 were PhD holders. By 2009, the network of scientific institutions encompassed 510 research units and 25 laboratories. Overall, institutional funding amounted to some 80 million euros, compared with 5 million euros a decade earlier.

One of the main objectives of the Associate Laboratories was to increase scientific employment by recruiting both doctorate researchers and technicians. As a result, the average number of PhDs in Associate Laboratories is twice that of general R&D units. A second objective was to develop critical mass in every scientific discipline by bringing together comparatively large research consortia engaged in thematic networks across a number of institutions, selected by international assessment. In addition, Associate Laboratories opened the way for a new science culture, grounded in institutional autonomy, upheld by incentives and urged on by regular and on-going recourse to independent scientific assessment, a culture that has been developed and implemented in OECD countries and in most established and mature science systems [46].

At this stage, by 2005, the process of institutional building was oriented towards improving links between the research system and higher education institutions and private companies, while at the same time fostering institutional internationalization, as well as a greater internationalization of the science base at an individual level. This was achieved by establishing selected partnerships with leading research universities and institutes worldwide in the form of relatively large consortia bringing together universities, R&D units, end users and companies.

Table 5 summarizes the activities of a sample group of programs implemented through the Portuguese Science and Technology Foundation, unprecedented in Portugal and with innovative features worldwide, that have opened the way for a number of thematic networks with industry and with various
Portuguese universities. In addition, these programs have been strategically important in their impact on doctoral education and advanced study programs, some of which were offered as dual degrees between leading US universities and Portuguese universities, and in bringing academics and researchers together in application-driven, collaborative research projects aimed at global markets [22].

Overall, the influence of incentive-based funding programs in promoting R&D is self-evident and has been decisive. Even so, different forms of incentives need complementary action to increase the intensity and scale of R&D. Contract research is particularly important. If research is to take on a new dynamic, other incentives have to be brought to bear. Private sector firms, in particular, require other forms of encouragement, such as tax breaks for those actively engaging in research and innovation. In short, development of the science system rests on a variety of incentives.

5.4 Remaining challenges

The above discussion is centered on building research capacity through people and institutions. We conclude by briefly introducing further challenges in the development of research investment, the intensity of which is important in the sense that it quantifies the long-term stability of the research system, as well as helping to extend our understanding of the varied ways to promote the development of a science base.

Table 6 shows that in Portugal (and also Slovakia), the growth of the research system followed a path of relatively low (and, at times, decreasing) levels of funding per researcher, in order to give priority to the need to attract people and build critical mass. Our results show the success of such a strategy, even in an international landscape increasingly characterized by high level of competition for talent. Nevertheless, it is interesting to note that other countries in different national and international contexts, such as Spain and the Czech Republic, have followed slightly different patterns, achieving higher levels of funding per researcher, although with smaller relative concentrations of researchers. Importantly, for the case of Portugal and with relevance for developing countries and regions throughout the world, our results show that it is possible to grow a research system with relatively low intensities and still be sufficiently attractive to foster brain gain.
6. Conclusions

This paper provides a dynamic approach to the process of building knowledge-based societies in developing and emerging regions and countries worldwide through a new understanding of the co-evolution of human capital formation and research capacity.

The main data refer to Portugal over the period 1970-2010 and the paper explores primary source data focused on flows of doctorates in and out of the country. It shows a notable process of brain gain by the end of that period, which is discussed in terms of a new contribution to the design of science policies in emerging regions worldwide. The analysis shows the need to consider the co-evolution of brain gain, brain drain and brain circulation over time and space. It goes beyond commonplaces, leading us to argue in favor of a few major counter-intuitive measures and science policy instruments, as follows:

1. A central finding is that public investment in science associated with policies facilitating the co-evolution of human capital formation and institutional capacity building can lead to a situation of brain gain. In the specific case of Portugal, it took almost four decades to achieve reasonable international levels of investment in science and technology and to overcome a situation of continuous lagging behind international standards. This is shown to be associated with patterns of relatively sluggish or fluctuating investments in R&D for many years, reaching unprecedented levels of development only after 2008. We argue other regions worldwide can accelerate this process, if appropriate policies are systematically adopted to facilitate the co-evolution identified in this paper. In our case, the number of researchers grew with relatively low levels of R&D funding per researcher, but at a level sufficiently attractive to foster brain gain;

2. In addition, our research suggests the key role of a major, long-term, publicly funded and centralized program of research grants, particularly for doctoral and post-doctoral grants based on independent national assessments of individual proposals, totally independently of any higher education institution. We argue that it is particularly important to implement this in the earlier and middle stages of development to prevent the investment being absorbed by the hierarchical and change-adverse environments that characterize many higher education institutions in developing regions;

3. Our analysis also shows the importance of public policies that foster a research career path in universities, independent of traditional academic careers, in order to facilitate the continuous recruitment of young researchers by academic research units following best international practices, but independent of internal university procedures. The relatively temporary and insecure nature of that research career path leads us to argue for the need for integrative measures to foster policies of gradual recruitment in the public and private labor markets;
4. Last, but not least, a public program to fund research units based on multi-annual contracts based on national research assessments, also totally independent of internal university hierarchies, helps build the institutional capacity needed in a knowledge-based society. This is again a long-term process, requiring different institutional speeds and types of multi-annual contracts.

Overall, our analysis shows that understanding aspects of time and space and processes associated with advanced training of human resources and institutional building helps to promote the absorptive capacity that regions and countries need to acquire in order to learn how to use science for economic development. In this developmental process, our data reveal that higher education institutions are the de facto main employers of PhDs and hence the advanced qualification of human resources needs to be considered in the context of full internationalization of their region/nation, in terms of facilitating the integration of those regions/nations in knowledge networks at the highest international level.

In attempting to extrapolate our results to developing and emerging regions worldwide, our analysis underlines the need to foster the advanced qualification of human resources and their international mobility during their training and early careers. In addition, we argue that science policies emphasizing the advanced qualification of human resources, together with democratizing access to science, help to build the conditions that drive modern societies.

References


[22] Heitor, M.V. (2012), "How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?", Science and Public Policy, submitted for publication.


LIST of TABLES

Table 1 – Main flows of doctorates (PhDs) in Portugal over the last four decades, 1970-2008

Table 2 – Main periods analyzed in this paper for the development of science policy in Portugal (1970-2010)

Table 3 – Major science policy instruments used over time in Portugal (1970-2010), as identified in this paper

Table 4 – Impact of national research assessments (every 3 to 4 years) in terms of number of research units, number of PhDs and average PhDs per research unit in Portugal

Table 5 – Sample of measures adopted through major international partnerships in science, technology and higher education in Portugal (2006-2011)

Table 6 – Evolution of the number of researchers (headcount) and funding per researcher (in 1000 US dollars PPP of 2008), in a sample of OECD countries for 2001 and 2008
Table 1 – Main flows of doctorates (PhDs) in Portugal over the last 40 years, 1970-2008

<table>
<thead>
<tr>
<th>1. PhDs awarded by Portuguese universities between 1970 and 2008</th>
<th>14,147</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 % PhDs working in R&amp;D-related activities in Portugal (2009)</td>
<td>86.8%</td>
</tr>
<tr>
<td>1.2 % PhDs working in non-R&amp;D-related activities in Portugal or in other circumstances (e.g. retired)</td>
<td>7.0%</td>
</tr>
<tr>
<td>1.3 % PhDs working abroad (2009)</td>
<td>3.7%</td>
</tr>
<tr>
<td>1.4* % PhDs with no identified workplace</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. PhDs awarded abroad and recognized by Portuguese universities between 1970 and 2008</th>
<th>4,206</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portuguese nationals</td>
<td>3491</td>
</tr>
<tr>
<td>Non-Portuguese nationals</td>
<td>313</td>
</tr>
<tr>
<td>2.1 % PhDs working in R&amp;D-related activities in Portugal (2009)</td>
<td>76.3%</td>
</tr>
<tr>
<td>2.2 % PhDs working in non-R&amp;D-related activities or in other circumstances (e.g. retired) in Portugal (2009)</td>
<td>13.3%</td>
</tr>
<tr>
<td>2.3 % PhDs working abroad (2009)</td>
<td>3.5%</td>
</tr>
<tr>
<td>2.4* % PhDs with no identified workplace</td>
<td>6.9%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3. Foreign PhDs with a PhD obtained abroad and not registered in Portugal or working in R&amp;D in Portugal (2009)</th>
<th>1,523</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow of doctorates:</td>
<td></td>
</tr>
<tr>
<td>Number of foreign PhDs or Portuguese nationals with a PhD obtained abroad working in Portugal (2009)</td>
<td>1,836</td>
</tr>
<tr>
<td>3.1 % foreign PhDs working in R&amp;D-related activities in Portugal (2009)</td>
<td>82.9%</td>
</tr>
<tr>
<td>3.2 % foreign PhDs working in non-R&amp;D-related activities in Portugal (2009)</td>
<td>17.1%</td>
</tr>
</tbody>
</table>

| 4. Holders of PhDs awarded or recognized by Portuguese universities working abroad in 2009 (1.3 + 2.3) | 669 |

* The maximum expected uncertainty of this analysis is associated with PhD holders with no identified workplace.

Table 2 – Main periods analyzed in this paper for the development of science policy in Portugal (1970-2010)

<table>
<thead>
<tr>
<th>Period</th>
<th>Mobility trend</th>
<th>GERD/GDP</th>
<th>Total researchers (PhDs)</th>
<th>S&amp;T policy instruments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1970-1985</td>
<td>Early attempts at growth, with 50% PhDs abroad; few PhDs in universities (brain drain); high academic inbreeding</td>
<td>0.28%</td>
<td>5,736 (NA)</td>
<td>Creation of several universities in the mid-1970s (higher education policy); University and polytechnic career statutes</td>
</tr>
<tr>
<td>1986-1995</td>
<td>Striving to increase knowledge capacity; greater mobility to international scientific organizations (e.g. CERN); high academic inbreeding</td>
<td>0.49%</td>
<td>12,675 (NA)</td>
<td>Infrastructure building, competitive R&amp;D projects and individual fellowship program (doctoral and post-doctoral)</td>
</tr>
<tr>
<td>1996-2005</td>
<td>Doctoral and post-doctoral fellowship program, increased brain circulation</td>
<td>0.76%</td>
<td>29,761 (12,152)</td>
<td>Performance-based funding of research units, through national research assessments, including the creation of large Associate Laboratories, to foster research excellence through networks of academic research centers; research career status; promotion of scientific culture and the public understanding of science</td>
</tr>
<tr>
<td>2006-2010</td>
<td>Increasing capacity; research contracts (moving towards brain gain)</td>
<td>1.7%</td>
<td>75,073 (23,125)</td>
<td>Scientific employment though competitive research contracts; University Chairs; reform of university governance and assessment systems; international partnerships promoting thematic networks of research and advanced training; further promotion of the public understanding of science</td>
</tr>
</tbody>
</table>
Table 3 – Major science policy instruments used over time in Portugal (1970-2010), as identified in this paper

<table>
<thead>
<tr>
<th>Public policy instrument</th>
<th>Characterization/focus</th>
<th>Beginning</th>
<th>Ongoing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doctoral Fellowships</td>
<td>Centralized program oriented towards the advanced training of human resources, independently of university hierarchies</td>
<td>Late 1960s with JNICT; the number of fellowships and R&amp;D projects increased substantially after 1986, through EU funds and particularly, after 1996, through FCT</td>
<td>Yes</td>
</tr>
<tr>
<td>Competitive funding program for R&amp;D Projects</td>
<td>Promoting research activities and research teams at national and EU levels</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-doctoral fellowships</td>
<td>Promoting the internationalization and mobility of doctorates; fostering knowledge production and participation in international knowledge networks</td>
<td>Mid 1990s after the creation of FCT in 1996</td>
<td>Yes</td>
</tr>
<tr>
<td>Promotion of scientific culture</td>
<td>Science education in schools and the public understanding of science</td>
<td>Since 1996</td>
<td>Yes</td>
</tr>
<tr>
<td>Performance-based funding of research units, through national research assessments (every 3 to 4 years), with 3-year contracts</td>
<td>Promoting research capacity through institutional building, independently of university hierarchies. Facilitating the creation of independent research units and the concentration of doctorates in research units</td>
<td>Mid 1990s with the creation of FCT; first assessment only with Portuguese reviewers; all the others are international assessments</td>
<td>Yes</td>
</tr>
<tr>
<td>Associate Laboratories: performance-based funding of large research units and networks, based on national research assessments, with 10-year contracts</td>
<td>Association and networking of the better qualified R&amp;D units in the assessment exercise; focus on critical mass and renewal of the researcher base.</td>
<td>Since 1999</td>
<td>Yes</td>
</tr>
<tr>
<td>International partnerships with leading universities and research institutes</td>
<td>Thematic research and advanced training networks, facilitating the internationalization of academic staff; increased R&amp;D collaboration between Portuguese universities; Increasing the qualification of academic staff; academic exchange programs and pedagogical and scientific improvements</td>
<td>Since 2006</td>
<td>Yes</td>
</tr>
<tr>
<td>Post-doctoral research contracts program</td>
<td>Attraction of researchers nationally and internationally with a doctorate and some years of research experience; renewal of university academia</td>
<td>Since 2007</td>
<td>Yes</td>
</tr>
<tr>
<td>Sponsored Research Chairs</td>
<td>Attraction of foreign and Portuguese senior academics to Portuguese universities, co-sponsored by firms</td>
<td>Since 2007</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Table 4 – Impact of national research assessments (every 3 to 4 years) in terms of number of research units, number of PhDs and average PhDs per research unit in Portugal

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of research units</td>
<td>269</td>
<td>337</td>
<td>462</td>
<td>423</td>
</tr>
<tr>
<td>Total number of PhDs</td>
<td>3,673</td>
<td>5,850</td>
<td>8,038</td>
<td>11,426</td>
</tr>
<tr>
<td>Average number of PhDs per research unit</td>
<td>13.7</td>
<td>17.4</td>
<td>17.4</td>
<td>27.0</td>
</tr>
</tbody>
</table>

Source: Portuguese Science and Technology Foundation

Table 5 – Sample of measures adopted through major international partnerships in science, technology and higher education in Portugal (2006-2011)

<table>
<thead>
<tr>
<th>Program</th>
<th>Launched/signed</th>
<th>Human resources focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT-Portugal</td>
<td>October 2006</td>
<td>Doctoral education: joint doctoral programs between Portuguese universities with research spells spent at MIT in bioengineering systems, engineering design and advanced manufacturing, sustainable energy systems, and transportation systems; Master's programs and advanced studies programs in bioengineering systems, engineering design and advanced manufacturing, sustainable energy systems, and transportation systems; in collaboration with the Sloan School of Management, establishment of the Lisbon MBA with Portuguese business schools; Creation of an educational consortium involving Portuguese engineering schools, other Portuguese schools (mainly of economics and social sciences), Associated Laboratories, state laboratories, industrial research laboratories and MIT</td>
</tr>
<tr>
<td>Carnegie Mellon-Portugal</td>
<td>October 2006</td>
<td>Doctoral education: dual doctoral programs between CMU and Portuguese universities in Computer science, electrical and computer engineering, software engineering, engineering and public policy, language technology, human-computer interaction, applied mathematics and technological change and entrepreneurship; Faculty exchange programs bringing Portuguese academics to teach at CMU, and to learn new pedagogical and curricula perspectives; Dual master's degrees between CMU and Portuguese universities in entertainment technology, human-computer interaction, information technology/security, and software engineering; Post-doctoral fellowships in mathematics (mainly applied mathematics, probabilities and stochastic methods)</td>
</tr>
<tr>
<td>Univ. Texas at Austin-Portugal</td>
<td>March 2007</td>
<td>Doctoral education: dual doctoral programs with the University of Texas at Austin and Portuguese universities, in Digital Media and Mathematics. Post-doctoral fellowships in Mathematics and Digital Media</td>
</tr>
</tbody>
</table>
Table 6 – Evolution of the number of researchers (headcount) and funding per researcher (in 1000 US dollars PPP of 2008), in a sample of OECD countries for 2001 and 2008

<table>
<thead>
<tr>
<th>S&amp;T system</th>
<th>Total researchers (headcount)</th>
<th>Funding per researcher (1000 US$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech Republic</td>
<td>29,216</td>
<td>44,240</td>
</tr>
<tr>
<td>Denmark</td>
<td>29,791</td>
<td>48,442</td>
</tr>
<tr>
<td>Estonia</td>
<td>4,803</td>
<td>7,226</td>
</tr>
<tr>
<td>Finland</td>
<td>47,534</td>
<td>55,195</td>
</tr>
<tr>
<td>France</td>
<td>217,173</td>
<td>289,478</td>
</tr>
<tr>
<td>Hungary</td>
<td>28,351</td>
<td>33,739</td>
</tr>
<tr>
<td>Iceland</td>
<td>3,231</td>
<td>4,158</td>
</tr>
<tr>
<td>Italy</td>
<td>100,442</td>
<td>145,623</td>
</tr>
<tr>
<td>Japan</td>
<td>792,699</td>
<td>890,669</td>
</tr>
<tr>
<td>Korea</td>
<td>178,937</td>
<td>300,050</td>
</tr>
<tr>
<td>Norway</td>
<td>34,864</td>
<td>44,145</td>
</tr>
<tr>
<td>Poland</td>
<td>89,596</td>
<td>97,474</td>
</tr>
<tr>
<td>Portugal</td>
<td>31,146</td>
<td>75,073</td>
</tr>
<tr>
<td>Slovakia</td>
<td>19,923</td>
<td>19,814</td>
</tr>
<tr>
<td>Slovenia</td>
<td>6,740</td>
<td>10,124</td>
</tr>
<tr>
<td>Spain</td>
<td>140,407</td>
<td>217,716</td>
</tr>
<tr>
<td>Turkey</td>
<td>67,190</td>
<td>106,423</td>
</tr>
</tbody>
</table>

Source: OECD
LIST of FIGURES

Figure 1 – Doctoral degrees concluded in Portugal and abroad, 1970-2010, and percentage of those obtained at the four oldest universities in Portugal (Porto, Coimbra, Lisbon and the Technical University of Lisbon)

Figure 2 – Ratio of PhDs obtained abroad to PhDs obtained in Portugal, by five-year period, 1970-2010

Figure 3 – Percentage of academic staff holding a PhD awarded at their own university (“inbreeding level”) - the case of the Faculty of Engineering of Porto University, 1979-2010

Figure 4 – Academic staff holding a PhD as a percentage of all academic staff, at faculties of engineering, medicine and law in three different Portuguese universities, 2001-2009

Figure 5 – Gross expenditure on R&D (GERD) as a percentage of GDP, and its co-evolution with major public policies and funding instruments aimed at promoting the Portuguese science base

Figure 6 – National program for doctoral fellowships granted on a yearly basis, by mobility options, 1995-2008, funded through the Portuguese Science and Technology Foundation (FCT)

Figure 7 – National program for post-doctoral fellowships granted on a yearly basis, by mobility options, 1995-2008, funded through the Portuguese Science and Technology Foundation (FCT)

Figure 8 – National program for doctoral research contracts, publicly funded (2007-2009) through the Portuguese Science and Technology Foundation (FCT)

Figure 9 – Employment of doctorates in higher education over the last decade (2000s)
Figure 1 – Doctoral degrees concluded in Portugal and abroad, 1970-2010, and percentage of those obtained at the four oldest universities in Portugal (Porto, Coimbra, Lisbon and the Technical University of Lisbon).

Figure 2 – Ratio of PhDs obtained abroad to PhDs obtained in Portugal, by five-year period, 1970-2010.

Note: FCT direct PhD fellowships may involve periods in Portugal and abroad. The trend previous to the period 1994-1998 is simulated.

Source: GPEARI - MEC
Figure 3 – Percentage of academic staff holding a PhD awarded at their own university ("inbreeding" level) - the case of the Faculty of Engineering of Porto University, 1979-2010

Source: FEUP, University of Porto

Figure 4 – Academic staff holding a PhD as a percentage of all academic staff, at faculties of engineering, medicine and law in three different Portuguese universities, 2001-2009

Source: REBIDES, GPEARI - MEC
Figure 5 – Gross expenditure on R&D (GERD) as a percentage of GDP, and its co-evolution with major public policies and funding instruments aimed at promoting the Portuguese science base.

Source: GPEARI - MEC
Figure 6 – National program for doctoral fellowships granted on a yearly basis, by mobility options, 1995-2008, funded through the Portuguese Science and Technology Foundation (FCT)

Source: Portuguese Science and Technology Foundation (FCT)

Figure 7 – National program for post-doctoral fellowships granted on a yearly basis, by mobility options, 1995-2008, funded through the Portuguese Science and Technology Foundation (FCT)

Source: Portuguese Science and Technology Foundation (FCT)
Figure 8 – National program for doctoral research contracts, publicly funded (2007-2009) through the Portuguese Science and Technology Foundation (FCT)

Source: Portuguese Science and Technology Foundation (FCT)
Figure 9 – Employment of doctorates in higher education over the last decade (2000s)

Note: The total researchers with a PhD in the Higher Education sector include those performing research at IPFSLs;
Sources: REBIDES, GPEARI - MEC; IPCTN, GPEARI - MEC
Chapter 12:
Miguel Amaral, Rui Baptista and Francisco Lima, “Serial entrepreneurship: impact of human capital on time to re-entry”.
Serial entrepreneurship: impact of human capital on time to re-entry

A. Miguel Amaral · Rui Baptista · Francisco Lima

Accepted: 31 August 2009 / Published online: 24 September 2009 © Springer Science+Business Media, LLC. 2009

Abstract This study uses a longitudinal matched employer–employee database to examine how ex-entrepreneurs’ levels of general and specific human capital influence their likelihood of re-entering entrepreneurship over time, in a different firm, thereby becoming serial entrepreneurs. The results reveal a negative effect of general human capital on the hazard of becoming a serial entrepreneur; the impact of entrepreneurial-specific human capital on the hazard of re-entering entrepreneurship is in general positive. This research provides a dynamic approach to serial entrepreneurship revealing that specific types of human capital play distinct roles on individuals’ entrepreneurial behavior.

Keywords Serial entrepreneurship · General and specific human capital · Longitudinal data · Piecewise constant proportional hazard

JEL Classifications M13 · J24 · L26 · C41

1 Introduction

Entrepreneurship is not confined to new business creation (Cooper and Dunkelberg 1986), nor is it a single-action event (Birley and Westhead 1993). Individuals may become business owners by acquiring or inheriting an existing business. Additionally, entrepreneurs may not limit themselves to one firm, instead choosing to run several of them. This broader perspective emphasizes the heterogeneity of entrepreneurship and highlights the importance of studying both firms and individual entrepreneurs. The need to better understand the behavior of individual entrepreneurs in a variety of settings that extend beyond one-time entrepreneurial experience is highlighted by, among others, MacMillan (1986), Westhead and Wright (1998), Carter and Ram (2003), Westhead et al. (2005) and Stam et al. (2008). Entrepreneurs involved in more than one venture, or habitual entrepreneurs, are gaining the attention of researchers.1

1 Habitual entrepreneurs are defined as individuals who have established, inherited and/or purchased more than one business, as opposed to novice entrepreneurs, who have established, inherited and/or purchased only one business. Habitual entrepreneurs include individuals who, after owning one venture in a specific moment, start, acquire or inherit another business in a subsequent moment, i.e., serial entrepreneurs, and individuals who own several businesses simultaneously, i.e., portfolio entrepreneurs (Birley and Westhead 1993; Westhead and Wright 1998).
The present study concentrates on the subset of habitual entrepreneurs who exit their first business and subsequently start or acquire a second one, thereby becoming serial entrepreneurs. For the purpose of this research, “entrepreneurs” includes all individuals who report themselves as business owners, with full or partial ownership (Parker 2004; Van Stel 2005) and who have started or acquired a business (Westhead and Wright 1998). A growing, but still narrow, stream of work explores the differences in characteristics between novice and serial entrepreneurs (Kalleberg and Leicht 1991; Westhead et al. 2003, 2005). Serial entrepreneurship is said to be important for its potential in fostering the process of wealth creation (Scott and Rosa 1996; Westhead et al. 2005), job generation (Flores-Romero 2006; Westhead et al. 2005) and economic performance (Westhead et al. 2003, 2005; Colombo and Grilli 2005). Several works in the entrepreneurship literature recognize that serial entrepreneurship signals high levels of general and entrepreneurial-specific human capital (e.g., Ucbasaran et al. 2003, 2008; Westhead et al. 2005) and that serial entrepreneurs learn from earlier entrepreneurial experiences, thereby increasing their initial endowment of skills (Stam et al. 2008; Ucbasaran et al. 2008). Therefore, serial entrepreneurs are expected to have better managerial and technical skills; better networks of contacts and access to market-specific information; and, consequently, be better equipped to identify and take advantage of new business opportunities (McGrath and MacMillan 2000; Ucbasaran et al. 2008).

Research on serial entrepreneurship is relevant for advancing the understanding of entrepreneurs’ learning patterns over time, as they go through multiple entrepreneurial experiences. So far, to our knowledge, there are no studies focusing on the effect of general and entrepreneurial-specific human capital on time to re-entry. This paper focuses specifically on the transition between the first and second entrepreneurial experiences. Given the inherently dynamic nature of serial entrepreneurship, we believe this makes a significant contribution to the literature on this subject. In particular, little attention is paid by previous research to the fact that this transition does not necessarily occur immediately upon exiting a first business. Indeed, there may be a prolonged interval between entrepreneurial exit and subsequent re-entry, which is of significance for a better understanding about entrepreneurial dynamics and for the design of public policies targeted to the specific types and needs of ex-entrepreneurs (Westhead et al. 2005).

The absence of a dynamic approach to serial entrepreneurship in the literature is possibly due to the fact that records of entrepreneurs and their businesses are mostly drawn from cross-sectional data, and, therefore, researchers usually look at the businesses currently owned by entrepreneurs. Rosa and Scott (1999) find that most empirical studies only observe one of the ventures of the habitual entrepreneur. The present paper addresses this gap in the literature by using appropriate data to explore different behaviors among former entrepreneurs and ascertain whether their human capital plays a role on the time to re-enter entrepreneurship.

This study uses longitudinal matched employer–employee data covering the period 1986–2003 to study entrepreneurs’ time to re-entry decision following exit from their previous business ownership experience. The analysis focuses primarily on ex-entrepreneurs’ general human capital (years of formal education and employment experience) and specific human capital (previous entrepreneurial, managerial and founding experience), while controlling for the individuals’ demographics, characteristics of their first business, industry and macroeconomic context. Results reveal different impacts of specific and general human capital on the decision to anticipate or delay re-entry into entrepreneurship.

The paper proceeds as follows. The next section discusses the literature on human capital and serial entrepreneurship, defining the research goals and propositions of the present study, and clarifying its contribution. Section 3 presents the data and provides a detailed discussion of relevant issues in data construction associated with entrepreneurial careers and serial entrepreneurship. Section 4 presents the empirical model used, as well as the variables influencing time to serial entrepreneurship. Section 5 displays and discusses the results from model estimates. Section 6 presents some concluding remarks and addresses limitations of the study; suggestions for future research avenues to be pursued are also presented.

2 Theory and hypotheses

The analysis is theoretically grounded in human capital theory. Human capital models propose a positive
association between formal and on-the-job training (i.e., education and professional experience) and the performance (i.e., productivity) of individual workers (Mincer 1974; Becker 1975). The human capital perspective can be applied to entrepreneurship since individuals’ formal education and labor market, managerial and entrepreneurial experience have a significant effect on their choice of entering and exiting entrepreneurship, as well as on their performance as entrepreneurs (see Parker 2004, for a review). We investigate how individuals’ human capital impacts on time to re-enter entrepreneurship. In particular, and in line with Brüderl et al. (1992) and Becker (1993), a distinction is made between general and specific forms of human capital. Ucbasaran et al. (2008) distinguish between general and specific forms of human capital in the case of serial entrepreneurship.

For this study, general human capital comprises the formal education and paid-employment experience of the entrepreneur, which may lead to skills that are useful across different occupations and economic settings. Specific human capital entails prior experiences that are more relevant for an entrepreneur to re-enter entrepreneurship in a subsequent firm. Having had previous experience in entrepreneurship and possessing senior management skills endow individuals with specific knowledge of business dynamics (e.g., knowledge about customers, suppliers, products and services), as well as business opportunities identification and the process of entering and running a firm (Bates 1990; Gimeno et al. 1997; Bosma et al. 2004; Ucbasaran et al. 2008).

Human capital theory poses that those individuals whose human capital is more specific to the venture will be less mobile across organizations (Becker 1975). Accordingly, those whose human capital is more specific to entrepreneurial activities would be less mobile from entrepreneurship and face higher opportunity costs of becoming wage workers. Hence, even if individuals with higher entrepreneurial-specific human capital decide to leave their firms, they will want to continue in the same occupation in the labor market, thereby becoming serial entrepreneurs. In order to disentangle the role played by individuals’ backgrounds and experiences on the time to become a serial entrepreneur, this study focuses on differences among general and specific forms of human capital.

2.1 General human capital

Previous studies provide mixed evidence regarding human capital and the likelihood of becoming an entrepreneur. Some authors argue that better educated individuals are endowed with better skills and abilities, and thus may have a higher probability of choosing entrepreneurship than the less educated (Lucas 1978; Carrasco 1999). Additionally, better educated individuals tend to be better informed, implying that they are more efficient at assessing new entrepreneurial opportunities (Rees and Shah 1986). In contrast, various authors find negative or non-linear effects on human capital and entrepreneurship (Blanchflower 2004; Minniti and Bygrave 2003; Koellinger et al. 2007). Some studies show—for the case in which human capital is specifically measured through tertiary education—that individuals with higher educational levels have less chance of being entrepreneurs than people holding primary and secondary education (Amaral and Baptista 2007; Livanos 2009).

One common explanation is that higher levels of general human capital can facilitate transitions into wage employment, thus reducing the likelihood of entrepreneurship (Evans and Leighton 1989; De Wit and Van Winden 1989). This view is substantiated by Lazear (2005), who argues that entrepreneurs should be endowed with a human capital that is more varied than formal education (i.e., it should comprehend other knowledge and forms of experience that are more applied and specific to entrepreneurship), while those who are paid employees should be specialists (i.e., with a more specialized educational background).

Literature specific to serial entrepreneurship suggests that individuals with multiple entrepreneurial experiences are, in general, more educated than novice entrepreneurs (Donckels et al. 1987; Kolvereid and
However, this literature focuses on individuals’ current occupational status instead of on their behavior and transition from an initial to a subsequent entrepreneurial experience. Therefore, the extant literature is not explicit about the role played by education on time to re-enter entrepreneurship.

The study of time to serial entrepreneurship focuses on individuals’ occupational paths after leaving their first entrepreneurial experience. According to Gimeno et al. (1997), those with more education have more personal opportunities available after exiting their current venture. In fact, individuals with entrepreneurial experience and higher education are better equipped to choose an alternative occupation in the labor market rather than directly starting or acquiring another business. In line with the human capital theoretical framework, ex-entrepreneurs with more education have greater occupational mobility and flexibility to opt for a different and eventually more attractive immediate occupation in the labor market. Therefore, we propose that being endowed with higher levels of education will lower the hazard of re-entry into entrepreneurship.

**Hypothesis 1** Ex-entrepreneurs with more education are more likely to delay re-entry into entrepreneurship.

The same reasoning applies to individuals’ previous experience in employment, which is usually highly valued by employers (Mincer 1974). Hence, ex-entrepreneurs with high levels of previous employment experience may be more attracted by a new opportunity as wage employees than by the prospect of a new entrepreneurial experience. Hyytinen and Ilmakunnas (2007) find that tenure in paid-employment is insignificant for individual’s aspirations and transitions to serial entrepreneurship; however, its negative effect suggests that prior employment experience hinders the decision to re-enter entrepreneurship. Some of these ex-entrepreneurs may be willing to re-enter entrepreneurship, but—given their high labor mobility—choose to spend some time working for an employer while planning and preparing their resources for engaging on a subsequent entrepreneurial experience. We can then formulate the following hypothesis:

**Hypothesis 2** Ex-entrepreneurs with more experience in paid employment are more likely to delay re-entry into entrepreneurship.

### 2.2 Entrepreneurial-specific human capital

According to Starr and Bygrave (1992), the skills and knowledge relevant to managing and operating a business are experiential by nature. One explanation for the importance of this kind of specific human capital is that it influences how individuals seek information (Cooper et al. 1995) and create or identify entrepreneurial opportunities (Shane 2000, 2003). The positive effect of entrepreneurial experience on entrepreneurial intentions and transitions (Kolvereid and Isaksen 2006) extends to the particular case of serial entrepreneurs (Hyytinen and Ilmakunnas 2007). Serial entrepreneurs, who are endowed with entrepreneurial-specific human capital built from their earlier experiences as business owners, feel better prepared to detect and pursue opportunities (Ucbasaran et al. 2003; Westhead et al. 2005). Therefore, we derive the next hypothesis:

**Hypothesis 3** Ex-entrepreneurs with more experience in business ownership are more likely to re-enter entrepreneurship more quickly.

In the same vein, individuals with managerial experience are more likely to have developed the necessary skills to pursue market opportunities and organize new businesses (Bates 1990; Gimeno et al. 1997; Eckhardt and Shane 2003). In general, studies analyzing the differences between serial and novice entrepreneurs with regard to their backgrounds sustain that experience of managing a previous business is important for the management of subsequent ones (Ucbasaran et al. 2006; Westhead et al. 2003) and that it has a positive effect on the probability of re-entering entrepreneurship (Stam et al. 2008). For example, Westhead et al. (2005) find that a larger proportion of serial, rather than novice entrepreneurs, had been self-employed or in a managerial position prior to gaining an equity stake in their current businesses.

---

3 An exception is the study by Westhead et al. (2005), which found no statistically significant differences between novice and serial entrepreneur types with regard to their educational level.
Hypothesis 4 Ex-entrepreneurs with managerial experience are more likely to re-enter entrepreneur-
ship more quickly.

While it is widely stated in the literature that individuals’ past entrepreneurial experience is impor-
tant in shaping their future entrepreneurial careers, studies usually fail to specify the entry mode(s) associated with that experience (Kolvereid and Isaksen 2006). Westhead and Wright (1998) discuss a conceptual framework for categorizing novice and habitual entrepreneurs (owners–managers) that incorporates the possibility of either firm founding or acquisition. This distinction is important because business founders and non-founders may have different behavioral characteristics (Busenitz and Barney 1997). However, the majority of literature in serial entrepreneurship tends to focus their empirical analysis on firm founders rather than firm acquirers (Kolvereid and Bullvag 1993; Birley and Westhead 1993; Westhead and Wright 1998).

One exception, provided by Ucbasaran et al. (2003), examines both habitual starters (i.e., those who have established more than one business) and habitual acquirers (i.e., those who have acquired more than one business). The authors find that individuals experienced in starting a firm are more likely to found a second one. Conversely, Westhead et al. (2005) argue that serial entrepreneurs who have founded a firm are more likely to acquire a subsequent business because they find the start-up process daunting and are less likely to report that they enjoy the early stages of starting a venture. Hence, evidence on entry modes and serial entrepreneurship suggests that founding experience can be considered a significant form of entrepreneurial-specific human capital and that it has a positive impact upon entrepreneurial re-entry, both through the start-up or acquirement of a subsequent firm. However, this is insufficient to predict the impact of founding experience on time to re-enter entrepreneurship. Founding experience may not be, per se, a positive outcome, as it brings not only assets but also liabilities (Starr and Bygrave 1992). Ucbasaran et al. (2003) propose that the extent to which individuals’ would retard or speed up their re-entrance into entrepreneurship is influ-
cenced by the success (asset) or failure (liability) associated with their previous entrepreneurial experience.

In order to disentangle the positive and negative effects of founding experience on the time to become a serial entrepreneur, founding experience is deter-
mained by the mode of exit from the first firm. Serial entrepreneurs who exit by selling out their initial venture may have generated sufficient funds to use personal resources for financing the subsequent venture (Westhead and Wright 1998). Moreover, they are likely to have a good reputation with financiers, customers and suppliers, and other stake-
holders (Ucbasaran et al. 2003). Therefore, it is proposed that founders who sell out their businesses will evaluate more positively their founding experience and, as a result, will hasten the decision to re-enter entrepreneurship, as opposed to founders who have failed and dissolved their businesses. For the present research, we term this form of entrepreneur-
ial-specific human capital, “positive founding experience,” i.e., founders who have sold their business, vis-à-vis “negative founding experience,” i.e., founders who have dissolved their business. Conse-
quently, we put forward the next hypothesis to be tested:

Hypothesis 5 Ex-entrepreneurs with “positive founding experience” are more likely to re-enter entrepreneurship more quickly.

3 Data description

The present investigation benefits from an extensive data set comprised of individuals’ backgrounds, career paths, and flows between firms and industries, originating from a longitudinal matched employer–employee database, the Quadros de Pessoal.4 The database is built from legally mandatory surveys submitted by firms with at least one employee to the Portuguese Ministry of Employment and Social Security and accounts for nearly the entire population of entrepreneurs within the country. Firms and individuals (workers and business owners) are iden-
tified and matched through a unique identification number, so they can be followed over time. Yearly data on business owners and paid employees

4 The Quadros de Pessoal database has been used in a diverse set of empirical works in labor and industry dynamics (e.g., Mata and Portugal 2002; Cabral and Mata 2003 or Varejão and Portugal 2007).
includes: gender, age, hierarchical qualification, tenure, formal education and skill levels.

For each firm, yearly data are available on ownership composition, size (employment), age and industry. The study considers a set of predictors for the choice of entering or re-entering entrepreneurship that, according to the literature, are found to be consistent and important in explaining the entrepreneurial process (see Parker 2004; Shane 2004; Van Praag 2005; Westhead et al. 2005; Ucbasaran et al. 2006, 2008). These variables can be grouped into four dimensions: human capital, demographics, firm and industry characteristics, and macroeconomic context.

The longitudinal and often all-inclusive nature of large surveys, such as the one used in the present study, can be used to answer research questions where interrelated heterogeneous factors concerning firms and individuals require large, unbiased samples with the possibility to simultaneously investigate a variety of factors, thus making the use of such data particularly appropriate for the research issue being studied here.

3.1 Choice of labor market status

Our data set covers the period 1986–2003 and accounts for male and female non-agricultural workers, aged 16 or more, who have exited from a first entrepreneurial experience. We start by categorizing individuals according to their professional status at moments in time: business owner, paid employee or non-employed.

Exit from business ownership experience from a given firm (exit stage) occurs at time $t$. The decision-making process modeled here occurs after the exit stage, meaning exit from the first experience as an entrepreneur. At time $t + n$ ($n \geq 1$), individuals decide whether and when to enter (or not) a subsequent experience as entrepreneur.

3.2 Issues in data construction

The initial data set for the present study is comprised of 176,747 entrepreneurs who are observed to have exited their firms and for whom we have complete information on all the variables under analysis. While 19,074 individuals re-enter entrepreneurship in a subsequent firm, thus becoming serial entrepreneurs, the remaining 157,673 do not. In an attempt to control for potential biases in our analysis, the empirical investigation will focus primarily on a subsample of our initial data set, which is comprised of those ex-entrepreneurs who were 30 or younger when exiting their initial firm.

In fact, it is expected that individuals with more human capital are simultaneously older and have an edge with respect to resource acquisition and opportunity exploitation over younger entrepreneurs. Moreover, for older entrepreneurs, the opportunity cost of time increases with age because of a shorter remaining lifespan (Becker 1965). Therefore, we perform estimates using a sub-sample of ex-entrepreneurs who left their initial business at the age of 30 or younger. The rationale for this methodological step is that, on the one hand, by controlling for the initial conditions, we are giving ex-entrepreneurs sufficient time to become a serial entrepreneur and, simultaneously, reducing potential left censoring issues regarding individuals’ eventual entrepreneurial experiences prior to 1986. On the other hand, we are disentangling professional experience from age, thus decreasing the prevalence of unobserved heterogeneity in our analysis.

After filtering, the final data set consists of 23,172 observations, including 21,076 cases at risk of

---

5 The data set registers a gap in the years 1990 and 2001, for which there is no information available on employees. We control for this fact in the estimation procedure.

6 “Non-employed” people are classified as those who are disengaged from any firm (i.e., exit the database), either because they are inactive, unemployed, employed in private non-agriculture firms or because they exited the private job market. The non-employed also include people who have not yet entered the job market (i.e., became engaged with a firm for the first time) at a certain moment, but will do so at a subsequent time.

7 This study focuses on time to serial entrepreneurship rather than differences between novice and serial entrepreneurs. However, by choosing to use 30 years old as the cutoff point, it is assumed with a high degree of certainty that individuals are leaving their first entrepreneurial experience, thereby being novice entrepreneurs. According to the descriptive statistics, on average, individuals re-enter entrepreneurship at 40 years old. Therefore, the specific choice of 30 years old as the cutoff point allows for individuals aged 30 or less to be at risk of re-entering entrepreneurship for a reasonably long period of 10 years on average.
re-entering entrepreneurship and 2,096 actual re-entrants. As the present study investigates time to re-enter into entrepreneurship after individuals terminate their first entrepreneurial experience, there are a number of groups excluded from the study. This includes: individuals who have never entered entrepreneurship, individuals who have never left their business and, finally, portfolio entrepreneurs.\(^8\)

Another important issue concerns the modes of entry and exit from entrepreneurship. The mode of entry into entrepreneurship is determined by differentiating between firm founders (or starters) and firm acquirers. If an individual enters a firm for the first time and that same firm is new in the market, (i.e., firm start-up and transition to business ownership happen simultaneously)\(^9\) it is assumed that entry into entrepreneurship occurs through founding; otherwise, we assume that a pre-existing firm has been acquired by the business owner. The mode of exit is examined in a similar fashion. It is more likely that simultaneous entrepreneurial exit and business dissolution associate with business failure (i.e., a negative performance),\(^10\) while entrepreneurial exit from a firm that continues operating in the market after such exit is more likely to result from sale of the entrepreneur’s share of the business to a third person and is therefore less likely to result from business failure (Gimeno et al. 1997; Heald 2003). Hence, we build a binary variable distinguishing firm dissolution (i.e., the period when the business owner exit coincides with the firm extinction) from sell-out (i.e., the business owner exits from a firm that continues operating in the market).

The data are limited with regard to the mode of exit, more specifically the absence of information distinguishing mergers from true dissolutions of firms. We estimate a proxy for merger accounting for dissolutions by looking at the extent to which a sizeable part of the workforce of each firm moves to a different one. We reach a similar conclusion to Mata and Portugal (2002) that, within the Portuguese private sector, <1% of the total dissolutions are due to merger. This suggests that the inability to track mergers is unlikely to impact the results significantly.

3.3 Descriptive statistics for all ex-entrepreneurs

Table 1 presents variable definitions and descriptive statistics for all ex-entrepreneurs. In the complete data set 8.5% of the total observations consist of individuals re-entering entrepreneurship, the event of interest. The sample is composed mostly of males (72%). The proportions for each gender are similar for those who re-enter and those who do not. While the average age at exit is 43, individuals who become serial entrepreneurs exit their initial entrepreneurial experience at a younger than average age, at 40.

While firm age averages 13 years, serial entrepreneurs come from younger firms, aged about 11 years. Serial entrepreneurs also come from smaller firms (8.5 employees vs. an overall average of 12.8). On average, 23% of serial entrepreneurs founded their first firm, while this proportion is 20% for non-serial entrepreneurs. About 63% of re-entrants have dissolved their first business, while the overall proportion of dissolutions is only 49%.

3.4 Descriptive statistics for the sub-sample of ex-entrepreneurs aged up to 30

The proportion of individuals re-entering entrepreneurship (about 9%) in the sub-sample of ex-entrepreneurs aged up to 30 is similar to that observed for the whole sample (about 8%). This means that the

---

\(^8\) As previously noted, portfolio entrepreneurs are those who remain as business owners in one firm while simultaneously starting or acquiring other firms. Although portfolio entrepreneurship plays a vital role in the economy, this study focuses solely on serial entrepreneurs. This is a methodological choice as this study focuses on individuals who exit one business to engage (or not) in a second one and the time (number of years) mediating those two events, which does not fit the portfolio definition.

\(^9\) Foundation and transition into business ownership are considered simultaneous if the year of firm foundation equals the year the individual becomes a business owner or if there is a difference of 1 year between occurrences, in order to account for possible asymmetries in data collection.

\(^10\) In fact, as shown in Tables 1 and 2, descriptive statistics reveal that a vast majority of start-ups were dissolved rather than sold out. The data do not provide information on the specific reasons leading to firm closure. According to the correlations table (Appendix Table 4), firm dissolution is highly correlated with unemployment rate, suggesting that, to a large extent, real failures are captured. Nevertheless, it is possible that a business failure may also be understood as failure to equal or exceed a performance threshold required by the entrepreneur to keep the business running, while not necessarily indicating failure to be economically viable. It is therefore possible that a business deemed to be failing by an entrepreneur will be acquired by another entrepreneur with a lower performance threshold (see Gimeno et al. 1997)
sub-sample has similar characteristics to the overall sample with regard to the incidence of serial entrepreneurs.

When limiting the analysis to entrepreneurs who have left their firms between the ages of 16 and 30 years old, inclusive, we find similar characteristics to those of the overall sample in terms of gender, top management experience and sector composition, as shown in Table 2. As this group focuses on younger entrepreneurs, variables accounting for firm age, number of years in paid-employment, non-employment and business ownership are, in general, lower.

Education, the proportion of founders (versus acquirers) and firm size are greater for younger ex-entrepreneurs. The subsample composed of younger ex-entrepreneurs leave their first firms with, on average, 8.5 years of education, compared with 7 years for all individuals. Among ex-entrepreneurs who left their preceding business before turning 31 years, exit occurs at an average age of 25 for those individuals not re-entering and 26 for those re-entering. While firm dissolution is more frequent among younger entrepreneurs than for the overall sample (55 vs. 49%), both display a higher rate of dissolutions among serial than non-serial entrepreneurs. Apart from the variable accounting for number of years in business ownership (which is slightly higher for serial entrepreneurs), differences between re-entrants and non-re-entrants are, in general, similar or proportional to those observed in the overall sample (see Table 1).

### 4 Empirical model and variables

#### 4.1 Dependent variable

The primary dependent variable is the time from exiting one entrepreneurial experience to the start of the next entrepreneurial experience. A spell starts when an individual exits entrepreneurship, and the duration of that spell corresponds to the time elapsed until the moment the individual re-enters entrepreneurship. Single-spell duration data are obtained by flow sampling: the beginning of a spell out of

<table>
<thead>
<tr>
<th>Variable</th>
<th>All observations</th>
<th>Do not re-enter</th>
<th>Re-enter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Female (dummy)</td>
<td>0.28</td>
<td>0.45</td>
<td>0.29</td>
</tr>
<tr>
<td>Age</td>
<td>43.32</td>
<td>12.57</td>
<td>43.55</td>
</tr>
<tr>
<td>Years in non-employment</td>
<td>0.64</td>
<td>1.76</td>
<td>0.64</td>
</tr>
<tr>
<td>Years as a paid-employee</td>
<td>2.30</td>
<td>4.76</td>
<td>2.27</td>
</tr>
<tr>
<td>Years as a business owner</td>
<td>6.81</td>
<td>7.41</td>
<td>6.89</td>
</tr>
<tr>
<td>Years of education</td>
<td>7.03</td>
<td>3.78</td>
<td>7.03</td>
</tr>
<tr>
<td>Top management (dummy)</td>
<td>0.79</td>
<td>0.40</td>
<td>0.79</td>
</tr>
<tr>
<td>Start-up (dummy)</td>
<td>0.20</td>
<td>0.40</td>
<td>0.20</td>
</tr>
<tr>
<td>Start-up × dissolve</td>
<td>0.13</td>
<td>0.34</td>
<td>0.13</td>
</tr>
<tr>
<td>Start-up × sell-out</td>
<td>0.06</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>Firm size (no. of workers)</td>
<td>12.83</td>
<td>144.91</td>
<td>13.23</td>
</tr>
<tr>
<td>Total no. of partners (owners)</td>
<td>1.70</td>
<td>1.42</td>
<td>1.72</td>
</tr>
<tr>
<td>Firm dissolution (dummy)</td>
<td>0.49</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Primary sector (dummy)</td>
<td>0.02</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>Manufacturing (dummy)</td>
<td>0.20</td>
<td>0.40</td>
<td>0.20</td>
</tr>
<tr>
<td>Energy and construction (dummy)</td>
<td>0.12</td>
<td>0.33</td>
<td>0.12</td>
</tr>
<tr>
<td>Services (dummy)</td>
<td>0.60</td>
<td>0.49</td>
<td>0.60</td>
</tr>
<tr>
<td>Unemployment variation rate</td>
<td>0.01</td>
<td>0.20</td>
<td>0.01</td>
</tr>
<tr>
<td>No. of observations</td>
<td>224,805</td>
<td></td>
<td>205,731</td>
</tr>
</tbody>
</table>
Table 2 Descriptive statistics for ex-entrepreneurs aged up to 30 years old

<table>
<thead>
<tr>
<th>Variable</th>
<th>All observations</th>
<th>Do not re-enter</th>
<th>Re-enter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>Female (dummy)</td>
<td>0.33</td>
<td>0.47</td>
<td>0.35</td>
</tr>
<tr>
<td>Age</td>
<td>24.75</td>
<td>2.78</td>
<td>24.67</td>
</tr>
<tr>
<td>Years in non-employment</td>
<td>0.59</td>
<td>1.34</td>
<td>0.59</td>
</tr>
<tr>
<td>Years as a paid-employee</td>
<td>1.85</td>
<td>2.55</td>
<td>1.85</td>
</tr>
<tr>
<td>Years as a business owner</td>
<td>2.53</td>
<td>2.15</td>
<td>2.49</td>
</tr>
<tr>
<td>Years of education</td>
<td>8.50</td>
<td>3.77</td>
<td>8.51</td>
</tr>
<tr>
<td>Top management (dummy)</td>
<td>0.72</td>
<td>0.45</td>
<td>0.71</td>
</tr>
<tr>
<td>Start-up (dummy)</td>
<td>0.34</td>
<td>0.47</td>
<td>0.34</td>
</tr>
<tr>
<td>Start-up × dissolve</td>
<td>0.24</td>
<td>0.43</td>
<td>0.24</td>
</tr>
<tr>
<td>Start-up × sell-out</td>
<td>0.10</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>Firm age</td>
<td>8.01</td>
<td>12.52</td>
<td>8.17</td>
</tr>
<tr>
<td>Firm size (no. of workers in the firm)</td>
<td>20.80</td>
<td>293.91</td>
<td>21.95</td>
</tr>
<tr>
<td>Total no. of partners (owners) in the firm</td>
<td>1.85</td>
<td>1.78</td>
<td>1.87</td>
</tr>
<tr>
<td>Firm dissolution (dummy)</td>
<td>0.55</td>
<td>0.50</td>
<td>0.54</td>
</tr>
<tr>
<td>Primary sector (dummy)</td>
<td>0.02</td>
<td>0.14</td>
<td>0.02</td>
</tr>
<tr>
<td>Manufacturing (dummy)</td>
<td>0.21</td>
<td>0.41</td>
<td>0.22</td>
</tr>
<tr>
<td>Energy and construction (dummy)</td>
<td>0.14</td>
<td>0.34</td>
<td>0.13</td>
</tr>
<tr>
<td>Services (dummy)</td>
<td>0.58</td>
<td>0.49</td>
<td>0.58</td>
</tr>
<tr>
<td>Unemployment variation rate</td>
<td>0.03</td>
<td>0.21</td>
<td>0.03</td>
</tr>
<tr>
<td>No. of observations</td>
<td>23,172</td>
<td></td>
<td>21,076</td>
</tr>
</tbody>
</table>

entrepreneurship is observed for all individuals in the sample—there is no left censoring. While for a group of individuals we observe the periods until transition back to entrepreneurship—completed spells—for another group of individuals there is right censoring as they remain out of entrepreneurship longer than is observable in the data.

A piecewise-constant exponential hazard rate model (Jenkins 2005; Van Den Berg 2001) is used to examine how human capital variables influence time to re-entry into entrepreneurship. In the present framework, the regressors are assumed to be time-invariant. In the hazard model, $T$ is a continuous random variable denoting the number of years an individual remains between two entrepreneurial experiences. The hazard function gives the instantaneous probability of re-entering entrepreneurship at $t + n (n \geq 1)$, given that the individual has exited an experience in entrepreneurship ($t = 0$) and stayed in paid-employment or non-employment occupational status until $t$,

$$
\theta(t) = \lim_{{\Delta t \to 0}} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},
$$

where the probability density function of $T$ is $f(t)$, $F(t)$ is its cumulative distribution function, and $S(t)$ denotes the survival function.

Rather than specifying the shape of the hazard function a priori, it is fitted from the data using a piecewise constant proportional hazard function (PCH). The PCH departs from the proportional hazards model given by $\theta(t, X_t) = \theta_0(t) \exp(\beta X_t)$. In the particular case of the PCH, we have,

$$
\theta(t) = \begin{cases} 
\bar{\theta}_1 \exp(\beta X_1) & \text{if } 0 \leq t < \tau_1 \\
\bar{\theta}_2 \exp(\beta X_2) & \text{if } \tau_1 \leq t < \tau_2 \\
\vdots & \\
\bar{\theta}_K \exp(\beta X_K) & \text{if } \tau_{K-1} \leq t < \tau_K 
\end{cases}
$$

where the time axis is partitioned into $K$ intervals denoted by $\tau_1, \tau_2, \ldots, \tau_{K-1}$. The baseline hazard rate($\bar{\theta}_k$) is constant within each of the $K$ intervals,
but can change between them. As the year is the time unit, in specifying the baseline hazard function, four interval specific intercept terms are defined in the overall hazard, with binary variables distinguishing one initial interval of 1-year length, two subsequent intervals of 3 years and one final interval of 10 years duration.\textsuperscript{11}

4.2 Independent variables

As discussed in Sect. 2.1, general human capital is empirically approached using education and paid-employment experience as independent variables. Education is measured using the years of education reported by individuals in the Quadros de Pessoal questionnaire. Similarly, paid-employee experience is measured by the number of years in the occupation of paid-employee, observed in our longitudinal data.\textsuperscript{12}

Entrepreneurial-specific human capital entails three dimensions: experience in entrepreneurship, managerial attainment and founding experience. Entrepreneurial experience is measured by the number of years as business owner. Previous managerial attainment and firm founding experience (vis-à-vis firm acquisition) are captured using binary variables.

The time spent in non-employment, measured in years, is included as a variable associated with individual human capital. Some literature suggests that individuals who experience unemployment prior to entering self-employment suffered a deterioration of their human capital and therefore should have relatively lower survival rates (Evans and Leighton 1989; Carrasco 1999; Georgellis et al. 2007). However, there is, to the authors’ knowledge, no literature distinguishing the general and entrepreneurial-specific human capital attributes of non-employment, or its impact upon serial entrepreneurship. Moreover, as the data do not allow observing the reasons explaining individuals’ non-employment, this variable is used as a control with no further assumptions about the expected effect.

\textsuperscript{11} Percentiles distribution is used in order to fit a better baseline hazard in order to put 25% of the data in each step (Cleeves et al. 2002).

\textsuperscript{12} Since information on individuals’ date of hiring for each firm is included in the data, it is possible to track individuals’ experience prior to 1986.

4.3 Control variables

4.3.1 Entrepreneurs’ demographics

There is a fairly small amount of theoretical and empirical work predicting the effects of personal characteristics on the decision to become a serial entrepreneur. Moreover, the more abundant literature on the occupational choice of becoming an entrepreneur finds somewhat ambiguous results with regard to some variables, thus making difficult any attempt to bridge the literatures on occupational choice and serial entrepreneurship.

One such example is age. The literature suggests several effects associated with age on the decision to become an entrepreneur. Some authors find that the transition into self-employment is positively correlated with age (for example, Van Praag and Van Ophem 1995). The reason for this is that older people have had more time to build better networks and to identify valuable opportunities (Calvo and Wellisz 1980), and are more likely to have accumulated capital that can be used to establish a business (Blanchflower and Oswald 1998). Another research stream suggests that self-employment is concentrated among young individuals because older people are more risk averse (Miller 1984) and because older individuals are inclined to embark upon the more demanding work required by self-employment (Rees and Shah 1986). The literature is, nonetheless, consistent in showing that serial founders start their first business at a younger age than other novice founders (Birley and Westhead 1993; Kolvereid and Bullvag 1993; Westhead and Wright 1998).

In general, women have a lower likelihood of becoming an entrepreneur than males (Wagner 2005). Studies of serial entrepreneurship find very few women becoming engaged in a second entrepreneurial experience (Kolvereid and Bullvag 1993; Westhead and Wright 1998; Westhead et al. 2005). It is therefore expected that being female impacts negatively on the hazard of becoming a serial entrepreneur.

4.3.2 Organizational and industry characteristics of the first business owned

Firm age is frequently included in empirical analyses as an indicator of the firm life cycle (Mitchell 1994; Holmes and Schmitz 1996). Westhead et al. (2005)
find that serial entrepreneurs’ firms are significantly younger than novice entrepreneurs’ firms. Hence, logarithm of firm age at the moment of exit is included in our empirical analysis. Firm size is used as a predictor for serial entrepreneurship (Westhead and Wright 1998; Westhead et al. 2005); therefore, this analysis includes the firm size log, as measured by number of workers in the firm. The logarithm of number of partners (owners) in the firm is included as a control for size of entrepreneurial team, since some studies regard each member of the founding team as having human capital that must be taken into account (Helfat and Peteraf 2003).

The mode of exit (i.e., whether the business was dissolved or sold as the entrepreneur exited) is deemed to be associated with the decision to re-enter entrepreneurship. The analysis also distinguishes between entrepreneurs that started a business from those who acquired one, following Ucbasaran et al. (2003) and Stam et al. (2008). In our estimates, industry is controlled for as specific business environments may impact differently upon firms’ profitability and performance. Moreover, different industries may be associated with different business opportunities (Shane 2003), and it may influence individuals’ decision to re-enter, or not, entrepreneurship (Westhead and Wright 1998). Dummy variables are employed to distinguish among Primary sector, Manufacturing, Energy and construction, Services, and Community, social and personal services.

4.3.3 Macroeconomic context

As individuals’ decisions to enter and exit entrepreneurship may be influenced by the unemployment level (Carree et al. 2007; Audretsch et al. 2005), a control variable for unemployment is also included, thus accounting for the macroeconomic environment. This variable was constructed by calculating the unemployment rate variations relative to the homologous last quarters in 1986–2003 using official data gathered by INE–Statistics Portugal.

5 Results

This section presents the results from estimating the impact of general and specific human capital on individuals’ hazard rate of re-entering entrepreneurship, thus becoming serial entrepreneurs. Figure 1 shows the baseline hazards for individuals aged up to 30 (dashed line) and for all individuals (thick line). The baseline hazards for re-entering entrepreneurship have negative duration dependence. Furthermore, individuals aged up to 30 have a slightly higher baseline when compared with all individuals, but the pattern remains the same. For both groups, the hazard of re-entering entrepreneurship is much higher during the first years following exit from a previous entrepreneurial experience. This negative relationship between time and the hazard of re-entry into entrepreneurship is also shown in Table 3, where in estimates I–II the time dummy controlling for re-entry between the 2nd and 4th years ($t_2$) is about 2.2 times smaller than the hazard for the first year ($t_1$); re-entry in the 5th to 7th years ($t_3$) is about 2.6 times smaller than in the 1st year, and, finally, re-entry between the 8th and 17th years ($t_4$) is three times smaller than in the first year.

As noted in Sect. 3.2, the empirical analysis focuses primarily on Table 3, which includes only entrepreneurs who have left their initial firm prior to turning 31.13 Estimates I and II in Table 3 differ in that while I includes individual, firm, industry and macro characteristics, II adds interaction terms between start-up experience and the mode of exit from the first firm (sell-out versus dissolution) in order to test the effect of a “positive entrepreneurial

---

13 Nevertheless, in Appendix Table 5 we provide similar estimations, but considering all observations within our data set. These results will be focused on further ahead in this section.
experience” (the interaction: “start up” × “sell-out” representing founders who have sold their business) on the hazard of re-entering entrepreneurship.

The estimates show that, as proposed in Hypothesis 1 (H1), ex-entrepreneurs who are better educated will postpone their decision to re-enter entrepreneurship. In effect, education is negatively associated with the hazard of re-entering entrepreneurship. A negative impact on the hazard means that...
education has lower conditional re-entry rates ($\exp(\beta) = 0.98$), and therefore, highly educated individuals experience longer periods in-between two entrepreneurial events. Hence, the first hypothesis is confirmed.

The argument, put forward in Hypothesis 2 ($H2$), that ex-entrepreneurs with more experience in paid-employment are more likely to delay re-entry into entrepreneurship is not confirmed because coefficients, although negative, are insignificant. One possible interpretation of this result is that entrepreneurs who left their business at up to 30 years old have not had sufficient time to build a significant stock of experience in the labor market.

The argument proposed in Hypothesis 3 ($H3$) that ex-entrepreneurs with more experience in business ownership will re-enter entrepreneurship more quickly is confirmed. Individuals with 1 additional year of experience in entrepreneurship are about 4% more likely to re-enter entrepreneurship ($\exp(\beta) = 1.041$).

Hypothesis 4 ($H4$) claims that ex-entrepreneurs with managerial attainment will re-enter entrepreneurship more quickly. Results provide empirical support for a strong negative impact of managerial attainment on time to re-enter entrepreneurship. Individuals with managerial experience are about 36% more likely to re-enter entrepreneurship ($\exp(\beta) = 1.355$). Therefore, the fourth hypothesis is confirmed.

Finally, Hypothesis 5 ($H5$) proposes that ex-entrepreneurs with “positive founding experience” are likely to re-enter entrepreneurship more quickly. As detailed in Sect. 2.2 (when discussing Hypothesis 5), the mode of exit is used to proxy for the positive or negative aspects of founding experience on serial entrepreneurship. Firm founding can lead to a negative outcome if there is firm closure/dissolution. Consequently, this particular type of founding experience can be seen as a liability rather than an asset. Interactions between founding experience and modes of exit show that individuals who have founded their own firm and, later on, dissolve it are more likely to delay their re-entry. Conversely, those who found their firm and then exit through sell-out are more likely to re-enter entrepreneurship sooner. This result suggests that a positive founding experience (i.e., the individual sells his firm as an exit strategy) contributes to speed up the decision to become a serial entrepreneur. On the contrary, starting a firm that later on is closed/dissolved has a negative effect on the hazard of re-entering (i.e., there may be a “scarring” effect, or a stigma of failure). Therefore, the results confirm Hypothesis 5, by showing those individuals who have had a “positive founding experience” (start-up and subsequent sell-out) are nearly 17% ($\exp(\beta) = 1.171$) more likely to re-enter entrepreneurship than individuals with a “negative founding experience” (start-up and subsequent dissolution).

A set of controls is included in the estimates. Variables controlling for entrepreneurs’ demographics show that, in general, males and older individuals tend to re-enter more quickly than females and the younger. The age effect is, however, insignificant, which is expected since the study focuses on a sample of individuals of similar age (younger than 31 years old at the time of exit).

The logarithm of firm age has a significant negative effect on the hazard of re-entering entrepreneurship, meaning that individuals who have left older firms will have a lower hazard of re-entering entrepreneurship. Individuals who leave larger firms are more likely to quickly re-enter entrepreneurship. One possible explanation is that, by leaving a large firm, the probability of being endowed with more human, financial and physical resources is greater. All estimations consistently show that entrepreneurial team size negatively impacts the hazard of re-entering entrepreneurship, so an individual leaving a firm with no partners will be more likely to rapidly become a serial entrepreneur. One possible explanation for this result is that these individuals do not feel they need to search for the complementary skills provided by entrepreneurial partners.

Exit with firm dissolution—instead of sell-out or transfer—has a positive coefficient, which means it decreases the time to re-enter entrepreneurship. However, when looking exclusively at ex-entrepreneurs who started their first business, firm dissolution has a negative effect on the hazard (Table 3, Estimate II) of re-entry. This finding is consistent with Stam et al. (2008).

When controlling for the macroeconomic context it is found that, in a context of high unemployment, individuals are more likely to postpone their decision to re-enter entrepreneurship. There is, to our knowledge, no study concentrating on unemployment dynamics and serial entrepreneurship. Nevertheless, this finding is consistent with the “prosperity pull” rather than the “unemployment push” hypothesis of
entrepreneurial entry (Evans and Leighton 1990; Blanchflower and Oswald 1991; Abell et al. 1995; Carrasco 1999). This suggests serial entrepreneurs are influenced by the more optimistic expectations or more numerous business opportunities that occur during economic booms.

Even though our analysis is focused on a sub-sample composed of ex-entrepreneurs aged up to 30 years old, we include supplementary estimations for all ex-entrepreneurs in Appendix Table 5. Results for all ex-entrepreneurs do not differ significantly from the ones obtained for ex-entrepreneurs aged up to 30 years old regarding the impact of education and “positive start-up experience,” thereby confirming H1 and H5.

While paid-employment experience has a negative but insignificant impact on the hazard of re-entry into entrepreneurship in estimations for the sub-sample including only individuals younger than 31, this negative effect is significant for the whole sample; it is, however, of small magnitude: a 1-year increase in paid-employment experience leads to a reduction of about 2% in the hazard of re-entering entrepreneurship. Therefore, H2 is confirmed when looking at all ex-entrepreneurs.

H3 and H4 are not confirmed for the sample including all individuals. The effect of the number of years in business ownership (H3) becomes negative and insignificant, which may be explained by the fact that this sample includes older individuals who are closer to retirement and eventually not up to the challenge of re-entering entrepreneurship. Top management experience (H4) has a significant but negative impact on the hazard of re-entry into entrepreneurship for the overall sample. One possible explanation is that although there is a high percentage of ex-entrepreneurs who are also top managers (nearly 80% of the sample), management experience among younger and less experienced individuals is shorter and may be less valued by hiring firms than for older and more experienced individuals. Therefore, these older people would benefit from their top management experience in the job market, having better perspectives of finding attractive paid employment. This would raise the opportunity cost of entrepreneurship, thereby delaying the decision to engage on a subsequent entrepreneurial experience.

As the impacts of business ownership experience and managerial attainment on the hazard of re-entering entrepreneurship become negative in the sample including all ex-entrepreneurs, particular attention should be paid to initial conditions when studying serial entrepreneurship, principally those related with individuals’ age.

The major findings for the group of interest—ex-entrepreneurs aged up to 30—show the distinct effects of general human capital, measured as formal education and paid-employment experience, as well as entrepreneurial specific human capital, as measured by years in entrepreneurship, managerial attainment and “positive founding experience.” While variables fitting the general human capital category have a negative impact upon the hazard of re-entering entrepreneurship, variables associated with specific human capital have a positive impact.

As a robustness check to our empirical approach, additional continuous-time models are estimated using a piecewise constant exponential model with interval-specific intercept terms to capture the hazard for each year (in a total of 17 years). The use of this alternative baseline does not impact significantly or change our findings. Furthermore, an unbalanced panel data set-up is used in order to apply a complementary log–log discrete specification for both ex-entrepreneurs aged up to 30 years and all ex-entrepreneurs (see Appendix Table 5). Results from the discrete approach are consistent with those of the continuous approach.

6 Concluding remarks

The present paper uses a longitudinal matched employer–employee database to examine serial entrepreneurship, more specifically the determinants of the time in-between entrepreneurial events. In particular, the study focuses on how individuals’ general and specific human capital influences their decision to re-enter entrepreneurship in a different (new or acquired) business, thereby becoming serial entrepreneurs.

Results from model estimation using ex-entrepreneurs aged up to 30 years old highlight important

---

14 The average age for all ex-entrepreneurs is of 43.32 years (with a high standard deviation of 12.57), while the figure for ex-entrepreneurs aged up to 30 is of 24.75 years. Moreover, as shown in the Correlations Table (Appendix Table 4), age is strongly correlated with business ownership experience.
differences between general and entrepreneurial-specific human capital on the time necessary to become a serial entrepreneur. The empirical analysis reveals negative effects of general human capital upon the hazard of becoming a serial entrepreneur and a clearly overall positive impact of entrepreneurial-specific human capital on time to re-entering entrepreneurship. While higher levels of education and employment experience are likely to delay ex-entrepreneurs’ decision to re-enter, more years of previous entrepreneurial experience, previous managerial attainment and having had a “positive founding experience” are likely to speed up individuals’ return to entrepreneurship. Results suggest that even though the success (asset) or failure (liability) of individuals’ previous entrepreneurial experience may play a role on re-entry, the speed with which individuals’ re-enter entrepreneurship is fundamentally related to their general and specific human capital.

These findings are consistent with human capital theory since individuals whose human capital is more specific to entrepreneurship are less likely to delay re-entering entrepreneurship. Another explanation is that ex-entrepreneurs endowed with higher entrepreneurial-specific human capital may also face higher opportunity costs of choosing a different occupation and are consequently more likely to accelerate the decision to re-enter entrepreneurship.

A concurrent view to the opportunity costs approach is the one of entrepreneurial-opportunity identification and pursuit. Our results are consistent with Ucbasaran et al. (2008) who find that, while general human capital variables have lower “explanatory” power with regard to opportunity identification and pursuit, several aspects of entrepreneurship-specific human capital are significantly associated with both a higher probability of identifying more opportunities and pursuing more opportunities.

From the evidence, some crucial implications emerge. First, this study contributes to the literature on serial entrepreneurship since, more than simply analyzing serial entrepreneurship as a static phenomenon, it focuses on time to re-enter as the main variable of interest. By monitoring the skills and experience of each type of entrepreneur, and the ‘flows’ across occupations over time, the research presented here reveals an enriched process of occupational choice for serial entrepreneurs, providing better contextual and empirical evidence. As both the theoretical approach and empirical results suggest, further studies on serial entrepreneurship should account for the fact that different forms of human capital impact how individuals learn from their first business ownership differently, affecting their propensity to delay or hasten entry in a subsequent entrepreneurial event.

There are some limitations that need to be acknowledged and addressed. The first limitation concerns the fact that this study looks at human capital and other variables’ effects on time to re-enter entrepreneurship, measuring such variables at the time of exiting the first entrepreneurial experience. Hence, it does not address individuals’ occupational choices and experiences in between exit and re-entry. Further research on serial entrepreneurship can benefit from studying the dynamics of choosing different alternative occupations upon exiting entrepreneurship. For example, competing risks models with time-variant variables can be applied to investigate in more detail the occupational path of ex-entrepreneurs.

This study is also limited by the fact that it focuses on transitions from entrepreneurship in one firm to entrepreneurship in a subsequent firm. Since serial entrepreneurship may entail multiple transitions (i.e., individuals may exit and re-enter more than once), future research can extend the present analysis by using mixed proportional hazard models and its multivariate extensions to account for this aspect.

Since time to entrepreneurial re-entry is not just associated with human capital factors, but also with a stigma of failure, there is a need to replicate this research using data from different countries, institutional settings and periods. Complementarily to time-to-event data modeling accounting for the time lag between entrepreneurial events, panel data estimation techniques can be used to assess sequences of different occupations and decisions of serial entrepreneurs through time, therefore shedding new light on the entrepreneurial process.

Appendix

See Tables 4, 5, 6.
Table 4 Correlations table (for all ex-entrepreneurs)

<table>
<thead>
<tr>
<th></th>
<th>Female (dummy)</th>
<th>Age</th>
<th>Years of education</th>
<th>Years in non-employment</th>
<th>Years as a paid-employee</th>
<th>Years as a business owner</th>
<th>Top management (dummy)</th>
<th>Start-up (dummy)</th>
<th>Log firm size (no. of workers)</th>
<th>Log no. of partners (owners)</th>
<th>Firm dissolution (dummy)</th>
<th>Primary sector (dummy)</th>
<th>Total manufacturing (dummy)</th>
<th>Energy and construction (dummy)</th>
<th>Services (dummy)</th>
<th>Unemployment variation rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female (dummy)</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.078</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of education</td>
<td>0.069</td>
<td>-0.230</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years in non-employment</td>
<td>0.023</td>
<td>-0.031</td>
<td>0.027</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years as a paid-employee</td>
<td>0.008</td>
<td>-0.025</td>
<td>0.040</td>
<td>0.251</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years as a business owner</td>
<td>-0.094</td>
<td>0.455</td>
<td>-0.123</td>
<td>-0.173</td>
<td>-0.246</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Top management (dummy)</td>
<td>0.011</td>
<td>0.094</td>
<td>0.175</td>
<td>0.073</td>
<td>0.047</td>
<td>0.012</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up (dummy)</td>
<td>0.021</td>
<td>-0.205</td>
<td>0.067</td>
<td>0.030</td>
<td>0.142</td>
<td>-0.257</td>
<td>0.077</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log firm age</td>
<td>-0.051</td>
<td>0.421</td>
<td>-0.112</td>
<td>-0.097</td>
<td>-0.129</td>
<td>0.550</td>
<td>-0.053</td>
<td>-0.607</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log firm size (no. of workers)</td>
<td>-0.067</td>
<td>0.043</td>
<td>0.039</td>
<td>-0.086</td>
<td>-0.018</td>
<td>0.160</td>
<td>-0.141</td>
<td>-0.132</td>
<td>0.303</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log no. of partners (owners)</td>
<td>-0.020</td>
<td>-0.023</td>
<td>-0.029</td>
<td>-0.035</td>
<td>-0.019</td>
<td>0.064</td>
<td>-0.080</td>
<td>-0.041</td>
<td>0.089</td>
<td>0.388</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Firm dissolution (dummy)</td>
<td>0.037</td>
<td>-0.072</td>
<td>0.041</td>
<td>0.134</td>
<td>0.085</td>
<td>-0.099</td>
<td>0.093</td>
<td>0.188</td>
<td>-0.215</td>
<td>-0.242</td>
<td>-0.131</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary sector (dummy)</td>
<td>-0.025</td>
<td>0.012</td>
<td>-0.023</td>
<td>0.001</td>
<td>-0.023</td>
<td>-0.016</td>
<td>-0.029</td>
<td>-0.013</td>
<td>-0.002</td>
<td>-0.011</td>
<td>-0.026</td>
<td>0.005</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total manufacturing (dummy)</td>
<td>-0.023</td>
<td>-0.001</td>
<td>-0.090</td>
<td>-0.016</td>
<td>0.031</td>
<td>0.060</td>
<td>-0.122</td>
<td>-0.021</td>
<td>0.075</td>
<td>0.280</td>
<td>0.089</td>
<td>-0.027</td>
<td>-0.074</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Energy and construction (dummy)</td>
<td>-0.129</td>
<td>-0.049</td>
<td>-0.084</td>
<td>0.036</td>
<td>-0.021</td>
<td>-0.055</td>
<td>-0.038</td>
<td>0.063</td>
<td>-0.095</td>
<td>0.051</td>
<td>-0.007</td>
<td>0.066</td>
<td>-0.055</td>
<td>-0.189</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Services (dummy)</td>
<td>0.040</td>
<td>0.039</td>
<td>0.094</td>
<td>-0.012</td>
<td>0.002</td>
<td>-0.003</td>
<td>0.172</td>
<td>-0.021</td>
<td>0.006</td>
<td>-0.239</td>
<td>-0.054</td>
<td>-0.023</td>
<td>-0.181</td>
<td>-0.618</td>
<td>-0.460</td>
<td>1.000</td>
</tr>
<tr>
<td>Unemployment variation rate</td>
<td>0.017</td>
<td>0.011</td>
<td>0.069</td>
<td>0.193</td>
<td>0.052</td>
<td>-0.035</td>
<td>0.057</td>
<td>0.074</td>
<td>-0.055</td>
<td>-0.044</td>
<td>0.001</td>
<td>0.345</td>
<td>0.003</td>
<td>-0.029</td>
<td>0.046</td>
<td>-0.014</td>
</tr>
</tbody>
</table>
Table 5 Re-entry into entrepreneurship: proportional hazards specification with piece-wise constants baseline hazard (for all ex-entrepreneurs)

<table>
<thead>
<tr>
<th></th>
<th>All ex-entrepreneurs</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate I</td>
<td>Estimate II</td>
<td></td>
</tr>
<tr>
<td><strong>Human capital</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years of education</td>
<td>-0.021***</td>
<td>-0.022***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.002]</td>
<td>[0.002]</td>
<td></td>
</tr>
<tr>
<td>Years as a paid-employee</td>
<td>-0.014***</td>
<td>-0.014***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.002]</td>
<td>[0.002]</td>
<td></td>
</tr>
<tr>
<td>Years as a business owner</td>
<td>-0.002</td>
<td>-0.002</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.002]</td>
<td>[0.002]</td>
<td></td>
</tr>
<tr>
<td>Top management (dummy)</td>
<td>-0.102***</td>
<td>-0.104***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.021]</td>
<td>[0.021]</td>
<td></td>
</tr>
<tr>
<td>Start-up (dummy)</td>
<td>-0.078***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.022]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Demographics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female (dummy)</td>
<td>-0.349***</td>
<td>-0.348***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.018]</td>
<td>[0.018]</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.089***</td>
<td>0.088***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.005]</td>
<td>[0.005]</td>
<td></td>
</tr>
<tr>
<td>Age squared/100</td>
<td>-0.116***</td>
<td>-0.115***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.005]</td>
<td>[0.005]</td>
<td></td>
</tr>
<tr>
<td><strong>Non-employment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years in non-employment</td>
<td>-0.010**</td>
<td>-0.010**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.004]</td>
<td>[0.004]</td>
<td></td>
</tr>
<tr>
<td><strong>Firm characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log firm age</td>
<td>-0.025**</td>
<td>-0.018*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.011]</td>
<td>[0.011]</td>
<td></td>
</tr>
<tr>
<td>Log firm size (no. of workers)</td>
<td>0.099***</td>
<td>0.094***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.009]</td>
<td>[0.009]</td>
<td></td>
</tr>
<tr>
<td>Log no. of partners (owners)</td>
<td>-0.174***</td>
<td>-0.174***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.018]</td>
<td>[0.018]</td>
<td></td>
</tr>
<tr>
<td>Firm dissolution (dummy)</td>
<td>0.275***</td>
<td>0.372***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.017]</td>
<td>[0.019]</td>
<td></td>
</tr>
<tr>
<td><strong>Industry characteristics</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary sector (dummy)</td>
<td>-0.253***</td>
<td>-0.254***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.064]</td>
<td>[0.064]</td>
<td></td>
</tr>
<tr>
<td>Total manufacturing (dummy)</td>
<td>-0.161***</td>
<td>-0.163***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.039]</td>
<td>[0.039]</td>
<td></td>
</tr>
</tbody>
</table>

Table 5 continued

|                                | All ex-entrepreneurs |          |          |
|                                |                       | Estimate I | Estimate II |
| Energy and construction        | -0.037                | -0.035     |
| (dummy)                        |                       | [0.040]    | [0.040]    |
| Services (dummy)               | -0.039                | -0.037     |
|                                | [0.036]               | [0.036]    |
| **Macroeconomic context**      |                       |          |          |
| Unemployment variation rate    | -2.771***             | -2.756*** |
|                                | [0.045]               | [0.045]   |
| **Interaction terms**          |                       |          |          |
| Start-up × firm dissolution    | -0.207***             |          |
| ("negative founding experience") |                       | [0.025]    |
| Start-up × firm sell-out       | 0.230***              |          |
| ("positive founding experience") |                       | [0.033]    |
| **Time dummies**               |                       |          |          |
| \( r2 = [3,5] \)               | -2.701***             | -2.691*** |
|                                | [0.019]               | [0.019]   |
| \( r3 = [6,9] \)               | -3.856***             | -3.835*** |
|                                | [0.028]               | [0.028]   |
| \( r4 = [10,17] \)             | -4.972***             | -4.947*** |
|                                | [0.047]               | [0.047]   |
| Constant                       | -2.833***             | -2.887*** |
|                                | [0.108]               | [0.108]   |
| Log likelihood                 | -52238                | -52168    |
| Log likelihood                 |                       | 54,598    |
| No. of observations            | 224,805               | 224,805   |
| No. of observations            |                       | 176,747   |
| No. of failures                | 19,074                | 19,074    |
| No. of failures                |                       |           |

* Significant at 10%; ** significant at 5%; *** significant at 1%; standard errors in brackets. Coefficients indicate variables’ effects on the hazard function; a positive coefficient increasing the probability of serial entrepreneurship and thereby decreasing expected time to re-entry. Reference categories for dummy variables: not a top manager, acquisition, male, firm continuance through sell-out. “Community Social and Personal Services” is the omitted reference for industry characteristics. The interval \( t1 = [0,1] \) is the omitted reference for our time dummies.
Table 6 Re-entry into entrepreneurship: complementary log–log discrete specification with piece-wise constants baseline hazard

<table>
<thead>
<tr>
<th>Human capital</th>
<th>Ex-entrepreneurs aged ≤30</th>
<th>All ex-entrepreneurs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate I</td>
<td>Estimate II</td>
</tr>
<tr>
<td>Years of education</td>
<td>−0.020***</td>
<td>−0.021***</td>
</tr>
<tr>
<td></td>
<td>[0.006]</td>
<td>[0.006]</td>
</tr>
<tr>
<td>Years as a paid-employee</td>
<td>−0.018*</td>
<td>−0.017*</td>
</tr>
<tr>
<td></td>
<td>[0.009]</td>
<td>[0.009]</td>
</tr>
<tr>
<td>Years as a business owner</td>
<td>0.042***</td>
<td>0.041***</td>
</tr>
<tr>
<td></td>
<td>[0.008]</td>
<td>[0.008]</td>
</tr>
<tr>
<td>Top management (dummy)</td>
<td>0.348***</td>
<td>0.346***</td>
</tr>
<tr>
<td></td>
<td>[0.066]</td>
<td>[0.066]</td>
</tr>
<tr>
<td>Start-up (dummy)</td>
<td>−0.083</td>
<td>−0.097***</td>
</tr>
<tr>
<td></td>
<td>[0.061]</td>
<td>[0.022]</td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female (dummy)</td>
<td>−0.456***</td>
<td>−0.454***</td>
</tr>
<tr>
<td></td>
<td>[0.054]</td>
<td>[0.054]</td>
</tr>
<tr>
<td>Age</td>
<td>0.008</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>[0.131]</td>
<td>[0.131]</td>
</tr>
<tr>
<td>Age squared/100</td>
<td>0.149</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>[0.267]</td>
<td>[0.268]</td>
</tr>
<tr>
<td>Non-employment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years in non-employment</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>[0.016]</td>
<td>[0.016]</td>
</tr>
<tr>
<td>Firm characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log firm age</td>
<td>−0.075**</td>
<td>−0.061*</td>
</tr>
<tr>
<td></td>
<td>[0.033]</td>
<td>[0.033]</td>
</tr>
<tr>
<td>Log firm size (no. of workers)</td>
<td>0.099***</td>
<td>0.101***</td>
</tr>
<tr>
<td></td>
<td>[0.027]</td>
<td>[0.027]</td>
</tr>
<tr>
<td>Log no. of partners (owners)</td>
<td>−0.151***</td>
<td>−0.150***</td>
</tr>
<tr>
<td></td>
<td>[0.050]</td>
<td>[0.050]</td>
</tr>
<tr>
<td>Firm dissolution (dummy)</td>
<td>0.236***</td>
<td>0.353***</td>
</tr>
<tr>
<td></td>
<td>[0.051]</td>
<td>[0.062]</td>
</tr>
<tr>
<td>Industry characteristics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primary sector (dummy)</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>[0.196]</td>
<td>[0.196]</td>
</tr>
<tr>
<td>Total manufacturing (dummy)</td>
<td>0.166</td>
<td>0.163</td>
</tr>
<tr>
<td></td>
<td>[0.127]</td>
<td>[0.127]</td>
</tr>
<tr>
<td>Energy and construction (dummy)</td>
<td>0.284**</td>
<td>0.286**</td>
</tr>
<tr>
<td></td>
<td>[0.130]</td>
<td>[0.130]</td>
</tr>
<tr>
<td>Services (dummy)</td>
<td>0.158</td>
<td>0.158</td>
</tr>
<tr>
<td></td>
<td>[0.119]</td>
<td>[0.119]</td>
</tr>
<tr>
<td>Macroeconomic context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment variation rate</td>
<td>−2.288***</td>
<td>−2.283***</td>
</tr>
<tr>
<td></td>
<td>[0.126]</td>
<td>[0.126]</td>
</tr>
</tbody>
</table>
Table 6 continued

<table>
<thead>
<tr>
<th>Interaction terms</th>
<th>Ex-entrepreneurs aged ≤30</th>
<th>All ex-entrepreneurs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate I</td>
<td>Estimate II</td>
</tr>
<tr>
<td>Start-up × firm dissolution</td>
<td>−0.171***</td>
<td>−0.235***</td>
</tr>
<tr>
<td>(“negative founding experience”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up × firm sell-out</td>
<td>0.161*</td>
<td>0.236***</td>
</tr>
<tr>
<td>(“positive founding experience”)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time dummies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t2</td>
<td>−2.324***</td>
<td>−2.326***</td>
</tr>
<tr>
<td></td>
<td>[0.060]</td>
<td>[0.060]</td>
</tr>
<tr>
<td>t3</td>
<td>−2.735***</td>
<td>−2.732***</td>
</tr>
<tr>
<td></td>
<td>[0.100]</td>
<td>[0.100]</td>
</tr>
<tr>
<td>t4</td>
<td>−3.224***</td>
<td>−3.209***</td>
</tr>
<tr>
<td></td>
<td>[0.181]</td>
<td>[0.181]</td>
</tr>
<tr>
<td>Constant</td>
<td>−3.260**</td>
<td>−3.377**</td>
</tr>
<tr>
<td></td>
<td>[1.607]</td>
<td>[1.611]</td>
</tr>
<tr>
<td>LR χ²</td>
<td>−6852</td>
<td>−6847</td>
</tr>
<tr>
<td>No. of observations</td>
<td>47,711</td>
<td>47,711</td>
</tr>
<tr>
<td>No. of failures</td>
<td>2,096</td>
<td>2,096</td>
</tr>
</tbody>
</table>

* Significant at 10%; ** significant at 5%; *** significant at 1%; standard errors in brackets. Coefficients indicate variables’ effects on the hazard function; a positive coefficient increasing the probability of serial entrepreneurship and thereby decreasing expected time to re-entry. Reference categories for dummy variables: not a top manager, acquisition, male, firm continuance through sell-out. “Community Social and Personal Services” is the omitted reference for industry characteristics. The interval t1 = [0, 1] is the omitted reference for our time dummies.

References


Cleeves, M., Gould, W., & Gutierrez, R. (2002). An introduction to survival analyses using Stata. College Station, TX: Stata Press.


Serial entrepreneurship: impact of human capital on time to re-entry


Chapter 13:

Miguel Preto and Rui Baptista, “Knowledge based start-ups, regional agglomeration and economic performance”.
INTRODUCTION

Economic analysis has increasingly focused on the economic benefits of entrepreneurship in terms of, for instance, employment generation and innovation (van Praag and Versloot 2007). A recent stream of studies examines whether there is a relationship between increases in new firm formation rates and subsequent employment growth at the regional level. These studies show that the impact of new business formation on regional development is distributed over a relatively long period of time, usually ten years. Comparable patterns of results are found for Germany (Fritsch and Mueller 2004, 2008), Great Britain (Mueller, van Stel and Storey 2008), the Netherlands (van Stel and Suddle 2008), Portugal (Baptista, Escária and Madruga 2008), Spain (Arauzo-Carod, Liviano-Solis and Martin-Bofarull 2008), and the US (Acs and Mueller 2008).

New firms are generally smaller than the average incumbent, and their direct contribution to the stock of jobs in an economy is relatively small (Van Stel and Storey 2004). Moreover, new businesses have a greater probability of failure than old businesses. According to Geroski (1995), the survival probability of most entrants is low and successful entrants may take more than a decade to achieve a size comparable to that of the average incumbent. It is therefore striking that a key finding shared by the aforementioned stream of studies for different countries is that there is a positive relationship between new firm formation and subsequent employment growth. These studies found a similar pattern for the lag structure of the effects of new firm formation on employment growth over time. First, the magnitude of direct employment creation in entry cohorts was found to be small. Second, new entrants subsequently crowd out inefficient firms, lowering employment. Third, positive supply-side effects increase overall employment significantly, through the growth of successful entrants and incumbents.

1 This chapter is based on a paper published in Small Business Economics, No 36, Vol. 4, pp. 419-442.
The remarkable similarity in the patterns of results observed across countries suggests that there are three different kinds of impacts of new firm formation on subsequent employment change, and that these impacts do not occur simultaneously, but in different points in time.

i. First, a direct impact of employment creation by firm $j$ in time $t$ is observed;  

ii. Second, there is a period when displacement of existing jobs occurs, possibly as a result of increased competition and market selection;  

iii. Third, there are positive long term impacts, possibly associated with increased competitiveness and innovation brought about by successful new firms that spills over to the industry.

Even though patterns of effects of new business formation on employment growth are similar across countries, there seem to be pronounced differences in the magnitude and specific timing of these effects. Moreover, differences across regions within countries are also observed (Fritsch and Mueller 2008). Such differences are significant even when one controls for different regional industrial structures. This finding suggests that, even when one controls for economic structure, different regions display different kinds of industrial dynamics, and attract different kinds of new firms.

In particular, the studies by Acs and Mueller (2008) and Fritsch and Mueller (2008) revealed pronounced differences in the magnitude and timing of the effects of new firm creation on subsequent employment growth across regions within countries. In their study for the US, Acs and Mueller (2008) find that large consolidated metropolitan areas are fertile ground for the growth of new businesses, whereas small towns and cities may register high rates of new firm entry but cannot support the expansion of rapidly growing firms.

Similar disparities exist across different countries. Baptista, Escária and Madruga (2008) find that the positive long term impacts of new firm formation on subsequent employment growth in the case of Portugal are smaller and take longer to occur than the same kind of effects in the comparable case of Germany, as found by Fritsch and

---

2 Economic activities in different regions differ accordingly to their composition in terms of sectors; therefore, regions with greater proportions of firms in declining sectors should display lower impacts of new firm formation on employment growth than regions with a greater proportion of firms in growing sectors. A shift-share procedure was used to correct for this kind of difference in, for instance, the studies by Baptista, Escária and Madruga (2008); and Mueller, van Stel and Storey (2008). Differences across regions in the effects of new firm formation on employment growth remained significant.
Mueller (2004). As these studies controlled for differences in regional industrial structures, such disparities are likely due either to differences in business dynamics (i.e. the qualities of the new firms being started) or to differences in regional/national characteristics that may bolster the positive long term impacts of new firm creation on employment growth.

The present paper uses data for Portugal to examine in detail how differences in regional and business dynamics may contribute to generate disparities among regions in the magnitude and timing of employment growth effects. In particular, we examine differences between the long-term effects on employment growth of new firm creation in knowledge-based sectors and in other sectors for regions with different levels of productivity and agglomeration of economic activity.

The paper proceeds as follows. The next section provides a brief exploration of the regional and business dynamics that may lead to differences in the magnitude and timing of the long term effects of new business formation on employment growth within and across regions. Section 3 focuses on the empirical analysis. Section 4 presents and discusses the results, and section 5 provides some concluding remarks.

EFFECTS OF NEW FIRM FORMATION ON EMPLOYMENT GROWTH: REGIONAL AND BUSINESS DYNAMICS

The relationship between new firm entry and economic growth has been addressed by recent theory and empirical work. Acs et al. (2004) argue that entrepreneurship (i.e. new firm creation) contributes to economic growth by penetrating the “knowledge filter” that prevents new knowledge from spilling over to economic agents. In this way, new firm creation facilitates the working of the fundamental mechanism behind sustained economic growth (Romer 1986, 1990). Recent empirical studies find a positive relationship between new firm entry and productivity growth (Disney, Haskel and Heden 2003; Aghion et al. 2004).

While the creation of new firms may play a significant role in spawning regional economic growth, the extent of the effects of entrepreneurial activity on subsequent growth should vary across regions according to the pools of innovative opportunities and human capital available in each region (Shane 1996; Baptista and Mendonça 2009). Business and regional dynamics are strongly inter-related, and play an important role in
determining the impact of new firm entry on economic development and employment growth. Regions may differ considerably with regard to the characteristics of new and incumbent businesses as well as with regard to their ability to absorb the positive effects of new business formation. In order to analyze such differences, they distinguish between types of regions according to different criteria, including the degree of agglomeration and the average labor productivity. Acs and Mueller (2008) look at differences in levels of business dynamics (i.e. proportion of rapidly growing firms) between regions; not surprisingly, they find that most rapidly growing firms are located in the larger (i.e. more agglomerated) metropolitan regions. It seems therefore reasonable to assume that there is a positive relationship between the degree of regional agglomeration and business dynamics levels.

The role played by agglomeration effects – or externalities – in promoting supply-side spillovers is widely addressed in the literature (see for instance Baptista 1998; and Audretsch 2003). There is a general belief that location matters to the development and growth of industries. Much literature has been developed around the notion that firms tend to concentrate in certain regions so they can benefit from co-location. Feldman (1994) argues that spillovers associated with innovation are stronger within relatively restricted geographical regions due to agglomeration externalities that increase the capacity of firms to tap into the local pool of new ideas, while Jaffe, Henderson and Trajtenberg (1993) provide evidence of geographical concentration of spillovers on innovative (patenting) activity; Baptista (2000) finds that the probability of a firm adopting an innovation depends positively on the local presence of other adopters.

Other studies have argued that agglomeration externalities influence business dynamics directly through the process of displacement that determines which firms survive and grow, and which firms fail (Acs and Mueller 2008). Agglomeration externalities influence firm competitiveness and growth through mechanisms that involve both concentration and diversity of industries (Glaeser et al. 1992; Blien, Südekum and Wolf 2006), and may also result from efficiency gains due to increased competition: several empirical studies support the conjecture of a relatively high level of competition in agglomerations. These studies find higher rates of start-ups (Fritsch and

---

3 According to Acs and Mueller (2008), 40 percent of all the rapidly growing firms are located in only 20 metropolitan regions, which are mostly the largest cities in the United States.

4 Glaeser et al. (1992) find a positive correlation between agglomeration externalities and firm growth levels.
Falck 2007) and a lower probability of survival (Fritsch, Brixy and Falck 2006) in more agglomerated regions.

Baptista, Escária and Madruga (2008) suggest that a substantial part of the differences across regions and countries in the size and structure of lagged effects of entry on employment change are likely to be due to differences in types and/or qualities of start-ups. The size of negative (market selection) and positive (supply-side spillovers) effects and the lag time for those effects to ensue should vary according to the type of entrant, as not all entrants are equally efficient and/or innovative, and, therefore, not all have the same impact. While it is acknowledged that the emergence of positive supply-side effects from new firm formation does not require that newcomers are successful, it is expected that different kinds of start-ups will have different impacts on the industrial re-structuring process according to the “quality” of new entrants with regard to innovation, efficiency, and product differentiation. New firms provide a vehicle for the introduction of innovations into an economy, therefore being a source of both industry turbulence (Beesley and Hamilton 1984) and productivity growth (Aghion et al. 2004).

Even though, as pointed out by Van Stel and Storey (2004), innovation in new firms seems to be not as frequent as expected, a significant contribution by new entrants to employment growth occurs through increased competitiveness and enhanced efficiency of incumbents. In a recent study for Germany, Fritsch and Noseleit (2009) find that the employment effects of new businesses on the incumbents are significantly positive and considerably larger than the employment that is directly generated in the start-ups.

When discussing the transition from the managed economy to the entrepreneurial economy, Audretsch and Thurik (2004) stress the role played by the increasing pace of technological progress. In the managed economy technological trajectories were relatively well-defined and firms were subject to relatively low uncertainty, while in the entrepreneurial economy product life-cycles are short and competitive conditions change rapidly. In a meta-analysis of the empirical evidence that net employment

5 Hoetker and Agarwal (2007) find that significant diffusion of knowledge may occur after an innovative firm exits an industry if other firms are able to use that firm’s activities as a template to successfully replicate and extend its innovative knowledge.
6 Baptista and Preto (2009) seek to examine the effects on employment growth of start-up rates according to different types of entrants (e.g. small versus large start-ups and domestics vs. foreign start-ups).
7 Disney, Haskel and Heden (2003) find that in the UK between 1980 and 1992 about half of productivity gain was because of internal factors, such as introducing new technology and organisational changes. The remaining half was because of external factors, most notably that the entrants were more productive than those exiting. However amongst single plant independent firms almost all the gains were attributable to external factors.
growth is generated by only a few rapidly growing firms, Henrekson and Johansson (2008) do not find that these firms are disproportionately high tech. However, the large majority fits in the less restricted qualification of knowledge-based enterprises (KBEs). A greater presence of knowledge-based entrants is likely to boost the introduction of innovations in the market. Knowledge-based industries tend to have shorter product and technology life-cycles and, being less focused on operational economies of scale, provide more opportunities for new, small firms to induce market re-structuring and change through innovation and efficiency improvements. It can therefore be argued that both employment destruction (due to increased competition and displacement) and employment creation (due to positive supply-side spillovers) will be greater the higher is the rate of entry in knowledge-based sectors.

EMPIRICAL ANALYSIS

Our study investigates whether there is a significant relationship between new firm start-up rates and subsequent employment growth at the regional level using longitudinal data for Portuguese regions. Following Fritsch and Mueller (2004, 2008), we look at the lag structure of these effects and at the total effect over time. Based on the discussion in the previous section, as well as on the works by Fritsch and Mueller (2008), and Acs and Mueller (2008), two main hypotheses are tested:

**Hypothesis 1:** Start-ups will have stronger impacts on subsequent employment change in regions with relatively high levels of agglomeration.

**Hypothesis 2:** Start-ups of knowledge-based enterprises will have a greater impact on subsequent employment change than start-ups of other firms, regardless of the regional level of agglomeration.

Data and measurement issues

Data on entry and employment come from the longitudinal matched employer-employee micro-data set Quadros de Pessoal (QP), which was built based on information gathered in annual mandatory surveys by the Portuguese Ministry of Labor.

---

8 This classification includes high technology and medium-high technology industries, post and communications, finance and insurance and business services (OECD, 2002).
and Social Solidarity. These surveys cover all business units with at least one wage-earner in the Portuguese economy.9

Following Fritsch and Mueller (2004) and Baptista, Escária and Madruga (2008) we use as indicator of regional development the relative change over a two-year period of employment in the private sector. By using changes over a two-year period we attempt to avoid disturbances due to short-run fluctuations. The specific form in which the data set was built enables us to distinguish between true start-ups and entry of new plants/business units. New firm formation is measured by yearly regional start-up rates. Start-ups were assigned to the 30 standardized (NUTS 3) regions of Portugal for the period 1983-2000. Start-ups in the agricultural sector are excluded.

In order to control for differences in regional size, regional entry levels are not gauged in absolute terms, but as rates measured relative to regional size (Ashcroft, Love and Malloy, 1991). Following Garofoli (1994), and Audretsch and Fritsch (1994), regional start-up rates are measured using the size of the regional workforce as denominator (“labor market” approach). This methodology has advantages over the use of the total number of firms in a region as the denominator (“business stock” approach) since the latter might be misleading in regions with a few large firms (in such case, small numbers of new firms would provide an artificially high birth rate, primarily because of the small denominator).

Knowledge-based start-ups, regional agglomeration, and economic performance

In our analysis, we follow the OECD classification of knowledge-based sectors, aggregated by technology level (OECD 2002). This is a very wide definition of technology-intensive sectors, encompassing high and medium technology industries, as well as knowledge-based services. Using this wide definition of technology intensity provides a more adequate measure of the business dynamics in a region than merely including entry in high tech industries, as these firms represent a very small share of the Portuguese industrial structure, and are therefore unlikely to impact significantly on regional innovation and efficiency levels.

9 The database is property of the Portuguese government and can be accessed on-site at the Observatory of the Ministry of Labor and Social Solidarity.
In order to measure regional agglomeration effects and business dynamics, we follow Fritsch and Mueller (2008). These authors use population density as a proxy for the level of agglomeration externalities in German planning regions, classifying these into “highly agglomerated;” “moderately congested;” and “rural.” The classification by Fritsch and Mueller is linked with the perspective taken by Acs and Mueller (2008), who look at levels of business dynamics in American metropolitan regions by measuring the shares of fast growing and slow growing firms. These authors revisit one of the main insights of David Birch’s (1981) seminal contribution about the role played by small firms in employment creation: the perception that a small number of rapidly growing establishments (so-called “gazelles”) are responsible for most of the employment growth in regional economies. These authors find that some regions – the more agglomerated metropolitan areas – have most of the rapidly growing companies. By contrast, less agglomerated regions have a predominance of slow growing companies.

Table A1 in the Appendix to this paper shows that, in the particular case of Portugal, it is easy to identify highly agglomerated regions as the ones generating greater entry. Only the large metropolitan regions of Greater Lisbon and Greater Oporto (NUTS III codes 10104 and 10302 in Table A1) display agglomeration levels that are susceptible of ranking in the German “highly agglomerated” group defined by Fritsch and Mueller (2008). Moreover, these two regions are also the ones that display higher entry rates. Hence, we start by classifying the 30 Portuguese NUTS 3 regions into two groups:

i. Highly agglomerated regions, corresponding to the metropolitan regions of Greater Lisbon and Greater Oporto, which are highly agglomerated and display high proportions of rapidly growing new firms (i.e. high levels of business dynamics);

ii. Modestly agglomerated regions, corresponding to all other 28 NUTS regions, which display below average levels of agglomeration and business dynamics.

The QP database allows us to use start-up and incumbent sizes to determine the proportions of rapidly growing start-ups per region. In order to compute regional population density, additional data on NUTS regions was gathered from the Portuguese National Institute of Statistics (INE).
In order to look closer at the business dynamics of modestly agglomerated regions – i.e. the ones displaying relatively low levels of agglomeration and start-up rates, and lower than average proportions of rapidly growing firms, we again follow Fritsch and Mueller (2008) and look at differences in labor productivity, as measured by GDP per working population. When drawing a distinction between regions according to their economic performance (i.e. labor productivity), these authors find that the differences between the effects of new business formation on employment are much more pronounced between higher and lower productivity regions than when regions are differentiated on the basis of agglomeration only. However, Fritsch and Schroeter (see this Special Issue) find that the main determinant of this effect is population density rather than labor productivity. Many of the German high productivity regions have high levels of population density while most of the low productivity regions are rural areas.

Figure A1 in the Appendix displays the Portuguese modestly agglomerated regions by labor productivity levels. By examining the differences in the effects of new business formation on employment between regions with high and low labor productivity, we observe whether entries in regions with relatively high levels of economic performance (as measured by labor productivity) have a greater impact on subsequent employment change than entries in regions with low economic performance. While regions with high economic performance may be dominated by efficient incumbents in mature industries which have relatively highly qualified employees (from which the founders of new firms are likely to be drawn), regions with low economic performance are likely to be dominated by less efficient incumbents, employing less qualified human capital. Firms founded by less qualified human capital are likely to have a lower impact on the business dynamics of a region and, therefore, a lower impact on subsequent employment growth, whether through enhanced efficiency or through amplified innovation.

**Econometric Methodology**

The basic relationship to be modeled is adapted from Baptista, Escária and Madruga. (2008), where the change in regional employment between period $t-2$ and period $t$ is explained by the firm birth rates in periods $t, t-1, t-2\ldots n$, and has the following form:
\[ \Delta \text{EMP}_{t,r} = [\alpha_0^I \cdot \text{BIR}_{t,r}^I + \alpha_1^I \cdot \text{BIR}_{t-1,r}^I + \ldots + \alpha_n^I \cdot \text{BIR}_{t-n,r}^I] + [\alpha_0^II \cdot \text{BIR}_{t,r}^{II} + \alpha_1^II \cdot \text{BIR}_{t-1,r}^{II} + \ldots + \alpha_n^{II} \cdot \text{BIR}_{t-n,r}^{II}] + X_{t,r} \beta + \epsilon_i \] (1)

where \( \Delta \text{EMP}_{t,r} \) is the change in regional employment between period \( t-2 \) and period \( t \) for region \( r \); \( \text{BIR}_{t-i,r}^I, \text{BIR}_{t-i,r}^{II} \) are the firm birth rates in period \( t-i \) for type I and type II start-ups (e.g.: type I—knowledge-based firms; type II—other firms), with \( i=0,\ldots,n \) being the lag periods considered for region \( r \); and \( X_{t,r} \) are the control variables. For the present study, yearly start-up rates at the beginning of the current employment change period and for the ten preceding years are included.

However, an additional problem arises due to the significance of path dependency of regional new firm formation over time. We find that there is persistency of new firm formation over time at the regional level. The start-up rate in period \( t \) is significantly correlated with the start-up rate in the previous year and is also significantly determined by new firm formation activity five, ten and 15 years previously. The initial strong pattern of path dependency weakens over time. Almost 50% of the variation of the start-up rate in \( t \) can be explained by new firm formation activity 1, 5, 10 and 15 years previously.\(^{10}\) This means that correlations between start-up rates over time are mostly significant, leading to multicollinearity that makes interpretation of coefficients in the models difficult. In order to deal with this problem, the lag structures for the effect of regional start-up rates on regional employment growth are estimated using Almon polynomials (see Trivedi 1978; and Van Stel and Storey 2004, for details). The Almon lag procedure reduces the effects of multicollinearity in distributed lag settings by imposing a particular structure on the lag coefficients. In the Almon method, parameter restrictions are imposed in such way that the coefficients of the lagged variables are a polynomial function of the lag, producing a more compact model that overcomes the problems of multicollinearity.

When estimating the effects of start-up rates in different types of sectors (knowledge-based and others), we are introducing additional sources of bias, as residuals become correlated over time and heteroskedasticity becomes more significant. While studies such as Fritsch and Mueller (2004) and Baptista, Escária and Madruga (2008) used Huber-White robust estimators (i.e. panel corrected standard errors), under the new circumstances, Feasible Generalized Least Squares (FGLS) may be more appropriate as

\(^{10}\) Tables containing the correlations of new firm entry over time are omitted due to space limitations, but are available from the authors upon request.
these estimators handle autocorrelation and heteroskedasticity simultaneously (Parks 1967). However, in order to obtain unbiased estimations using FGLS the total number of time observations must be at least as large as the total number of panels (Beck and Katz 1995). In this case we have a panel of 18 years with 30 regions, which may lead to biased estimations. We therefore follow the methodology employed in previous papers, so the results from the Huber-White robust estimators are presented and discussed in Section 4. The results from FGLS estimation are quite similar, suggesting that the Huber-White estimations are robust.\footnote{The estimation results and corresponding figures for FGLS are omitted from due to space constraints but can be made available by the authors upon request.}

A shift-share procedure\footnote{The relative importance of incumbents and start-ups varies systematically across both regions and industries. For example, start-up rates are systematically higher in services than in manufacturing. Entrepreneurial activity could be systematically overestimated in regions with a high share of industries where start-ups play an important role, while the role of new firm formation in regions with a high share of industries where start-ups are relatively few would be underestimated. Following previous studies – pointed out in Section 1 – in order to account for differences in regional industrial structures, and in the relative importance of start-ups and incumbents across industries a shift-share procedure (Ashcroft, Love and Malloy 1991; Audretsch and Fritsch 2002) is applied in order to derive a measure of sector-adjusted start-up activity. The shift-share measure adjusts the raw data by imposing the same industry composition in each region (See Baptista, Escária and Madruga 2008 for a detailed explanation). Thus, the sector-adjusted number of start-ups is defined as the number of new firms in a region that can be expected to be observed if the composition of industries was identical across all regions.} is used to account for regional differences in industrial structure. Estimation of region-specific fixed effects is expected to capture regional asymmetries including differences in local labor market conditions, house prices and the extent of knowledge/innovation spillovers, as well as different cultural attitudes towards entrepreneurship – regions may differ in how they favor entrepreneurial activity and how they react to business failure. Also, two control variables are included in estimation, namely population density and average size of the firms. The objective of incorporating population density (number of inhabitants per square km) in our models is to control for regional characteristics which might affect the relationships between new firm formation and employment change. Fritsch and Mueller (2008) argue that regional population density is highly correlated with a number of factors such as the wage level, real estate prices, quality of communication infrastructure, qualification of the workforce, and diversity of the labor market. By incorporating the regional average firm size we are controlling for regional market structure, and intensity of regional competition.

Model estimations also correct for spatial autocorrelation. Following Anselin (1988), and Anselin and Florax (1995), the average of the residuals in adjacent regions is
included in the estimation. These residuals provide an indication of unobserved influences that affect larger geographical entities than NUTS3 and that are not entirely reflected in the explanatory variables.\textsuperscript{13}

**RESULTS**

Results are presented in Tables 1 to 6 and Figures 1 to 6. We begin by examining differences in the impact of new firm formation on subsequent employment growth between the highly agglomerated regions – Greater Lisbon and Greater Oporto – and other regions, as displayed by Tables 1 to 3 and Figures 1 to 3. Table 1 and Figure 1 present the effects of the total start-up rate on subsequent employment change in highly agglomerated vs. other regions. Table 2 and Figure 2 display the results for the same two kinds of regions when only knowledge-based firms are considered. Table 3 and Figure 3 present the results when the effects of the start-up rates for knowledge-based firms and other firms are estimated simultaneously for each type of region.

Results are presented for the unrestricted and restricted (Almon polynomial lag) models. Estimation of the Almon polynomial lag model assumes that the effect of changes in yearly start-up rates is distributed over eleven periods (\(t\) to \(t-10\)). Almon lag models were estimated for the second through to the fifth order of the polynomial. A critical issue in applying the Almon lag procedure is determining which order of polynomial to choose.\textsuperscript{14}

As can be seen in Figures 1-3, the patterns of effects of new business formation on employment change are different for highly agglomerated and modestly agglomerated regions. New firm formation in highly agglomerated regions initially has a negative effect, suggesting that displacement effects occur rapidly upon entry of new firms. The direct effect of new business formation in other regions is generally positive. In highly agglomerated regions, however, positive effects of new firm creation on employment

\textsuperscript{13} Estimations showed spatial autocorrelation to be insignificant, therefore not affecting the coefficients for the other variables. To correct for this, following Fritsch and Mueller (2004), we compute for each region the average of the residuals in the neighbouring regions and include this variable as an explanatory variable in the model.

\textsuperscript{14} An appropriate way to do this is to use Likelihood Ratio tests. Comparing the Nth order Almon polynomial model with the \((N+1)\textsuperscript{th}\) order Almon polynomial model comes down to a Likelihood Ratio test with one restriction, since each additional order of the polynomial adds one restriction to the model. In the present case, we find that the 4th order polynomial provides the best fit for the lag structure of the effects of new firm formation on regional employment change in each of the cases under analysis, so we present the estimation results for that model.
change become dominant after the second year, and their magnitude is higher than that of positive indirect effects in other regions. In both cases, the effect tails off from the sixth year onwards. This means that hypothesis 1 of our study is confirmed.

The pattern of effects when only knowledge-based firms are considered is somewhat different. Knowledge-based firms seem to play a more significant role in business dynamics and displacement effect than other firms, since negative selection effects occur both in highly and modestly agglomerated regions, and the decline in total employment goes on until the third period. Only after that do positive spillovers become dominant. These effects are much stronger for highly agglomerated regions, leading to a clearly positive overall effect on total employment. Hence, hypothesis 2 of our study is confirmed in particular for highly agglomerated regions.

When the effects of the start-up rates for knowledge-based firms and other firms are estimated simultaneously for highly agglomerated and modestly agglomerated regions (Table 3 and Figure 3) the pattern of results suggests that the type of start-up (knowledge-based versus other firms) plays a more important role in stimulating displacement and indirect positive spillovers than the type of region. Knowledge based start-ups have an initial negative effect on employment change, followed by significant, positive indirect effects occurring from the fourth (highly agglomerated regions) and fifth (modestly agglomerated regions) periods onwards. The overall effect of knowledge based start-ups seems to be clearly positive regardless of the regional degree of agglomeration while the overall effects of other types of start-ups appear not to be significant. The patterns of the effects of knowledge based start-ups on employment growth estimated simultaneously with the effects of other types of startups (depicted in Figure 3) appear to have a fairly similar shape to those obtained from estimation of their single effect across regions (depicted in Figure 2). The main difference seems to be the downward tail of the effect of knowledge-based start-ups on employment change in agglomerated regions occurring from about year nine, suggesting that the more intense business dynamics observed in these regions may lead to shorter-lived effects of these start-ups.

Despite this irregularity, which occurs only some nine years after start-up, it seems clear that Hypothesis 2 (differences in type of start-up) plays a more important role in explaining variations on the impact of new business formation on subsequent employment growth than Hypothesis 1 (differences in the type of region). In fact, while
the overall effect of start-up rates on employment growth appears to be clearly greater in highly agglomerated regions than in other regions when knowledge-based start-ups are concerned, the same conclusion cannot be clearly drawn for other types of start-ups. This suggests that the creation of knowledge-based firms imparts greater positive indirect effects on employment change in regions with high levels of agglomeration and business dynamics. The same is not clearly true for start-ups that are not knowledge-based. This is possibly due to the fact that start-ups in these sectors are likely to be less innovative, so other firms have less to gain from spillovers.

In order to shed further light on the nature of regional dynamics, we look more closely at the economic performance of modestly agglomerated regions, differentiating between those with relatively high labor productivity (upper quartile) and those with relatively low labor productivity (lower quartile), as can be seen in Figure A1. Table 4 and Figure 4 present the effect of the total start-up rate on subsequent employment growth in high labor productivity and low labor productivity regions. Table 5 and Figure 5 display the results for the same two kinds of regions when only knowledge-based firms are considered. Table 6 and Figure 6 present the results when the effects of the start-up rates for knowledge-based firms and other firms are estimated simultaneously for high labor productivity and low labor productivity regions.

While differences in the effect of total start-ups on subsequent employment change between higher and lower labor productivity regions are not large, the pattern of effects again appears to suggest that more dynamic (i.e. productive) regions experience stronger displacement effects and stronger subsequent positive supply-side effects (thus confirming hypothesis 1). The difference becomes clearer when we focus our analysis exclusively on knowledge-based start-ups. Comparing Figure 5 with Figure 4 we see that when only knowledge-based start-ups are considered, displacement effects dominate overall employment change immediately after entry in higher productivity regions, while positive indirect effects become dominant after only two years. When all start-ups are considered, positive indirect effects occur after three years for higher productivity regions. In any case, the overall effect of knowledge-based start-ups on subsequent employment growth is clearly greater than that of other types of start-ups for both kinds of regions (thus confirming hypotheses 2).

Table 6 and Figure 6 present the effects of the start-up rates for knowledge-based start-ups and other firms for high labor productivity and low labor productivity regions.
These results confirm our previous observation that the type of start-up plays a more important role than the type of region when determining the effects of new business formation on subsequent employment growth. In both higher and lower labor productivity regions the negative displacement effects and positive indirect spillover effects are of greater magnitude for knowledge-based start-ups than for other types of start-ups. The pattern of the effects requires some interpretation, however. In lower productivity regions, the negative selection effect that originates from the entry of knowledge-based start-ups is very strong indeed, and goes on until after the fifth year after entry. This is possibly due to the fact that new firms represent a significant efficiency improvement over existing firms in low economic performance regions, and their entry brings about the displacement of incumbents and the concomitant increase in unemployment. Selection effects brought about by knowledge-based entrants in higher labor productivity regions are of lower magnitude, and occur earlier, after the third year.

The simultaneous estimation of the impact of the formation of knowledge-based firms and other firms on regional employment change in both highly agglomerated and modestly agglomerated regions reveals a pattern of dynamics (depicted in Figure 6) which is to some extent different from what is obtained from the estimations for the total start-up rate as well as for the knowledge based firms start-up rate across regions (presented in Figures 4 and 5). In particular, the curve corresponding to the effect of knowledge based start-ups on highly agglomerated regions shows a somewhat unlikely upper tail from year 9 onwards. This pattern may be caused by the simultaneous estimation of the effects of both entry rates for knowledge-based and other industries (which add up to the total start-up rate), creating some sort of autocorrelation bias. However estimations using FGLS – supposed to correct for AR(1) autocorrelation – result in a similar pattern of results, so we keep the robust Huber-White estimators as the reference for our discussion.

Table 7 presents the sum of coefficients of the 11 periods under analysis for all models estimated, taken as an approximation of the overall effect of new firm formation on subsequent employment growth (following the approach taken by Fritsch and Mueller 2008). The sums of the regression coefficients for both the unrestricted models and the fourth order Almon polynomial lag models confirm that the type of start-up plays a more important role in explaining differences in competitive selection and
indirect spillover effects across regions than differences in agglomeration and economic performance between regions.

Focusing on the results of the Almon lag estimation, it is possible to observe on Table 7 that knowledge-based start-up rates have an overall positive effect on employment growth in the years after entry in both highly agglomerated and modestly agglomerated regions. When modestly agglomerated regions are divided according to economic performance, the overall positive effect of knowledge-based entry is positive in high economic performance (i.e. high labor productivity) regions, but negative in low labor productivity regions, where a very strong selection/increased competition effect is not completely compensated by the subsequent positive indirect effect.

An important observation that can be made from statistically significant coefficients of Table 7 is that the increases in entry rates for start-ups that are not knowledge-based have very small (or even negative) overall effects on subsequent employment change, and these effects do not change significantly according to the type of region. While it is true that the type of start-up (knowledge-based versus others) matters more than the type of region, it is also true that the levels of agglomeration and labor productivity in regions matter more for the effects of knowledge-based start-ups than for the effects of other start-ups.

CONCLUDING REMARKS

This study examined differences in the effects of start-up rates on subsequent employment change across regions. In particular, two sources of such differences – types of start-ups and types of regions – were analyzed, leading to two main hypotheses. Firstly, the impact of increases in start-up rates on subsequent employment change will be greater in regions with higher levels of agglomeration and business dynamics; secondly, increases in the start-up rates of knowledge-based firms will have a greater impact on subsequent employment change than increases in the start-up rates of other firms regardless of the type of region where these start-ups occur.

We find that differences between types of start-ups – namely between knowledge based and other firms – dominate differences in regional agglomeration and economic performance (as measured by labor productivity). Knowledge-based start-ups in high
business dynamics regions have essentially two effects on subsequent employment change:

i. First, a displacement (selection) effect (which occurs right from entry), likely brought about by increased competition and efficiency gains, leading to the exit of firms and a negative impact on employment;

ii. Second, an indirect, positive spillover effect, likely brought about by amplified innovation, increased efficiency and greater product variety, leading to increases in employment.

Start-ups in knowledge-based sectors have greater effects on subsequent employment growth than other start-ups, regardless of the type of region where these start-ups occur. This result suggests that knowledge-based start-ups have a greater potential to induce change in markets, bringing about both negative selection effects and positive spillover effects on overall employment.

Regional business dynamics, as measured by agglomeration levels and by labor productivity also matter, however. Differences in the effects of new start-ups on subsequent employment growth between more agglomerated, higher firm growth regions and less agglomerated, lower firm growth regions are greater for knowledge-based start-ups than for other types of start-ups. A particularly interesting result is obtained when modestly agglomerated regions are examined according to their levels of economic performance, as measured by labor productivity, the overall positive effect of knowledge-based entry is positive in high labor productivity regions, but negative in low labor productivity regions, where a very strong selection and increased efficiency effect offsets the subsequent positive indirect effect.

The results suggest that, while knowledge-based start-ups (which almost certainly include those more likely to be innovative and have a greater potential for high growth) are likely to impart greater overall benefits on employment than other types of start-ups (likely including the less innovative, low growth ones), these benefits are significantly larger when those start-ups locate in stronger, more dynamic (high agglomeration, high labor productivity) regions. The effects of other types of (non-innovative) start-ups on subsequent employment growth do not change significantly with the type of region where they locate.
Further research should concentrate on other sources of differences between types of start-ups, in order to better ascertain which types of start-ups have a greater impact on subsequent employment growth. For instance, the literature finds that larger, better financed entrants are more likely to survive and grow (Geroski 1995). It is therefore possible that these types of start-ups will have a greater impact on subsequent employment growth than smaller ones. Other sources of differences that may be examined are associated with the innovative potential of start-ups, and include human capital (of both founders and employees) and direct foreign investment (usually associated with technology spillovers).

ACKNOWLEDGEMENTS

The authors thank the GEP (Gabinete de Estratégia e Planeamento from Ministério do Trabalho e Solidariedade Social) for providing the data used in this study.

REFERENCES


<table>
<thead>
<tr>
<th>Table 1: Impact of lagged start-up rates on regional employment growth by agglomeration/business dynamics levels – robust fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Highly agglomerated regions</strong></td>
</tr>
<tr>
<td>Start-up rate t</td>
</tr>
<tr>
<td>Unrestricted</td>
</tr>
<tr>
<td>Almon method (4th order polynomial)</td>
</tr>
<tr>
<td>$\alpha_0$</td>
</tr>
<tr>
<td>0.298</td>
</tr>
<tr>
<td>(0.73)</td>
</tr>
<tr>
<td>$\alpha_1$</td>
</tr>
<tr>
<td>0.404</td>
</tr>
<tr>
<td>[-3.48]</td>
</tr>
<tr>
<td>$\alpha_2$</td>
</tr>
<tr>
<td>0.388</td>
</tr>
<tr>
<td>[3.36]</td>
</tr>
<tr>
<td>$\alpha_3$</td>
</tr>
<tr>
<td>0.729</td>
</tr>
<tr>
<td>[-4.05]</td>
</tr>
<tr>
<td>$\alpha_4$</td>
</tr>
<tr>
<td>1.013</td>
</tr>
<tr>
<td>[4.69]</td>
</tr>
<tr>
<td>$\sum$ coefficients start-up rate t to t-10</td>
</tr>
<tr>
<td>3.316</td>
</tr>
<tr>
<td>2.797/4.254</td>
</tr>
<tr>
<td>Notes: Robust t statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *. $^+$ sum of coefficients excluding negative coefficients after third phase.</td>
</tr>
<tr>
<td>Knowledge-based firms</td>
</tr>
<tr>
<td>-----------------------</td>
</tr>
<tr>
<td><strong>Highly agglomerated regions</strong></td>
</tr>
<tr>
<td>Start-up rate t</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-5</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-6</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-7</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-8</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-9</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Start-up rate t-10</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>∑ coefficients start-up rate t to t-10</td>
</tr>
</tbody>
</table>

<p>| <strong>Notes:</strong> Robust t statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *. ** sum of coefficients excluding negative coefficients after third phase. |</p>
<table>
<thead>
<tr>
<th></th>
<th>Unrestricted</th>
<th>Almon method (4th order polynomial)</th>
<th>Unrestricted</th>
<th>Almon method (4th order polynomial)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Highly agglomerated regions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge-based firms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up rate t</td>
<td>25.272***</td>
<td>α₀  3.852</td>
<td>3.852</td>
<td>Other firms</td>
</tr>
<tr>
<td></td>
<td>[3.36]</td>
<td>[0.90]</td>
<td></td>
<td>Start-up rate t</td>
</tr>
<tr>
<td></td>
<td>-13.384***</td>
<td>α₁  -1.109</td>
<td>2.152</td>
<td>-3.883***</td>
</tr>
<tr>
<td></td>
<td>[-3.03]</td>
<td>[-0.14]</td>
<td></td>
<td>[α₁] -1.341</td>
</tr>
<tr>
<td></td>
<td>5.8</td>
<td>α₂  -1.027</td>
<td>0.698</td>
<td>-0.139</td>
</tr>
<tr>
<td></td>
<td>[1.40]</td>
<td>[-0.25]</td>
<td></td>
<td>[-0.20]</td>
</tr>
<tr>
<td></td>
<td>-8.394***</td>
<td>α₃  0.475</td>
<td>0.929</td>
<td>1.307***</td>
</tr>
<tr>
<td></td>
<td>[-2.74]</td>
<td>[0.77]</td>
<td></td>
<td>[α₃] 1.134***</td>
</tr>
<tr>
<td></td>
<td>14.788***</td>
<td>α₄  -0.039</td>
<td>3.343</td>
<td>-0.225</td>
</tr>
<tr>
<td></td>
<td>[3.75]</td>
<td>[-1.38]</td>
<td></td>
<td>[-0.86]</td>
</tr>
<tr>
<td></td>
<td>-2.175</td>
<td></td>
<td>7.499</td>
<td>1.245***</td>
</tr>
<tr>
<td></td>
<td>[-0.41]</td>
<td></td>
<td></td>
<td>[2.76]</td>
</tr>
<tr>
<td></td>
<td>-20.887***</td>
<td></td>
<td>12.017</td>
<td>2.601***</td>
</tr>
<tr>
<td></td>
<td>[-6.97]</td>
<td></td>
<td></td>
<td>[5.56]</td>
</tr>
<tr>
<td></td>
<td>46.017***</td>
<td></td>
<td>14.575</td>
<td>-4.261***</td>
</tr>
<tr>
<td></td>
<td>[4.08]</td>
<td></td>
<td></td>
<td>[-5.40]</td>
</tr>
<tr>
<td></td>
<td>4.191</td>
<td></td>
<td>11.914</td>
<td>-1.501*</td>
</tr>
<tr>
<td></td>
<td>[0.42]</td>
<td></td>
<td></td>
<td>[-2.00]</td>
</tr>
<tr>
<td></td>
<td>6.295</td>
<td></td>
<td>-0.168</td>
<td>-1.668**</td>
</tr>
<tr>
<td></td>
<td>[0.84]</td>
<td></td>
<td></td>
<td>[-2.32]</td>
</tr>
<tr>
<td></td>
<td>-24.508***</td>
<td></td>
<td>-26.808</td>
<td>0.546</td>
</tr>
<tr>
<td></td>
<td>[-4.51]</td>
<td></td>
<td></td>
<td>[0.98]</td>
</tr>
<tr>
<td></td>
<td>33.015</td>
<td></td>
<td>30.002/56.978*</td>
<td></td>
</tr>
<tr>
<td><strong>Modestly agglomerated regions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge-based firms</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up rate t</td>
<td>5.452**</td>
<td>α₀  2.459</td>
<td>2.459</td>
<td>Other firms</td>
</tr>
<tr>
<td></td>
<td>[2.20]</td>
<td>[0.95]</td>
<td></td>
<td>Start-up rate t</td>
</tr>
<tr>
<td></td>
<td>-5.602</td>
<td>α₁  -3.412</td>
<td>-1.328</td>
<td>-0.966**</td>
</tr>
<tr>
<td></td>
<td>[-1.30]</td>
<td>[-0.54]</td>
<td></td>
<td>[α₁] 1.230***</td>
</tr>
<tr>
<td></td>
<td>-4.988</td>
<td>α₂  -0.686</td>
<td>-4.794</td>
<td>0.508***</td>
</tr>
<tr>
<td></td>
<td>[-1.28]</td>
<td>[-0.18]</td>
<td></td>
<td>[α₂] 0.582*</td>
</tr>
</tbody>
</table>

Table 3: Impact of lagged knowledge-based and other firms’ start-up rates on regional employment growth by agglomeration/business dynamics levels – robust fixed effects.
| Start-up rate t-3 | -5.35 | \( \alpha_3 \) | 0.333 | -6.731 | Start-up rate t-3 | 0.691** | \( \alpha_3 \) | -0.103** | 0.875 |
| Start-up rate t-4 | -3.107 | \( \alpha_4 \) | -0.022 | -6.458 | Start-up rate t-4 | 1.020*** | \( \alpha_4 \) | 0.005** | 0.985 |
| Start-up rate t-5 | -7.079 | \( \alpha_5 \) | 0.333 | -6.731 | Start-up rate t-5 | 1.392*** | \( \alpha_5 \) | 0.103** | 0.875 |
| Start-up rate t-6 | -5.928 | \( \alpha_6 \) | 0.333 | -6.731 | Start-up rate t-6 | 0.264 | \( \alpha_6 \) | 0.103** | 0.875 |
| Start-up rate t-7 | 12.510*** | \( \alpha_7 \) | 0.333 | -6.731 | Start-up rate t-7 | -1.160*** | \( \alpha_7 \) | -0.103** | 0.875 |
| Start-up rate t-8 | 15.975** | \( \alpha_8 \) | 0.333 | -6.731 | Start-up rate t-8 | -1.160*** | \( \alpha_8 \) | -0.103** | 0.875 |
| Start-up rate t-9 | 14.045*** | \( \alpha_9 \) | 0.333 | -6.731 | Start-up rate t-9 | -1.160*** | \( \alpha_9 \) | -0.103** | 0.875 |
| Start-up rate t-10 | 12.137* | \( \alpha_{10} \) | 0.333 | -6.731 | Start-up rate t-10 | -2.118*** | \( \alpha_{10} \) | -0.103** | 0.875 |
| \( \sum \) coefficients start-up rate t to t-10 | 28.065 | \( \sum \) | coefficients start-up rate t to t-10 | 26.781 | \( \sum \) | coefficients start-up rate t to t-10 | 0.431 | \( \sum \) | coefficients start-up rate t to t-10 | 0.331 |

Firm size | 0.782*** | \( \gamma_{11} \) | 0.751*** |
Population density | 0.038 | \( \gamma_{12} \) | 0.032 |
Constant | -20.974** | \( \gamma_{13} \) | -17.552* |
Log-likelihood | -1528.78 | \( \gamma_{14} \) | -1545.71 |
Adjusted R-squared | 0.3007 | \( \gamma_{15} \) | 0.2895 |
No. of observations | 510 | \( \gamma_{16} \) | 510 |

Notes: Robust \( t \) statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *. \( \sum \) sum of coefficients excluding negative coefficients after third phase.
Table 4: Impact of lagged start-up rates on regional employment growth by labor productivity – robust fixed effects

<table>
<thead>
<tr>
<th>High labor productivity regions</th>
<th>Unrestricted</th>
<th>Almon method (4th order polynomial)</th>
<th>Low labor productivity regions</th>
<th>Unrestricted</th>
<th>Almon method (4th order polynomial)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start-up rate t</td>
<td>0.540</td>
<td>$\alpha_0$ 0.549</td>
<td>Start-up rate t</td>
<td>0.644**</td>
<td>$\alpha_0$ 0.797*** 0.797</td>
</tr>
<tr>
<td></td>
<td>[1.26]</td>
<td>[1.38]</td>
<td></td>
<td>[2.39]</td>
<td>[2.88]</td>
</tr>
<tr>
<td>Start-up rate t-1</td>
<td>0.198</td>
<td>$\alpha_1$ -1.038</td>
<td>Start-up rate t-1</td>
<td>-0.291</td>
<td>$\alpha_1$ -2.564*** -0.730</td>
</tr>
<tr>
<td></td>
<td>[0.81]</td>
<td>[-1.54]</td>
<td></td>
<td>[-0.47]</td>
<td>[-3.46]</td>
</tr>
<tr>
<td>Start-up rate t-2</td>
<td>-0.330</td>
<td>$\alpha_2$ 0.527*</td>
<td>Start-up rate t-2</td>
<td>-1.140**</td>
<td>$\alpha_2$ 1.209*** -0.805</td>
</tr>
<tr>
<td></td>
<td>[-1.11]</td>
<td>[1.72]</td>
<td></td>
<td>[-2.55]</td>
<td>[3.65]</td>
</tr>
<tr>
<td>Start-up rate t-3</td>
<td>0.131</td>
<td>$\alpha_3$ -0.083*</td>
<td>Start-up rate t-3</td>
<td>-0.289</td>
<td>$\alpha_3$ -0.180*** -0.209</td>
</tr>
<tr>
<td></td>
<td>[0.37]</td>
<td>[-1.82]</td>
<td></td>
<td>[-0.63]</td>
<td>[-3.89]</td>
</tr>
<tr>
<td>Start-up rate t-4</td>
<td>0.916***</td>
<td>$\alpha_4$ 0.004*</td>
<td>Start-up rate t-4</td>
<td>0.767**</td>
<td>$\alpha_4$ 0.008*** 0.474</td>
</tr>
<tr>
<td></td>
<td>[3.33]</td>
<td>[1.87]</td>
<td></td>
<td>[2.19]</td>
<td>[4.05]</td>
</tr>
<tr>
<td>Start-up rate t-5</td>
<td>1.041***</td>
<td></td>
<td>Start-up rate t-5</td>
<td>0.891*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[3.03]</td>
<td></td>
<td></td>
<td>[2.03]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-6</td>
<td>0.240</td>
<td></td>
<td>Start-up rate t-6</td>
<td>0.835*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.47]</td>
<td></td>
<td></td>
<td>[1.72]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-7</td>
<td>-0.755*</td>
<td></td>
<td>Start-up rate t-7</td>
<td>-0.102</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-1.86]</td>
<td></td>
<td></td>
<td>[-0.21]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-8</td>
<td>0.037</td>
<td></td>
<td>Start-up rate t-8</td>
<td>-0.537</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.14]</td>
<td></td>
<td></td>
<td>[-1.63]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-9</td>
<td>-0.018</td>
<td></td>
<td>Start-up rate t-9</td>
<td>-0.896*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.04]</td>
<td></td>
<td></td>
<td>[-1.93]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-10</td>
<td>-1.469***</td>
<td></td>
<td>Start-up rate t-10</td>
<td>-1.021***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-4.01]</td>
<td></td>
<td></td>
<td>[-4.98]</td>
<td></td>
</tr>
<tr>
<td>$\sum$ coefficients start-up rate t to t-10</td>
<td>0.531</td>
<td>0.557/2.560*</td>
<td>$\sum$ coefficients start-up rate t to t-10</td>
<td>-1.139</td>
<td>-1.177/1.408*</td>
</tr>
</tbody>
</table>

Notes: Robust t statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *. $\sum$ sum of coefficients excluding negative coefficients after third phase.
### Table 5: Impact of lagged knowledge-based start-up rates on regional employment growth by labor productivity – robust fixed effects

<table>
<thead>
<tr>
<th></th>
<th>High labor productivity regions</th>
<th>Low labor productivity regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unrestricted (4th order polynomial)</td>
<td>Almon method (4th order polynomial)</td>
</tr>
<tr>
<td>Start-up rate t</td>
<td>6.058 [1.33] α₀ 1.958 [0.52] 1.958</td>
<td>Start-up rate t</td>
</tr>
<tr>
<td>Start-up rate t-3</td>
<td>6.741 [0.98] α₃ -2.031*** [-5.99] 2.815</td>
<td>Start-up rate t-3</td>
</tr>
<tr>
<td>Start-up rate t-4</td>
<td>16.278*** [4.38] α₄ 0.094*** [4.47] 11.892</td>
<td>Start-up rate t-4</td>
</tr>
<tr>
<td>Start-up rate t-5</td>
<td>15.005*** [2.84] 16.353</td>
<td>Start-up rate t-5</td>
</tr>
<tr>
<td>Start-up rate t-7</td>
<td>4.120 [-0.59] 5.621</td>
<td>Start-up rate t-7</td>
</tr>
<tr>
<td>Start-up rate t-8</td>
<td>0.583 [0.13] -6.825</td>
<td>Start-up rate t-8</td>
</tr>
<tr>
<td>Start-up rate t-9</td>
<td>-3.032 [0.34] -18.392</td>
<td>Start-up rate t-9</td>
</tr>
<tr>
<td>∑ coefficients start-up rate t to t-10</td>
<td>-13.985 -10.874/36.397⁺</td>
<td>∑ coefficients start-up rate t to t-10</td>
</tr>
</tbody>
</table>

Firm size
-0.051 [-0.25] -0.051 [-0.25]
Population density
-0.052 [-1.66] -0.052 [-1.67]
Constant
18.611** [2.25] 18.611** [2.27]
Log-likelihood
-1623.77 -1626.92
Adjusted R-squared
0.0311 0.0427
No. of observations
510 510

Notes: Robust t statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *.⁺ sum of coefficients excluding negative coefficients after third phase.
### Table 6: Impact of lagged knowledge-based and other firms’ start-up rates on regional employment growth by labor productivity – robust fixed effects

<table>
<thead>
<tr>
<th></th>
<th>Unrestricted Almon method (4th order polynomial)</th>
<th>Unrestricted Almon method (4th order polynomial)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High labor productivity regions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge-based firms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up rate t</td>
<td>0.751</td>
<td>Other firms</td>
</tr>
<tr>
<td></td>
<td>[0.13]</td>
<td>Start-up rate t</td>
</tr>
<tr>
<td></td>
<td>-6.583</td>
<td>-6.583</td>
</tr>
<tr>
<td></td>
<td>[-1.14]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-1</td>
<td>-20.340***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-4.70]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-8.711</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.58]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-2</td>
<td>-4.901</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.64]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7.773</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.01]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-3</td>
<td>10.440***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2.89]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-1.547</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-1.16]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-4</td>
<td>13.030**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[2.53]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.089</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[1.23]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-5</td>
<td>-3.883</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.74]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.302</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-6</td>
<td>-1.705</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.51]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.903</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-7</td>
<td>-2.204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.46]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4.094</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-8</td>
<td>5.713</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.44]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7.114</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-9</td>
<td>5.107</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.34]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.449</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-10</td>
<td>5.974</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>24.741</td>
<td></td>
</tr>
<tr>
<td>∑ coefficients start-up rate t to t-10</td>
<td>7.982</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low labor productivity regions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge-based firms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up rate t</td>
<td>6.141</td>
<td>Other firms</td>
</tr>
<tr>
<td></td>
<td>[1.53]</td>
<td>Start-up rate t</td>
</tr>
<tr>
<td></td>
<td>4.458</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.89]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-1</td>
<td>1.859</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.27]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.437</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.13]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-2</td>
<td>1.757</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[0.29]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-3.515</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-0.60]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-0.806</td>
<td></td>
</tr>
<tr>
<td>∑ coefficients start-up rate t to t-10</td>
<td>11.173</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-3</td>
<td>$a_3$</td>
<td>0.841</td>
</tr>
<tr>
<td>------------------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td></td>
<td>[-1.00]</td>
<td>[0.82]</td>
</tr>
<tr>
<td>Start-up rate t-4</td>
<td>$a_4$</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>[-0.65]</td>
<td>[-0.89]</td>
</tr>
<tr>
<td>Start-up rate t-5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>[-1.14]</td>
<td>[-0.89]</td>
</tr>
<tr>
<td>Start-up rate t-6</td>
<td>2.457</td>
<td>2.965</td>
</tr>
<tr>
<td></td>
<td>[0.35]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-7</td>
<td></td>
<td>9.906</td>
</tr>
<tr>
<td></td>
<td>[2.00]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-8</td>
<td>13.158</td>
<td>15.449</td>
</tr>
<tr>
<td></td>
<td>[1.03]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-9</td>
<td>9.248</td>
<td>15.586</td>
</tr>
<tr>
<td></td>
<td>[1.19]</td>
<td></td>
</tr>
<tr>
<td>Start-up rate t-10</td>
<td>12.231</td>
<td>5.096</td>
</tr>
<tr>
<td></td>
<td>[0.84]</td>
<td></td>
</tr>
<tr>
<td>$\sum$ coefficients start-up rate t to t-10</td>
<td>45.702</td>
<td>43.919</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Firm size: -0.025 [-0.12] -0.034 [-0.17]
Population density: -0.038 [-1.22] -0.039 [-1.30]
Log-likelihood: -1593.55 -1601.55
Adjusted R-squared: 0.0984 0.1156
No. of observations: 510 510

Notes: Robust t statistics in brackets. Significant at: 1%-level ***; 5%-level **; 10%-level *. $\sum$ sum of coefficients excluding negative coefficients after third phase.
Figure 1: Estimated lag structure (4th order Almon polynomial) of the impact of start-ups (all industries) on regional employment change in highly agglomerated regions and modestly agglomerated regions – robust fixed effects

Figure 2: Estimated lag structure (4th order Almon polynomial) of the impact of the formation of knowledge-based firms on regional employment change in highly agglomerated and modestly agglomerated regions – robust fixed effects
Figure 3: Estimated lag structure (4th order Almon polynomial) of the impact of the formation of knowledge-based firms and other firms on regional employment change in highly agglomerated and modestly agglomerated regions – robust fixed effects

Figure 4: Estimated lag structure (4th order Almon polynomial) of the impact of the formation of start-ups (all industries) on regional employment change in high- and low labor productivity regions – robust fixed effects
Figure 5: Estimated lag structure (4th order Almon polynomial) of the impact of the formation of knowledge-based firms on regional employment change in high- and low labor productivity regions – robust fixed effects

Figure 6: Estimated lag structure (4th order Almon polynomial) of the impact of the formation of knowledge-based firms and other firms on regional employment change in high- and low labor productivity regions – robust fixed effects
Table 7: Overall effect of new business formation on regional employment change over time

<table>
<thead>
<tr>
<th>Agglomeration/business dynamics levels</th>
<th>Sum of coefficients (in parentheses: without negative coefficients after phase III)</th>
<th>Feasible generalized least squares</th>
<th>Robust fixed effects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Unrestricted</td>
<td>Unrestricted</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Almon method (4\textsuperscript{th} order polynomial)</td>
<td>Almon method (4\textsuperscript{th} order polynomial)</td>
</tr>
<tr>
<td>Highly agglomerated regions (all industries)</td>
<td>-0.180 (1.442)</td>
<td>3.316</td>
<td>2.797 (4.254)</td>
</tr>
<tr>
<td>Modestly agglomerated regions (all industries)</td>
<td>0.435</td>
<td>0.256 (1.840)</td>
<td>1.698</td>
</tr>
<tr>
<td>Highly agglomerated regions (knowledge based industries)</td>
<td>-1.132 (14.224)</td>
<td>5.586</td>
<td>5.647 (24.468)</td>
</tr>
<tr>
<td>Modestly agglomerated regions (knowledge based industries)</td>
<td>-1.904 (17.966)</td>
<td>0.561</td>
<td>1.341 (26.335)</td>
</tr>
<tr>
<td>Highly agglomerated regions (knowledge based industries)</td>
<td>18.448</td>
<td>14.838 (31.924) (n.s)</td>
<td>33.015</td>
</tr>
<tr>
<td>Modestly agglomerated regions (knowledge based industries)</td>
<td>-1.579</td>
<td>-2.364</td>
<td>-3.144</td>
</tr>
<tr>
<td>Highly agglomerated regions (all industries)</td>
<td>15.948</td>
<td>15.542 (n.s)</td>
<td>28.065</td>
</tr>
<tr>
<td>Modestly agglomerated regions (all industries)</td>
<td>-0.238</td>
<td>-0.520</td>
<td>0.431</td>
</tr>
<tr>
<td>Labor productivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High labor productivity (all industries)</td>
<td>0.078</td>
<td>0.085 (1.359)</td>
<td>0.531</td>
</tr>
<tr>
<td>Low labor productivity (all industries)</td>
<td>-0.730</td>
<td>-0.721 (1.527)</td>
<td>-1.139</td>
</tr>
<tr>
<td>High labor productivity (knowledge based industries)</td>
<td>-2.197</td>
<td>-1.702 (26.632)</td>
<td>-13.985</td>
</tr>
<tr>
<td>Low labor productivity (knowledge based industries)</td>
<td>-2.829</td>
<td>-1.750 (7.713) (n.s)</td>
<td>-13.218</td>
</tr>
<tr>
<td>High labor productivity (knowledge based industries)</td>
<td>0.294</td>
<td>-6.085 (5.026)</td>
<td>7.982</td>
</tr>
<tr>
<td>Low labor productivity (knowledge based industries)</td>
<td>-5.552</td>
<td>-8.711</td>
<td>45.702</td>
</tr>
<tr>
<td>High labor productivity (other industries)</td>
<td>-0.122</td>
<td>0.639 (n.s)</td>
<td>0.140</td>
</tr>
<tr>
<td>Low labor productivity (other industries)</td>
<td>0.398</td>
<td>0.557</td>
<td>-3.172</td>
</tr>
</tbody>
</table>

Notes: n.s. = coefficients are not statistically significant at the 10% level
Appendix

*Table A1: Portuguese NUTS III regions – population density and start-up rates*

<table>
<thead>
<tr>
<th>NUTS III</th>
<th>Region Name</th>
<th>Population density (number of residents per square km)</th>
<th>Average share of start-ups 1983-2001 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10101</td>
<td>Minho-Lima</td>
<td>120</td>
<td>2.4</td>
</tr>
<tr>
<td>10102</td>
<td>Câvado</td>
<td>304</td>
<td>4.1</td>
</tr>
<tr>
<td>10103</td>
<td>Ave</td>
<td>397</td>
<td>5.3</td>
</tr>
<tr>
<td>10104</td>
<td>Greater Oporto</td>
<td>1529</td>
<td>12.1</td>
</tr>
<tr>
<td>10105</td>
<td>Tâmega</td>
<td>211</td>
<td>5.3</td>
</tr>
<tr>
<td>10106</td>
<td>Entre Douro e Vouga</td>
<td>312</td>
<td>3.1</td>
</tr>
<tr>
<td>10107</td>
<td>Douro</td>
<td>63</td>
<td>1.5</td>
</tr>
<tr>
<td>10108</td>
<td>Alto Trás-os-Montes</td>
<td>32</td>
<td>1.5</td>
</tr>
<tr>
<td>10201</td>
<td>Baixo Vouga</td>
<td>208</td>
<td>3.1</td>
</tr>
<tr>
<td>10202</td>
<td>Baixo Mondego</td>
<td>170</td>
<td>2.5</td>
</tr>
<tr>
<td>10203</td>
<td>Pinhal Litoral</td>
<td>138</td>
<td>3.0</td>
</tr>
<tr>
<td>10204</td>
<td>Pinhal Interior Norte</td>
<td>57</td>
<td>1.2</td>
</tr>
<tr>
<td>10205</td>
<td>Dão-Lafões</td>
<td>87</td>
<td>2.5</td>
</tr>
<tr>
<td>10206</td>
<td>Pinhal Interior Sul</td>
<td>29</td>
<td>0.4</td>
</tr>
<tr>
<td>10207</td>
<td>Serra da Estrela</td>
<td>66</td>
<td>0.3</td>
</tr>
<tr>
<td>10208</td>
<td>Beira Interior Sul</td>
<td>31</td>
<td>0.9</td>
</tr>
<tr>
<td>10209</td>
<td>Beira Interior Sul</td>
<td>23</td>
<td>0.7</td>
</tr>
<tr>
<td>10210</td>
<td>Cova da Beira</td>
<td>73</td>
<td>0.8</td>
</tr>
<tr>
<td>10301</td>
<td>Oeste</td>
<td>175</td>
<td>4.0</td>
</tr>
<tr>
<td>10302</td>
<td>Greater Lisbon</td>
<td>1466</td>
<td>19.4</td>
</tr>
<tr>
<td>10303</td>
<td>Península Setúbal</td>
<td>449</td>
<td>6.0</td>
</tr>
<tr>
<td>10304</td>
<td>Médio Tejo</td>
<td>105</td>
<td>2.2</td>
</tr>
<tr>
<td>10305</td>
<td>Lezíria do Tejo</td>
<td>58</td>
<td>2.3</td>
</tr>
<tr>
<td>10401</td>
<td>Alentejo Litoral</td>
<td>20</td>
<td>0.9</td>
</tr>
<tr>
<td>10402</td>
<td>Alto Alentejo</td>
<td>22</td>
<td>1.3</td>
</tr>
<tr>
<td>10403</td>
<td>Alentejo Central</td>
<td>25</td>
<td>2.0</td>
</tr>
<tr>
<td>10404</td>
<td>Baixo Alentejo</td>
<td>18</td>
<td>1.4</td>
</tr>
<tr>
<td>10501</td>
<td>Algarve</td>
<td>73</td>
<td>5.5</td>
</tr>
<tr>
<td>20101</td>
<td>R. A. Açores</td>
<td>110</td>
<td>2.1</td>
</tr>
<tr>
<td>30101</td>
<td>R. A. Madeira</td>
<td>321</td>
<td>2.0</td>
</tr>
</tbody>
</table>
Figure A1: Map of Portuguese NUTS III regions by labor productivity levels
Chapter 14:
Muriel Pádua, João Sousa, Hugo Horta and Manuel Heitor, "Managing uncertainty: on risk prevention of non-communicable diseases in the context of primary intervention among the elderly in vulnerable communities".
Managing uncertainty in urban contexts:
looking at risk prevention of non-communicable diseases in the context
of primary intervention among the elderly in vulnerable communities

Muriel Pádua, João Santos, Hugo Horta and Manuel V. Heitor

Center for Innovation, Technology and Policy Research, IN+
Instituto SuperiorTécnico, Technical University of Lisbon, Portugal

Abstract
This paper deals with uncertainty in urban contexts and shows that scientific literacy of the elderly, as
quantified in terms of formal education levels, impacts on the risk of getting diabetes. To our
knowledge, these results are new and are relevant to the design of actions to look at risk perceptions,
risk communication and stakeholder engagement of lay people from vulnerable communities.

The rational for the paper is driven by the fact that the rate at which diabetes increase worldwide is
gaining momentum and, as a result, it is becoming a key issue in national policies health agenda, with
emphasis on vulnerable communities. As a consequence, the paper examines potential links between
scientific literacy, as measured through formal education levels, and health outcomes in the context of
primary intervention with regard to diabetes. It should be noted that other health outcomes considered
in the literature are mortality rates, but not risk of getting the disease. However, formal education has
very little impact on preventive behavior such as the adoption of risk modifiable behaviors or health
screening behaviors. On the other hand, education is known to impact risk perception.

While it is acknowledged that scientific literacy (and, therefore, education) can play a role in reducing
risks associated with non-communicable diseases, namely diabetes (namely in terms of improving
health outcomes and through the adoption of risk modifiable behavior), we still know very little on this
issue with regard to the elderly population, for which the burden of this disease is the highest.

We carried out an observational cross sectional study and collected data on the awareness of the risks
associated with diabetes and related prevention habits of the elderly (such as adoption of risk behavior
and health screening habits). Demographic data including education levels, age and gender were
collected as well. A final sample of 356 subjects was gathered.

Our analysis shows that the relationship between risk perception and the adoption of modifiable risk
behaviors is complex, and it is sensitive to different risk conditions (being at risk or not). The results
suggest there are new needs for promoting science culture among the elderly with regard to increasing
awareness towards risk factors, which are not necessarily dependent on formal education. Moreover
the promotion of scientific culture might consider taking into account different boundary conditions
(i.e., being at risk or not).
1. Introduction
Urban development and the growth of cities and the activities they encompass impose strains over various dimensions of cities' life. Social cohesion and the adequate integration of the elderly is just one of the areas in which risks affecting cities at a systemic level can arise. Tackling those risks requires designing resilient cities, with emphasis on inclusiveness and the potential open access of the elderly and the most vulnerable communities to knowledge and knowledge networks.

It is under this broad context that this work was oriented towards assessing risk perceptions of the elderly in vulnerable communities and, in the process, to examine strategies of risk communication towards those groups. While early studies on risk perception were based on “unilateral” expert views (i.e., “methods of expert elicitation”), it has become more and more clear that the involvement of lay people in the process is critical for the governance of risk (i.e., to ensure their participation). But, and despite these new trends, vulnerable groups remain an outlier category of this type of analysis.

Only recently, these studies have started to address risk perceptions of vulnerable groups (e.g., exploring differences in terms of risk perception across genders, white/non white population). But, efforts in this direction remain rather limited. For example, it is quite surprising that the argument regarding the need to communicate risks regarding non-communicable diseases, such as diabetes, to vulnerable groups constitutes by itself a topic for a recent publication in Science by Ezzati and Riboli (1). However, the literature is scarce in providing recommendations how to communicate risks in such environments.

1.1 Building a theoretical background: a brief overview of sample theoretical approaches

Different approaches have been put forward to explore risk perceptions of lay people towards risk. We distinguish the biases approach from other perspectives centred on values and attitudes. While the former approach argues that, by uncovering people’s biases in processes of decision making, we can raise their awareness and change behaviour accordingly (2), other approaches argue that information per se does not help shaping behaviours. The biases approaches focuses on information based strategies of risk communication.

In these models it is considered that the debate has to be focused on the communication of uncertainty itself. In some ways, it is assumed that people are capable of handling uncertainty, provided they have the right information. Other perspectives consider this approach as limited. They argue that information is not good enough to shape people’s behaviour. For example, other approaches argue that to mitigate risk behaviour towards non-communicable diseases, one has to consider people’s values in life or to consider their attitude towards diabetes (3).

The biases approach explores, on one hand, the cognitive and emotional biases of lay people towards risk perception, and, on the other hand, their representation of risk (by contrast to expert’s views). These studies have shown that people assessment of risk do not conform to Bayesian rules of probabilistic decision-making. Instead, people systematically deviate from these standard norms of rationality. Their reasoning is biased and they use heuristics or rules
of thumb when taking decisions under risk. People’s cognitive biases include overconfidence, anchoring and availability (availability refers to the fact that they ignore initial bases rate due to the information that is available in their memories).

Moreover, it has been shown that decisions are emotionally overloaded (people’s attention tends to be focused on catastrophic events - people’s attention to cancer is much higher than other diseases that kill more). On the knowledge representation side, it has been shown that representations of risk and experts differ. While experts examine risk perception using statistical evidence, lay people use other components such as, dread, exposure and controllability (4).

The values approach is currently used in medicine to mitigate people’s behaviour in diabetes. Rather than asking people their perceptions on long-term risks, this methodology focuses on current values of people. The theory of planned behaviour argues that attitudes, perceived behaviour control and social pressure are the dimensions to be considered when looking at the need to change behaviours. While these models are useful, it is certain correct to assume that information is still important, especially for lay people from vulnerable communities. More likely, this people lack basic knowledge on diseases.

We take, as a starting point, the bounded rational hypothesis, but we argue that it needs to be stretched alongside a new dimension. We need to consider the social fabric in which behaviours take place, namely communities (behaviour is shaped by context interdependences). For example, analysis has shown that social cognition (e.g., people’s perceptions of risk) depends on their affiliation groups (5).

In other words, people with more conservative background think that risk does not need to be shared, while people from more equalitarian background think the opposite. The idea that people have cultural biases has also been shown to be correct (their perceptions towards risk depends on group identity and culture, an area called cultural cognition). We purpose this line of search and argue that we need to contextualize behaviour into their communities (6).

1.2 An initial experiment and hypothesis

Our initial hypothesis was driven by the idea of "indwelling", firstly introduced by Polanyi (7) and recently explored by John Seely Brown (8) in terms of understanding learning through processes of knowing, playing and making. We attempted to provide new evidence on related learning processes through one specific set of experiments, as follows.

Lay people has been encouraged to ride a "stand-alone" bicycle, facilitating physical exercise and making a smoothie (healthy fruit based drink) through the preparation of crushed fruit, yoghurt or milk. This begins a learning and revelatory process concerning adequate diet, exercise and successful ways of reducing their risk by having fun together. But, above all, it facilitates making people aware of diabetes risks and to introduce the debate on the topic among specific and target communities. Experiments were successfully conducted with the elderly and specific muslin groups in “Mouraria”, a vulnerable zone in old Lisbon.
Analysis has shown that a “smoothie bike” was indeed a device that disrupted people’s perception towards diabetes. It enabled people to talk about the problem of diabetes, generated considerable interaction between people, and produced considerable enthusiasm amongst lay people. Different communication channels were used to provide a trust context. The contexts included day centres for elderly, shopping malls, small restaurants, public spaces, among others.

We identified a limitation of this approach because knowledge of lay people in relation to diabetes was so low and full of misconceptions. In addition, potential associations between diabetes prevention and physical exercise, as promoted through bike usage, or adequate diet, as promoted by drinking fruit juices, was rather limited.

People had considerable misconceptions towards diabetes. For example, one misconception was to consider that one has a “little bit” of diabetes. Moreover, they did not understand the concept of chronic disease. They assume, that once their level of glycaemia decreases, as a result of the medication, they stopped having the disease (the misconception here concerns the absence of the concept of “chronic disease” in their mental models); all sorts of projection and denial mechanism were taking place (even the people that were diagnosed with diabetes denied they had the disease).

Overall, a number of interviews with targeted communities reveal that many misunderstand the risks they face, in particular regarding non-communicable diseases, such as diabetes. This lead to further analysis and the requirement to build a proper sample of elderly people.

1.3 Looking deeper at the elderly in vulnerable urban zones

The elderly, adults aged 65 years and older, represent the age group with the lowest scientific literacy or formal education levels (9,10). They also have the highest proportion of risk of chronic illness of any age-group. Change in behavioral habits is effective in reducing the risk of developing this pathology among older adults (11) but the elderly are known to resist change (12). In this context, they comprise a vulnerable population that is most likely to be negatively affected by limited education levels (13). But whether and how this relationship applies remains poorly understood in a primary intervention context regarding the elderly population. In this scope, understanding the relationship between risk perceptions and preventive behavior is particularly important because of the known knowledge gap in the context of primary intervention, particularly concerning non-communicable diseases (14). Research on this issue in a primary intervention context can be useful for the design of health prevention oriented programs.

It should be noted that significant differences in health status due to limited education among older adults are well documented in the context of secondary and tertiary intervention (15-22). Although scarce, new concerning evidence shows that limited education levels can increase the risk of poor health outcomes in the context of primary intervention (23). For example, Sabates and Feinstein (24) show that people with poor education levels use less preventive measures, leading to higher mortality rates. Yet, despite calls for a better
understanding of the relationship between education and poor health outcomes in the context of primary intervention, reckoning of a direct link is unclear as it is the variables that explain it (25-27).

In trying to identify this link, e.g., the role of education and health outcomes, previous studies examine the role of genetic factors, anthropometric data and life-style effect (28-31), but not the effect of scientific literacy or education in the risk of getting the disease. Perceiving the effect of education on the risk of getting the disease seems relevant since obesity and life-style are found to be risk factors associated with low social economic and educational background, and thus may contribute to increased risk of getting the disease (32). Yet, the link between education and risk to get diabetes remains unexplored, and it is focused in this study.

The variables that compose the link between education and health outcomes within a primary prevention context also require to be analyzed. In this context, the role of risk perception might be critical, but its relationship with preventive behaviors – as well as with other variables, such as education - remains unclear. The greater number of studies focused on secondary and tertiary intervention, have shown that changes in health outcomes are associated to a multifaceted set of factors such as knowledge, awareness of risk, costs (33,34), and that possibly the link that connects education with preventive behavior is more complex and nonlinear than expected (35). Studies that can help us gain a better understanding of the knowledge gaps through which people with different lifestyles – particularly vulnerable populations such as the elderly - apprehend risks are rare, and this study contributes to this literature (36).

We may sum-up that the purpose of this study is threefold:

- First, to examine whether there is a statistical link between scientific literacy, as quantified through formal education, and the risk of getting diabetes. A better understanding of the link between limited education and health outcomes might be of help to understand the specific needs of at-risk groups, which is essential for the design of health primary intervention programs.
- Second, to examine whether scientific literacy or education is associated with preventive behavior. By examining these variables simultaneously, one expects to gain a better understanding of how complex interdependences between these variables work in different settings. This might help us to design health intervention programs more specific to distinct contexts.
- Third, to assess whether there is a social dimension in risk perception of elderly people. An understanding of knowledge gaps and on how the mechanisms of apprehension of risk are affected by education levels may be helpful in better define targets of intervention. Thus, potentially supporting the effectiveness of tools such as scoreboard risk or campaigns used to raise awareness.

2. Research methodology and sample

Focusing the analysis on at risk populations such as the elderly in urban contexts is critical and analysis demonstrate that diabetes is a disease that affects increasingly large cities in China, India, UK and Mexico. Analysis has shown that without prevention programs, the rate at which
this disease affects people tends to increase (e.g., 14). Diabetes, in addition to being prevalent, prevention of chronic diseases requires the adoption of habits of self-care that have standard definitions. A proxy used to measure scientific literacy is the level of formal education, known to be highly correlated with health literacy (37). Education levels are therefore used in this study to assess whether there is a link between literacy levels and risk of getting the disease. In this framework, a test on awareness of risks associated with diabetes is also applied.

In order to adjust the questions to the target population, a team of 9 nurses and 2 nutritionists from Misericórdia, a Portuguese NGO that carry out prevention/education programs to elderly populations for several years (each of these nurses has seen more than 4000 patients of this age) were extensively consulted. Based on these consultations, a set of questions regarding risks associated with diabetes was prepared. The questions focused on risk factors associated with diabetes, pathophysiology and extreme complications associated with diabetes (38). Based on the literature on risk perception regarding knowledge gaps amongst low education population (39) and the comments from the nurses, it is expected that people would know more on consequences of the diseases, e.g., the extremely high risks associated with diabetes that are directly observable. It was also expected that people with low levels of education would have substantial difficulties in identifying risk factors because the latter are not directly observable.

Our sample consists of 356 people and is mostly comprised by the elderly (M=62, 25, SD=15, 64). The majority of the individuals has basic education as the highest educational attainment: 16% are illiterate, 73% have basic education and 11% secondary and higher education. Based on Epidemiologic data on diabetes (39), we confirm that our sample is representative of the overall Portuguese population at risk of having diabetes. A risk assessment of the subjects performed in collaboration with the nurses involved in our analysis, according to a standard scale used by the medical community in general, shows that 19% scored very low in the risk of getting diabetes, 30% scored low, 21% average and 27% high and 2% very high. According to the nurses assessment, 15% of individuals are expected to get diabetes, a value similar to the incidence of diabetes in Portugal in 2010, of about 12%. In Portugal, diabetes affects people with ages between 20 and 79 years old, corresponding to approximately 991 thousand individuals (the Portuguese population is about 11 million).

3. Results

Figure 1 quantifies the potential relationship between education and risk of getting diabetes for the elderly sample and suggests the existence of a statistical link between educational levels and risk of getting diabetes. It is found that populations with no level of formal education levels have a higher risk of getting diabetes than those with basic and secondary/higher education. The estimate difference of the risk of getting diabetes between those without education and those with secondary and higher education is 20%.

Our analysis has also shown that individuals holding a higher education degree had an average decrease in risk of 1 point in a 5-point scale, which can be considered as a large effect. This is a
new and relevant finding because the literature does not consider this type of health outcome, usually measured by mortality rates. The following results represent exploratory attempts to understand this link.

The first explanation explored for this link is the relation of education with preventive behaviors. Yet, our results in Table 1 show that most of the preventive behaviors are considered to be independent of the education level. Only the level of physical activity (p<0.001) and daily water consumption (p<0.05) were shown to differ between the different education levels of the elderly sample used. Both these variables were also found to be correlated between them (r=0.131; p<0.05), as it is probably explained by the fact that people who do more exercise also tend to drink more water (because of exercising). This suggests that education is less likely to be the most critical variable in this comparison. However, the relationship between education and prevention behaviors has to be considered as weak, as educational levels are just significant for 2 of the 7 preventive behaviors under analysis.

Table 1: Relation between education levels and the adoption of risk modifiable behaviors

<table>
<thead>
<tr>
<th>Risk modifiable behaviors</th>
<th>P-Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level of physical activity</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of daily meals</td>
<td>0.068</td>
</tr>
<tr>
<td>Units of fruit eaten daily</td>
<td>0.228</td>
</tr>
<tr>
<td>Eats fiber with sugar</td>
<td>0.582</td>
</tr>
<tr>
<td>Units of vegetables eaten daily</td>
<td>0.836</td>
</tr>
<tr>
<td>Daily water consumption</td>
<td>0.006</td>
</tr>
<tr>
<td>Other drinks</td>
<td>0.769</td>
</tr>
</tbody>
</table>

* Fisher’s exact test was used.
A second possible explorative analysis to explain the link between educational levels and risk of getting diabetes was focused on how risks are perceived by the “elderly at risk” with different levels of education. Table 2 shows that no differences were found regarding educational levels in relation to risks on extreme levels of diabetes and to risk factors associated with diabetes. From the interviews with the subjects, it was expected that the population with no education or with basic levels of education were well aware that diabetes can cause death or other major physical alterations (i.e., the amputation of a leg). In general, individuals scored high on the perception of risks on extreme levels of diabetes. Inversely, all the individuals scored low on risk factors, independent of education levels, a result also validated by the interviews. It is in the risk perceptions associated to pathophysiology that differences between those more educated and those less educated or without education emerge (see Figure 2).

These results suggest that education levels play a role in risk perceptions concerning diabetes but that this role is somewhat limited, and knowledge gaps remain even amongst the more educated.

<table>
<thead>
<tr>
<th>Variable</th>
<th>(I) education</th>
<th>(J) education</th>
<th>P-Value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk perceptions on extreme complications of diabetes</td>
<td>No education</td>
<td>Basic</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>No education</td>
<td>0.716</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.385</td>
</tr>
<tr>
<td></td>
<td>Secondary and Higher</td>
<td>No education</td>
<td>0.231</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic</td>
<td>0.385</td>
</tr>
<tr>
<td>Risk perceptions on Pathophysiology</td>
<td>No education</td>
<td>Basic</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>No education</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>Secondary and Higher</td>
<td>No education</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic</td>
<td>0.139</td>
</tr>
<tr>
<td>Risk Factors associated with diabetes</td>
<td>No education</td>
<td>Basic</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.229</td>
</tr>
<tr>
<td></td>
<td>Basic</td>
<td>No education</td>
<td>0.775</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Secondary and Higher</td>
<td>0.337</td>
</tr>
<tr>
<td></td>
<td>Secondary and Higher</td>
<td>No education</td>
<td>0.229</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Basic</td>
<td>0.337</td>
</tr>
</tbody>
</table>

*Tukey’s HSD test for pair-wise comparisons was used. Note: (I) Education and (J) Education indicate which pair is being compared.
A third possible exploratory explanation for the link between educational levels and risk of getting diabetes was performed regarding health screening. The education level was found to not differ significantly amongst the non-health screening and health screening groups of subjects (p=0.832, for a Kruskal-Wallis test). These results indicate that the link between education levels and prevention of a non-communicable disease, such as diabetes, is complex. This is confirmed by the fact that no significant differences were found between the risk-aware and the non-risk-aware individuals on most prevention-related variables, except for physical activity level and type of drinks consumed (p<0.05).

Yet, significant differences were found between the screening and non-screening groups regarding their risk scores (p<0.001), with the group of subjects who perform health screening having a higher mean score on the risk scale. Moreover, individuals who perform regular health screening (at least once a year) have more preventive behaviors when compared to individuals who disregard health screening. Specifically, they eat more often each day (p<0.005, with 95 individuals on the health screening group having 5-6 meals each day, versus 2 subjects on the non-health screening group), eat more units of fruit each day (p<0.05, 16 individuals on the >5 category for the health screening group versus 2 individuals on the non-screening group), and drink more water (p<0.005, 39 subjects on the >1.5L category for the health screening group, versus 10 individuals for the non-screening group).

The fact of being at risk associated with health screening behavior leads to more prevention independently of education levels. The explanation of the link between education levels and risks of getting diabetes remains to be explored in future studies with a greater sample. For our observation it is not possible to single-out the effect of each variable and establish a cause-effect relationship. Future studies are needed to better assess potential relationships, but at this stage, it is possible to confirm the existence of an association between limited education and risk of getting the disease.
4. Discussion and conclusions

This study focuses on the elderly in vulnerable urban zones, an age strata with the highest incidence of diabetes among all age strata. For example, with reference to the Portuguese population at large, a recent report indicates that 27% of people in the 60-75 year-old strata is afflicted with diabetes, with a prevalence of 13% in the 40-59 year-old strata and 2% in younger subjects (39). As a result, it has been emphasized that patients’ educational levels must be considered in the care and information of patients with chronic diseases such as diabetes (e.g, 26). However, this is the first study, to the best of our knowledge that empirically demonstrates the link between educational level and risk of getting diabetes.

Despite the existence of a correlation between education levels and risk of getting diabetes, a significant relationship between risk perception associated with diabetes and preventive behavior was found to be limited.

These results need to be interpreted carefully. It could be argued that, as our sample population does not know much on risk factors, there is no association between perceptions on risk factors and prevention. Alternatively, it might be suggested that there is a greater level of complexity at stake (see 35). The analysis shows that people at risk of having diabetes engage more often in health screening, while those engaging often in health screening are those doing systematic prevention. In other words, people do not systematically consider prevention measures, because they do it only when they are at risk.

Our analysis also confirms that prevention behaviors of those more educated are indistinguishable of those less educated. This suggests that there is an important gap at this level and that there is need for risk education concerning diabetes, which is known to be successfully carried out through campaigns, namely with regard to risk factors as a recent study in China has shown (39).

The fact that no relation was found between education and preventive behavior may also be related with lack of health literacy. A recent study demonstrated that health literacy is positively correlated with the adoption of risk modifiable behaviors (40), leading us to argue that educated and less elderly educated people at risk continues to lack health literacy.

While exploring how education levels affect the risk of getting diabetes, it was found that the impact of education on perceptions and on behaviors is dissimilar highlighting the complexity of this relation. This complexity is further underlined by the fact that no correlation was identified between risk awareness, preventive behaviors and health outcomes. This complexity at the level of primary intervention might be associated with the fact that these variables might correlate with each other in some settings but not in others, as suggested by studies focused on secondary and tertiary intervention (e.g, 36). One of the boundaries conditions found in the latter studies concerns the need for health awareness programs for minorities.

In addition to these results, our work also presents preliminary evidence of new ways to assess risk perception associated with diabetes for the elderly. This study indicates that typologies of risk perception (risk factors, pathophysiology and the consequences and the diseases) could be
taken into account. By using these typologies, a better understanding on how risk perceptions are structured can be achieved. The same holds true regarding the creation of design scales that capture differences and similarities in this population in this group and consequently design more effective prevention programs.

To conclude, it should be noted that the generalizability of our findings cannot be easily discerned. Larger samples, as well as samples including differences across groups with different educational levels, maybe required to test the relationship between awareness of risks and preventive behavior. Yet, limited educational levels, especially among the elderly, must be considered in patient education. By studying such relationships, some of the most promising areas for intervention research start to be highlighted, as well as important gaps in our current understanding of the pathways linking science literacy and health.

References


31. Elizabeth J. Mayer-Davis PhD, Tina Costacou MSc Obesity and sedentary lifestyle: Modifiable risk factors for prevention of type 2 diabetes, Current Diabetes Reports, 2001, Volume 1, Issue 2, pp 170-176 


Chapter 15:
Carlos Silva, Gonçalo A. D. Pereira and Susana Vieira, "Diversity in energy consumer taxonomies: understanding user capabilities for sustainable cities".
Diversity in energy consumer taxonomies: understanding user capabilities for sustainable cities

Carlos A. Santos Silva¹, Gonçalo A. D. Pereira¹, Susana Vieira²

¹ IN+ Center for Innovation, Technology and Policy Research, Instituto Superior Técnico
² Center of Intelligent Systems, IDMEC - LAETA, Instituto Superior Técnico

Abstract

Energy consumers are changing the way they acquire energy. It is relevant for energy providers to understand the main drivers of energy consumption. In this paper, we suggest that energy consumption can be divided into two parts: a part related to the household characteristics and the rest to the consumers characteristics. The paper analyses data from an ongoing project to verify the assumptions regarding household characteristics and proposes a taxonomy to classify consumers based on a literature review. The paper ends by suggesting a methodology to test the validity of the proposed taxonomy.

Introduction

Energy providers (utilities, ESCOS) around the globe are facing fundamental changes in the relationship with energy consumers, as these are changing their expectations towards the use of energy as new technological advances, regulatory frameworks or market drivers step in.

Traditionally, energy providers have a very straightforward, standardized relationship with their customers, in general restricted to bill payments and service complaints. However, as it is happening in with other areas like entertainment, energy consumerism is increasing with consumers seeking for smarter solutions for their home/company energy management.

In terms of technology, the main game changer increasing consumer expectations is the information availability at all times based on new communication platforms. In terms of regulation, regulators are defining new frameworks to impose requirements to upgrade and invest in infrastructure that enables more efficient use of energy with direct economic impact on consumers and suppliers. Finally, in terms of market, with the deregulation of the former energy markets monopolies, companies are competing against each to get/maintain clients.

Energy providers must therefore reshape their relationship with customers to a more customer-centric approach (Accenture, 2011). However, this requires a deep knowledge of the consumers and their behaviour in terms of energy, in particular those that may induce some kind of rebound effect, (IRGC, 2013).

This paper analyses consumers’ taxonomies in different areas and suggests that energy consumers at the household level have to be studied under the specific context of energy
consumption, especially in the current transitional period where energy is no longer perceived as a simple commodity, but more and more as a service. The paper describes then some initial evidences found from an experiment with 50 electricity household consumers in the area of Lisbon, Portugal and proposes a methodology to define in concrete a precise energy consumer taxonomy that captures the capabilities of the consumers to participate in the development of sustainable cities.

Understanding households energy use

According to the International Energy Agency (IEA) (International Energy Agency, 2007), energy use in the residential sector can be categorized in the following items: space heating (space cooling data is still very scarce for most of the countries), water heating, cooking, lighting and appliances use (refrigerators, freezers, dish and clothes washing, TV and other multimedia equipment’s).

In general, over the period that was analysed (from 1990 to 2004), the energy consumption increased 14% (at an average rate of 1% per year), albeit an increase of the efficiency of most equipment – without this, the consumption would have increased 25%.

The increase has been made mostly at the expenses of the growth of appliances use, which grew 50% in this period, while the other uses had modest increases. Overall, space heating represents know 54% of energy use, appliances 20%, water heating 17%, lighting 5% and coking 4%.

By source, the use of electricity and natural gas has been increasing, 35% and 23% respectively, and represent now 74% of all use, although this varies from region to region – in southern European countries, electricity and gas account in general for more than 90%.

The main drivers for energy consumption identified by the IEA in a house are floor area (mostly related to the needs of space heating), number of occupants (impact on water heating, cooking and appliances), climate (measured in Heating Degree Days (HDD)) income and energy costs.

Overall, considering the normalization by the number of HDD due to climate variations from year to year and country to country, the following conclusions can be drawn:

- There is an obvious relationship between residential floor area per capita and income and this has put upward pressure on per capita space heating and cooling needs. heating/cooling needs of a given dwelling are not;
- Average household occupancies have fallen in most of the countries (from 2.8 to 2.6) and the energy needs, specially space heating and cooling, are not necessarily reduced because fewer people are living in it;
- Higher energy costs are likely to restrain household energy consumption to some extent by encouraging more energy-efficient behaviour and purchases. However, energy costs continue to be a small proportion of household expenditure (between 2 and 3% for most European countries)
- Per capita consumption varies from country to country, but for European countries is in general within the interval of 15 to 35 GJ/per capita /per year (normalized by HHD)
From this analysis, we propose the hypothesis that energy consumption in a household can be divided in two drivers, the household characteristics and the consumer type:

\[ E_{\text{household}} = \text{Household}_{\text{characteristics}} + \text{Consumer}_{\text{characteristics}} \]

The household characteristics refers to the floor area, number of occupants, type of house, and other characteristics that influence energy consumption, like insulation, windows, etc.) and have an impact on energy heating and cooling needs. The impact of these characteristics is well understood, as it follows physical principles and has been documented in detail in the literature.

The consumer characteristics refers to the use that depends on the behavioural aspects of energy consumption, like the response to information, willingness to save energy, the concern regarding environmental impacts or knowledge about energy efficiency. The assumption is the even for the same household characteristics, different occupants use energy differently and behave differently according to the context of use. The impact of these characteristics is recognized, mentioned and studied in many different studies, as for example in (Darby, 2006), but the knowledge in this area is still diffuse and immature. This paper proposes a methodology to understand the relative weight of consumer characteristics impact on energy use compared to household characteristics and to identify the main driver of consumptions regarding consumer characteristics.

Energy consumption due to household characteristics

In order to study the impact of household characteristics in the energy consumption of a household, we analysed the energy consumption of 59 houses in the Lisbon area data from one year (January 2012 to December 2012) with a resolution of 15 minutes. This data was collected under the project SMARTGALP (GALP Energia S.A., 2013), an internet based interface to aid final customers of the residential sector to take advantage of smart metering. The project aims to reduce energy consumption by 10%, the CO2 associated emissions and the users expenses by using a continuous energy assessment process using energy meters (electricity and natural gas).

A survey was performed to all the 59 households upon the installation of smart meters in the user’s houses. The inquiry contained questions covering information about:

- Household characteristics (location, type, size, construction year)
- Inhabitants (number and occupation);
- Electric equipment inventory and usage;
- HVAC inventory and usage.

In order to successfully compare the energy consumption for different houses and identify the main drivers for energy consumption, relation between the extracted features and the energy consumption, a feature selection methodology that relates the household characteristics and the energy consumption, we applied a methodology called hierarchical clustering (P. Berkhin, 2006).

Hierarchical clustering is best for small datasets because this procedure computes a proximity matrix of the distance/similarity of every case with every other case in the dataset. An agglomerative or divisive method can be used to cluster cases. The agglomerative method
begins with each case being a cluster by itself and continues until similar clusters merge together. The divisive method begins with every case into one cluster and continues until each case is divided into individual clusters. In this case, we use the agglomerative method.

To represent the results, we present using dendrograms. Dendrograms, or tree diagrams, represent the process of going from individual cases to one large cluster. These graphs allow depicting the links between cases and its structure enables the visualization of how the different cases form the clusters. The proximity of the clusters is visible in the dendrogram by the height of the union between features, small heights translate to proximity while large heights represent remoteness.

The average household

The average house can be described as an apartment from 1995 to 2005 with 3 rooms. The house windows are double glazed with an aluminum frame and the blinds have an interior box/frame. The dominant direction of the rooms is south except for the bedrooms, that are facing east. The average family is composed by 3 people and there is 1 occupant during the day (it is not clear in the questionnaire if the occupant during the day is part of the family or does housekeeping).

The average household has a contracted power of 6,9 kVA with a dual tariff. Looking at the kitchen appliances, the average house has A++ or A class combined freezer-fridge unit, a A++ or A class dishwasher and a B class washing machine. For cooking, it has an electric oven and a gas stove and for heating water or baths it use a natural gas based tankless boiler.

The lighting equipment in the house is mainly CFL, except in the living room where incandescent lights are more common. Regarding heating equipment, the average house has no central heating but it has at least a portable electric heater.

Looking at entertainment systems, the average house has a 1 to 5 year old TV in the living room with DVR equipment and a less than 1 year old TV in the bedroom. It also has a PC and a laptop with 1 to 5 years. A gaming system is also present with less than 1 year and a 5 to 10 year old Hi-fi system. The average house leave the equipment in stand-by mode, not shutting the devices down completely.

Household characteristics analysis

The results of the application of the hierarchical clustering to the dataset are presented in the form of a dendrogram in Figure 1. This figure shows that the first major breakdown is on the left hand side related to small apartments (3 rooms or less) and small number of inhabitants, while the right hand side is related to houses with a larger number of occupants.
Analysing the data in more detail, we can observe that two other features stand out: the year of construction and the existence of central heating in small apartments. Thus, we can divide our sample in four clusters, described in detailed in Table 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>House</td>
<td>1990</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>3274</td>
<td>1803</td>
<td>5077</td>
<td>4.6</td>
</tr>
<tr>
<td>Old apartment</td>
<td>1985</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2505</td>
<td>1800</td>
<td>4305</td>
<td>5.2</td>
</tr>
<tr>
<td>New apartment</td>
<td>2003</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1090</td>
<td>1280</td>
<td>2370</td>
<td>2.8</td>
</tr>
<tr>
<td>New apartment with central heating</td>
<td>2003</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1344</td>
<td>2244</td>
<td>3588</td>
<td>6.5</td>
</tr>
</tbody>
</table>

Table 1 - Main clusters of households characteristics

It is very interesting to notice that, compared to the IEA study (International Energy Agency, 2007), the energy consumption per capita is very low – 6.5 or less compared to the 15 to 35. These results have not been normalized in terms of heating degree days, but as in Lisbon the number of heating degree days is small compared to other regions in Europe, the results should not vary considerably. This means that in Lisbon, we consume much less energy in households that other countries in Europe, not only because the climate is moderate, especially in winter. However, it may also mean that the comfort level is lower than in other countries.

Another interesting fact is that the existence of a natural gas central heating system changes significantly the structure of consumption. The general household consumes more electricity
than gas (60% to 40%), but in case there is a central heating system, the supply structure is the opposite (40% of electricity and 60% of gas).

As a conclusion, the main consumption drivers from the household characteristics that have been identified in the IEA study are verified for the analysed sample, albeit the order of magnitude of the energy consumption per capita is smaller. In general, older and bigger households with more occupants use more energy.

However, the values within these clusters have some variation. The assumption is that the household characteristics that impact the energy use are the same, thus the differences are related to the consumers characteristics.

Energy consumption due to consumers characteristics

Energy has been consumed as a good. In that sense, a possible hypothesis would be to consider that the way users consume energy is similar to the way they consume other goods, thus energy consumers could be characterized as general consumers. However, as (Aaker, 1995) describes, the task of identifying segments is difficult, because in any given context there are literally millions of ways to divide up the market and its consumers.

The hypothesis that we follow here is that energy is a special type of good and therefore their users have special features that have to be considered. To identify these features, we could not found in the literature a systematized work about the characterization of energy consumers, albeit the topic has been studied by many over the last three decades. We present here a brief review that highlights the common findings of different studies in different contexts and regions to identify the common features of energy consumers.

Energy consumer characteristics

(P, Ellis P and Gaskell, 1978) appears to be the first reference to the topic and one of their main conclusions was the importance of providing information to energy consumers to enable them to develop a learning process regarding energy use, acknowledging that users do not fully understand the good they are consuming.

The lack of knowledge has been addressed in many other works. For example, (Bang, Hae-Kyong, Ellinger, Alexander E., Hadjimarcou, John, Traichal, 2000) analyse the relationship between three variables - concern with the environment, knowledge about renewable energy, and beliefs about salient consequences of using renewable energy - with consumer attitude toward paying a premium for renewable energy. Using the theory of reasoned action as a theoretical framework, a positive relationship between beliefs about salient consequences and attitudes toward paying more for renewable energy was found. Overall concern levels were quite high for consumers in the sample, whereas knowledge levels were relatively low. Interestingly, it was found that consumer concern failed to translate into heightened knowledge about renewable energy. The study suggests that the consumer’s environmental concern and beliefs about renewable energy to date are more emotionally charged than fact- or knowledge-based.
(Weber & Perrels, 2000) analyse and quantify the impact of lifestyle factors on current and future energy demand. They analyse directly environmentally relevant consumer activities such as space heating, but also expenditure, using household survey data from the national statistical offices and results are presented for West Germany, France and the Netherlands. Again, they suggest that energy use is influenced by technical knowledge, behavioural aspects like frequency of use and economic level (like energy budget).

(Bin & Dowlatabadi, 2005) analyse also the lifestyle of consumers and its impact on green house gases emissions, looking into detail into socio-economic characteristics, personal beliefs and technical knowledge and conclude that a large influence of how consumers operate their house and that is strongly correlated to their purchasing behaviour.

In conclusion, different studies suggest that the energy use is influenced by purchasing power, personal beliefs, and knowledge. In this sense, we propose an energy consumer taxonomy that covers these dimensions.

**Energy consumer taxonomy**

Based on the literature review, the proposed taxonomy for energy consumers is the one presented in Figure 2. The figure describes the tridimensional space of users’ beliefs, knowledge and purchasing power and proposes a preliminary classification of users into three groups:

- **Bill payers**: these are the consumers that perceive energy as a good, and that have no strong beliefs regarding its use (environmental or other type). They are also characterized by having a very low technical knowledge about energy and its use. They are called *Bill payers* as the only interaction they have with the energy provider is when they pay the bill. We can find this type of user whether the purchasing power is high or low. These consumers are expected to be the most inefficient.

- **Green**: these are the consumers with very strong believes regarding the energy use and its impact, regardless of their purchasing power or knowledge. In principle these consumers will be more efficient than the bill payers.

- **Smart**: these are a subgroup of the green consumers that have high purchasing power and strong technical knowledge and are characterized by adopting practices around energy consumption to reduce it, like buying efficient equipment as early adopters, even when the economic rationale does not justify it. These consumers are expected to be the most efficient consumers. However, this users have a higher risk of causing rebound effect (IRGC, 2013).
To verify if the proposed taxonomy has any adhesion to reality and quantify the impact on energy consumption, we propose a methodology to verify its validity.

Methodology to validate energy consumer taxonomy

To validate the proposed taxonomy, we plan to develop a survey to the consumers of the SMARTGALP project. The survey will consist on a set of questions that will help to classify them within one of the three groups. The questions will try to assess the knowledge depth regarding energy use in general and energy efficiency at the household level, the strength of their beliefs regarding environmental constraints and impacts of energy use and their purchasing power, including elasticity to price and relative weight of energy budget compared to total budget.

The following limitations have been identified. Regarding the purchasing power assessment, the current sample from the SMARTGALP project includes many GALP employees and project partners, which in general belong to the medium-high socio-economic class. This can be easily checked indirectly through some elements in the preliminary survey, like the household neighbourhood, or the model of the car. In any case, the purchasing power may be high or low.

Another limitation has been identified regarding the methodology. The survey cannot be made personally and excludes therefore direct interviews. Eventually, it will be possible to perform the interviews by phone, but from previous experience, it is very hard to engage people in such a survey. On the other hand, if the survey is sent per email or posted only, it is reasonable to expect that very few participants really answer the survey.

Conclusions

In this paper, we analyse energy consumption drivers at the residential level and conclude that it can be divided into two parts: a part related to the household characteristics and the rest to the consumers characteristics. Regarding the first part, the data from an ongoing project was analysed and it was possible to aggregate the consumers into 4 groups and conclude that the
main drivers identified in other studies are also observed in the analysed sample. Regarding the second part, a review of the literature was perform and a taxonomy for energy consumers was proposed.

The paper also suggest a methodology to test the validity of the proposed taxonomy, but highlight few limitations within the sample group.

Bibliography


Chapter 16:

Paulo Ferrão, Samuel Niza and André Pina, "Urban metabolism and sustainable cities".
Abstract

The metabolism of a city is dependent on anthropogenic and natural physical inputs of energy and materials, processes for transforming those inputs for urban activities, additions to the stocks contained within its spatial boundaries, and the waste and emissions handling.

The main purpose of this work consists on further developing methods that allow identifying the material consumption of activity sectors within a city. The method presented in this article allows the use of internationally available data (e.g. OECD Input Output matrices), which enables analysing a large range of urban areas. The method is applied to metropolitan economies in Europe (Lisbon and Paris) and Asia (Seoul-Incheon and Shanghai), and the variability of the physical structure of these economies is assessed. The urban areas are compared in terms of total and type of material input, destination of material inputs within the economy and analysis of the manufacturing sectors.
Introduction and objectives

As urban expansion and new patterns of economic activity have fed on each other, novel configurations for urban areas have emerged, such as mega urban regions, urban corridors and city-regions. Currently, urban areas account for over 48 per cent of the global GDP (BI, 2012), with many of these cities having become centres of international trade and commerce and hubs for regional and international connectivity.

Cities are engines of growth but also the key drivers for stronger and more relevant global environment-economy interactions (Seto & Satterthwaite, 2010). During their development stages, cities encourage or discourage the development of particular economic activities within their boundaries, and at each stage this defines their signature (typology), including the jobs, the economic output (Spence et al. 2009) or their dependence of material resources from elsewhere and, depending on how they process them, their impact on the environment (Kennedy et al, 2007).

The quantification of the requirements of materials and commodities needed to support human tasks in cities, including the removal and disposal of waste, is known as the urban metabolism (Wolman, 1965), as a parallel to the metabolisms of ecological systems. To fully describe the metabolism of a city, several factors need to be considered: anthropogenic and natural physical inputs of energy and materials, processes for transforming those inputs for urban activities, additions to the stocks contained within its spatial boundaries, and the fluid and reliable dispersion of wastes (Niza et al, 2009).

Progress towards urban sustainability will then depend, at least in part, on the reduction of resource input on the one hand and the further reduction of pollutant output on the other hand. Since material flows through the anthroposphere connect both of those ends, the Material Flow Analysis or Material Flow Accounting (MFA) approach offers considerable potential in helping to assess the path towards sustainability (Bringezu, 1999). The metrics provided by this type of approach allow understanding how natural resources uses correlate with urban economic activities and can enable the definition of priorities for action towards decreasing the pressure cities exert on the environment.

While some research has been made on the metabolism of urban areas, they are generally applied to a small set of urban areas as they require very detailed and specific data which is not commonly available (e.g. Barles, 2009; Schulz, 2007; Hammer and Giljum, 2006; Barrett et al. 2002). As such, it becomes very difficult to compare the results obtained for different cities, limiting the conclusions that can be drawn from these studies.

The main purpose of this work consists on further developing the methods developed at IN+/IST, Portugal, (Niza et al., 2009; Rosado et al., 2012) that allow decomposing material consumption through activity sectors stemming from the underlying idea of assuming a linear relation between the material flows and the number of workers of economic activities existing in an urban area. The method presented in this article adapts and simplifies those methods to allow the use of internationally available data (e.g. OECD Input Output matrices) for a large range of countries. The method is applied to metropolitan economies in Europe (Lisbon and Paris) and Asia (Seoul-Incheon and Shanghai), allowing assessing the variability of the physical structure of these economies. Urban areas are compared in terms of total and type of material input,
destination of material inputs within the economy and analysis of the manufacturing sectors.

**Urbanization and Sustainability**

Sustainability is increasingly identified with both a process and a goal to ensure long-term human well-being that does not threaten the continued availability of critical natural resources, as recently discussed in a National Academies Publication entitled “Sustainability and the U.S. EPA” (2011). It is assumed that everything that humans require for their survival and well-being depends, directly or indirectly, on the natural environment. The environment provides the air we breathe, the water we drink, and the food we eat. Our health and well-being, our economy, and our security all require a high quality environment.

Sustainability as a goal provides the motivation for the definition of integrated regional and global policies, whereas as a process, sustainability requires a deep and multidisciplinary understanding of the intertwined nature of its major dimensions: social, economic and environmental, and the places and tools to achieve the desired goals.

The regional-global interactions are increasingly determined by urbanization, as urban systems are increasingly becoming the locus of consumption and engines of economic growth in a globalized world which everyday sees more people flowing from rural areas to cities.

Since the first modern wave of urbanization in Europe and North America between 1750 and 1950, cities are once again strongly attracting people and industry (Pieterse, 2008). However, while the first wave resulted in a substantial increase in the population of cities in rich developed nations, this second wave is without precedent in scale and characterized by even greater population growth in cities of low- and middle-income countries, with Asian cities account for half of the world’s urban population and three-quarters of the world’s 100 fastest-growing large cities are found in Asia and Africa (Satterthwaite, 2007). While Latin America and the Caribbean do not show similar urban growth, these regions still host some of the world’s largest and fastest growing cities (especially Brazil and Mexico).

Furthermore, it has been observed that while many cities continue to increase their population, urban densities have been decreasing, with global cities from developed and developing countries having achieved their peak density particularly around 1890, and experiencing a continuous decline in population density since then (Angel et al., 2010). Generally, peak densities were preceded by significant increases in population densities, mainly due to rapid population growth. On the other hand, the decline in population density is typically connected with rapid income growths.

This effect of decreasing densities represents one of the aspects of urban sprawl, which can also be assessed or measure by several other factors such as:

- The expansion over time of the total built area;
- The decline in both parameters of the density curve (intercept, corresponding to maximum densities at the urban center, and gradient, rate of decline in density as distance from the city center increases);
• The relative amount and the spatial structure of the open spaces that are fragmented by the non-contiguous and non-compact expansion of cities into the surrounding countryside;
• Compactness or accessibility metrics.

Urban sprawl is one of the key issues that cities are facing, with the amount of construction in terms of buildings and infrastructure largely surpassing the increase in population and leading to an increased per capita environmental footprint, as exemplified in Figure 1.

![Figure 1 – Examples of compact and low density city distributions. Source: UN-Habitat (2011)](image)

However, the debate on the social-environmental impacts of compact city patterns has pointed out that compact cities have important benefits but can also have significant problems (Chen et al., 2008). The main benefits identified concern the reduced land use, reduction of travel distances and car dependence, lower material and energy use for infrastructure construction and more diverse urban functional mixture of employment, recreational and commercial facilities. In addition to the economies of scale in terms of providing infrastructure, education, health care, and sanitation services, there is evidence of increasing returns to innovation and wealth creation as urban areas become larger (Bettencourt et al., 2007).

The main problems that can be identified are overcrowding of spaces which can result in traffic congestions, pollution of streets and buildings, bad air quality, reduced personal communication between neighbours with crucial effects on elderly and children, use of higher energy intensive materials for construction and higher energy demand for lighting, ventilation and refrigeration due to high-density buildings.

However, the physical components of urbanization, particularly in what concerns the conversion of land cover to urban uses, is not well understood, especially at global scales. Most of the research performed on this topic focus on individual case studies of cities or metro regions (Seto et al., 2009), making it clear that urbanization processes differ significantly among regions and countries, and even within countries.

While many studies have analysed the rate and scale of urban land change, few have evaluated the changing form of urban areas and its impact on the environment. While urban form, or urban morphology, is central to understand the impact of urbanization on the environment, as discussed by Seto et al. (2010), the study of the urban metabolism
of an urban area enables estimating the materials requirement and the waste generation that those areas inflict on the environment.

With the increase in suburbanization, generally characterized by low population density, there is also an increase in peri-urbanization areas (Kombe, 2005; Yeboah, 2000), which refer to spatially and structurally dynamic transition zones where land use, populations, economic activities and lifestyles are neither fully urban nor rural. The creation of peri-urban areas is usually initiated by an influx of capital for industries or housing development and although it is occurring throughout the world, some researchers suggest that it is the dominant form of urbanization in East Asia. This style of urban development presents new challenges for planning and governance (Maneepong et al., 2008; Hudalah et al., 2007), as well as for assessing the environmental impacts of urbanization due to their increasing demand for infrastructure growth and their rising economic power.

Furthermore, material accumulation in cities is rapidly becoming a key concern for urban actors, along with similar impacts that may result from the cumulative energy, water, plastics, chemicals, and organic materials used in cities and the wastes they generate. The use of urban metabolism frameworks for estimating material use within urban areas can therefore provide important metrics for assessing the sustainability of cities.

Methodology

The comparison between the four metropolitan areas is performed using metrics as demographics, economic structure and physical structure. The physical structure of an urban economy is described by the material throughput of that economy. To measure these flows, it is necessary to consider: Inputs – domestic extraction of resources, imports of raw materials and products; Internal processes – intermediate and final consumption; Addition to stock – accumulation of materials in the system; Outputs – emissions and wastes, exports of raw materials and products.

The material inputs of an urban area may derive from locally extracted materials, imports from the rest of the country or imports from abroad. The raw materials and intermediate goods imported are going to be used by economic activities to produce final goods that will eventually be used for final consumption, either by the economic activities themselves or by the citizens and the government, or exported. Allocating materials to economic activities allows describing the production structure of the urban area in mass units.

Part of the consumption is accumulated to the material stock of the local economy (namely as buildings, infrastructures, durable goods). The remaining leaves the economy as valuable products, waste and emissions (be it to the local environment or outside it). Additionally, there is a large fraction of materials that are imported and to a great extent re-exported, designated transit or crossing flows. This fraction hardly becomes part of the urban economy since the region (through its harbours, train-stations and airports) works essentially as gateway to other regions.

Commonly, the structure of statistical records does not include data at urban level; consequently the accounting includes estimating values from existing data, usually at higher scales (namely regional or even country level). Niza et al. (2009) and Rosado et
al. (2013) developed a method to account and disaggregate urban flows using data produced for several scales. The method is based on EUROSTAT’s economy-wide material flow accounting (EUROSTAT, 2001), but requires detailed statistics, particularly at the urban area level, such as International Trade Statistics: to the urban area; Transport Statistics: within the urban area and between the area and the other regions of the country; Industrial Production Statistics; Mineral extraction; Agricultural harvest; Forestry; Fisheries; Industrial waste; Municipal solid waste and Emissions. While some statistics are available for many urban areas around the world, transport statistics are rare, making it difficult to know which materials and products are imported to an urban area from the rest of the country.

The method developed in this article consists on estimating the metabolism of an urban area by scaling down the metabolism calculated for the country. The calculation of the country urban metabolism is performed using published Input-Output tables (IO tables) (OECD1). These tables characterize the sales from each economic sector to the other, the consumption of households, the consumption of the government, the acquisition of buildings and machinery by companies (gross fixed capital formation, GFCF) and the exports – these are the demand parameters of IO tables. The input of materials to the country is obtained through domestic extraction and trade statistics (FAO2, UN Comtrade3 among others).

In order to distribute the input of materials by the economic sectors, the final consumption and the exports, each material and product entering the economy is first allocated to the economic sectors that produce them. This is performed using correspondence tables linking commodities (expressed in SITC4, EW-MFA5, HS6 or CN7 nomenclatures) to economic activities (expressed in nomenclatures such as ISIC8 and NACE9), and conversion tables for nomenclatures of materials and for nomenclatures of economic activities.

The distribution of materials through the economy is then made based on the sales registered by each economic sector, with the materials coming from each sector being split by all the economic sectors, final consumption and exports according to the purchases made from that sector. These calculations enable the estimation of how materials are distributed within a country economy. Due to the large variety of materials and products that enter an economy, making it difficult to even analyse the results, the products are converted to a nomenclature representing categories of materials named MATCAT (Rosado et al., 2013). This nomenclature was created at IST, and establishes a correspondence between products listed in the CN and the materials that constitute them (Table 1). The MATCAT nomenclature considers 6 categories of materials (fossil fuels, metallic minerals, non-metallic minerals, biomass, chemical and others), with a total of 28 subcategories. This enables an easier analysis of the results, and allows identifying which types of materials is an economy more dependent on.

4 Standard International Trade Classification
5 Economy-Wide Material Flow Accounts
6 Harmonized System Codes
7 Combined Nomenclature
8 International Standard Industrial Classification
9 Statistical Classification of Economic Activities in the European Community
The key methodological step to scale down consists on using the share of the country’s workers that exist in the urban area, per economic activity, as well as the share of population. The share of workers is used to estimate the amount of materials consumed by each economic sector, as well as the amount of materials and products produced by each economy sector that are for international export. The final consumption by households and by the government was estimated using the share of population.

This method determines that the material input to an urban economy is composed of the materials locally consumed by the economic sectors (households, government and companies) and the materials used for international exports: Material input = local consumption + international exports, where local consumption= consumption by households + consumption by government + consumption by companies.

Though the materials used for exports to the rest of the country are included in the direct material input they are not identified through this method. Therefore, when comparing to the EUROSTAT material flows indicators (EUROSTAT, 2001) the material input is a value between the Direct Material Input (Domestic Extraction + Imports) and the Domestic Material Consumption (DMI – Exports), with the local consumption being equal to the DMC. Consequently, comparison of the results of this study with other urban metabolism studies should be performed carefully.

Moreover, the methodology described in this work does not identify the crossing flows within an urban area. However, as the total material input to an urban area is determined by the total final consumption and the exports, the materials that cross the urban area are also not accounted in the inputs, thus enabling a material balance.
Results and discussion

The methodology was applied to four urban areas (Seoul-Incheon, Shanghai, Paris and Lisbon) for the year 2000. Results comparing the total material input, the local consumption, the weight of the different economic sectors in terms of material consumption and analysis of the manufacturing sectors, are shown in the next paragraphs.

The socioeconomic characteristics of the chosen metropolitan areas are very diverse, in terms of sheer size, population and relative wealth, as can be seen on Table 2. From the four urban areas under analysis, Seoul-Incheon was the largest metropolitan region with an area above thirteen thousand square kilometres, more than four times the shortest, Lisbon. Seoul-Incheon was also the most populous of the metro areas, followed by Shanghai, with eight and six times more population, respectively, than the least populated, Lisbon.

Shanghai was also the densest metro area with more than two thousand inhabitants per square kilometre, which is more than two times denser than Lisbon. It is interesting to notice that while Seoul-Incheon is the most populous urban area, Shanghai is a much more compact city, with the population density of Seoul-Incheon being three quarters as that of Shanghai.

Table 2 – Demographic and output indicators of the four metropolitan areas

<table>
<thead>
<tr>
<th>Area</th>
<th>Population</th>
<th>Share of national population</th>
<th>Nominal GDP</th>
<th>Share of national GDP</th>
<th>GDP per capita</th>
</tr>
</thead>
<tbody>
<tr>
<td>km²</td>
<td>(thousand)</td>
<td>(hab/km²)</td>
<td>(mils $)</td>
<td>(%)</td>
<td>(thousands $)</td>
</tr>
</tbody>
</table>

The wealthiest area per capita was Paris, followed by Lisbon. The metro economies that contributed more to the country wealth were Seoul, with almost half of the country output, and Lisbon, with more than a third. Shanghai, on the other hand, represented only 2% of the country GDP, which can be explained by the large country that is China, with Shanghai having only 1% of the population of the country at the time.

The application of the methodology showed that there is not a correlation between the wealthiest areas per capita and the ones that consume more materials per capita, as can be seen in Table 3. While Paris was the richest urban area and had the highest consumption per capita (both in terms of material input and local consumption), the

---

10 http://www.ine.pt/
11 http://www.insee.fr/
12 http://kosis.kr/eng/
13 http://www.stats-sh.gov.cn/
Seoul-Incheon urban area had the second highest material input and material consumption but a much lower GDP per capita when compared to Lisbon. Furthermore, Shanghai had a very low GDP per capita when compared to the other urban areas under analysis, but its material input and local consumption was of the same order as all other urban areas.

<table>
<thead>
<tr>
<th></th>
<th>Material input</th>
<th>Local consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total [kt]</td>
<td>Per capita [t/cap]</td>
</tr>
<tr>
<td>Lisbon</td>
<td>34.880</td>
<td>13.1</td>
</tr>
<tr>
<td>Paris</td>
<td>260.698</td>
<td>23.8</td>
</tr>
<tr>
<td>Seoul-Incheon</td>
<td>369.521</td>
<td>17.3</td>
</tr>
<tr>
<td>Shanghai</td>
<td>175.806</td>
<td>10.9</td>
</tr>
</tbody>
</table>

When assessing the material input structure of the four urban areas, as shown in Figure 2, it is possible to see that non-metallic minerals and biomass are the most important material categories. This can be explained by the large and complex infrastructures that are required for urban areas to function, as well as the use of biomass products to feed the population or food industries. These two types of materials account for between 58%, in the case of Seoul-Incheon, and 83%, in the case of Lisbon, of the total material input.

In the case of the non-metallic minerals share, the MatCat subcategory NM4 (stone), represents more than 95% of the material input in Seoul-Incheon and Shanghai, while in Paris and Lisbon it corresponds to below 80% with NM1 (sands), being responsible for between 12% and 18%.

In terms of biomass, the subcategories BM1 (agricultural biomass), BM6 (wood), and BM8 (non-specified biomass), are the most relevant. Together, they account for over 85% of all biomass used in each urban area.

Figure 2 – Material input structure of the four urban areas, 2000. Own calculations.
The urban areas of Lisbon and Paris rely on very high shares of non-metallic minerals use, above 50%, and have significant biomass needs, around 25%. The urban area of Seoul-Incheon relies also significantly on non-metallic minerals, close to 50%, but also uses a high amount of fossil fuels, reaching up to 30% of the total material input. Finally, the urban area of Shanghai is characterized by using a very significant amount of biomass, close to 40%, relying also on non-metallic minerals (around 35%).

The high share of non-metallic minerals in Lisbon and Paris can be explained by the boom of construction that these cities were experiencing during the beginning of the 2000’s. In the particular case of Lisbon, the city was undergoing significant transformations in some areas, such as the Parque das Nações area, with several new buildings being constructed.

The high share of fossil fuels consumption in Seoul-Incheon can suggest that the industry in the area is highly energy intensive, there is specific industry for fossil fuel products or there is a heavy use of private transport (or a mix of these hypotheses).

To assess the diversity of material usage in the urban areas under analysis, Figure 3 shows the cumulative share of total material input, with the 28 materials subcategories defined in MATCAT. For each urban area, the subcategories are ordered by their share of total material input, from the highest to the lowest. The results show that only a small number of subcategories have a significant share of total material input in all urban areas.

![Figure 3 – Cumulative share of the 28 material categories in the four metro areas, 2000. Own calculations.](image)

The urban areas of Lisbon and Paris are the more diverse in terms of material input subcategories, while Shanghai is the more dependent on specific materials subcategories. Interestingly, this contrasts with the results obtained from the material input structure, shown in Figure 2, in which Shanghai was the urban area that present the more balanced structure. This is due to the fact that there is one subcategory in each major material category that is responsible for more than 60% of the input of those categories in Shanghai.
This diversity of materials used in the urban areas of Lisbon and Paris can be observed by the fact that to achieve a share of 90% of the material input, they required 10 and 9 different subcategories of materials, respectively. Seoul-Incheon and Shanghai, on the other hand, required only 7 and 5 subcategories, respectively. The division of these subcategories in the main categories is as follows:

- Lisbon: 10 subcategories, being 4 related to fossil fuels, 3 to biomass and 3 to non-metallic;
- Paris: 9 subcategories, being 3 related to fossil fuels, 2 to biomass, 3 to non-metallic and 1 to metallic minerals;
- Seoul-Incheon: 7 subcategories, being 4 related to fossil fuels, 1 to biomass, 1 to non-metallic and 1 to metallic;
- Shanghai: the 5 subcategories are divided into 1 of fossil fuels, 2 of biomass, 1 of non-metallic and 1 of metallic.

It is important to notice that while the diversity of material input may indicate a diversified economy, results need to be considered carefully, as they may also indicate that some urban areas do not have material intensive industries, leading to no type of material being dominant. To better understand this issue, the material input per sector of the economy is shown in Figure 4.

![Figure 4](image)

**Figure 4 – Material input per sector category in the four metro economies, 2000. Own calculation.**

As can be seen from the figure, Paris and Lisbon present again very similar shares of the material input with the manufacturing sector representing only between 13% and 17%, the final consumption between 22% and 25% and the gross fixed capital formation (GFCF) between 13% and 15%. This high share of material input to GFCF will have significant impacts in the evolution of the urban area, as it mainly includes material with long life spans that are not consumed in that year, but rather stay within the city for several years (material stock).

Seoul-Incheon and Shanghai have a material input to the secondary sector (manufacturing + construction and utilities) of around 50% of the materials that enter these economies. The manufacturing sector, in particular, uses a very high share of the
materials of these economies, around 30%. These are also the economies that have the least share of materials consumed by the gross fixed capital formation.

In what concerns exports, Shanghai is the metro area whose share of materials is the lowest, with only 5% of the materials input being exported from the urban area. On the other hand, Lisbon, Paris and Seoul-Incheon have very similar shares of materials being exported, with 10%, 14% and 13% respectively.

The low share of material use in agriculture and mining is not surprising, as the areas under analysis are densely populated, much more than the rest of country, with little space available for exploring raw materials such as biomass or minerals.

The results obtained support the last decades developments of these economies where the Asian metro areas specialized in producing tradable products (Shanghai and Seoul) while the Western Europe metro areas specialized in services and non-tradable products (see for instance the share of Final Consumption and GFCF in Paris and Lisbon).

In this analysis, the material input to the manufacturing sector represents the self-consumption of products by companies. Using the consumption of materials by the different types of industries as a proxy of the importance of each industrial sector, Figure 4 shows the industrial structure of each urban area.

As can be seen from the figure, Lisbon had the least diversified manufacturing sectors, with the biomass products and the construction products industries accounting for 40% and 30% of the total material input. On the other hand, Seoul-Incheon appears to be the more diverse economy, with all types of industries accounting for at least 10% of the material input to the manufacturing sector.

![Figure 5 – Share of material input for the manufacturing sector, used by each industry type, for the four metropolitan areas, 2000. Own calculations.](image)

Generally, the biomass products industry is very significant in the urban areas under analysis, being responsible for more than 30% of the products consumed by manufactures in all the metro areas, with the exception of Seoul-Incheon, in which this share was only 14%. The next most consuming industries are the chemicals and fuels (e.g. 32% in Paris, 27% in Shanghai and 25% in Seoul) and machinery and equipment
(e.g. 31% in Seoul, 26% in Shanghai and 24% in Paris), which can have high economic value.

When comparing the diversity of the industries with the socioeconomic characteristics of the urban areas, there does not seem to exist a direct relation between the structure of the manufacturing sector and the GDP per capita. This can also be due to the large impact that commerce and services have on the GDP of urban areas, while they are low material intensive. However, the low number of urban areas under analysis does not allow drawing strong conclusions on the correlation between these factors. Nonetheless, it is interesting to notice that urban areas with more diversified manufacturing sectors are also the ones that have more population, thus not being required to specialize in only one or two economic activities.

**Conclusions**

The study of urban areas in terms of how to drive their economic growth and the impact they have on the environment has become one of the more important topics in the world due to the growing urbanization trends worldwide. However, more than focusing on CO2 emissions that occur in urban areas, cities are being viewed from a holistic approach, with research trying to understand the metabolism behind a city, i.e. which materials enter the urban area, how they are used within the city and how they leave it.

This research work presents a method for estimating the metabolism of urban areas, based on the analysis of national metabolisms and the assumption that the flow of materials is proportional to the share of workers within each economic sector. The methodology was applied to four urban areas (Lisbon, Paris, Seoul-Incheon and Shanghai), using the year 2000 as reference.

The results showed that the metabolisms of Lisbon and Paris are very similar, being more dependent on non-metallic minerals, while Seoul-Incheon and Shanghai have different metabolisms, consuming significant shares of fossil fuels and biomass, respectively. In terms of the destination of the materials that enter the urban areas, Shanghai and Seoul direct most of it to their economic sectors, particularly the manufacturing sector, while Paris and Lisbon direct them more towards final consumption, GFCF and exports.

In terms of the development of the manufacturing sectors, the results point to the indication that metropolitan areas with higher population are able to diversify their manufacturing sectors, and thus extend their economic activities. However, the conclusions that can be drawn from this study need to be taken lightly as only four urban areas where analysed.

The potential for this methodology will continue to be studied through its application to more urban areas, which can enable the clustering based on their metabolism. Furthermore, the comparison of urban areas and other regions of a country can provide insights on how prepared is a region to support and trade with its neighbours, which might have a significant impact on its economic development. Finally, the methodology will also be used to analyse urban areas over several years, which can help define evolution patterns for urban areas, and therefore help guide the development of the new urban areas appearing throughout the world.
Referências


Chapter 17:
Tânia Sousa, Tiago Domingos, J.C. Poggiale and S.A.L.M. Kooijman "Dynamic Energy Budget Theory Restores coherence in biology".
Dynamic energy budget theory restores coherence in biology

Tânia Sousa, Tiago Domingos, J.-C. Poggiale and S. A. L. M. Kooijman

*Phil. Trans. R. Soc. B* 2010 **365**, 3413-3428
doi: 10.1098/rstb.2010.0166

References

This article cites 101 articles, 10 of which can be accessed free
http://rstb.royalsocietypublishing.org/content/365/1557/3413.full.html#ref-list-1

Rapid response

Respond to this article
http://rstb.royalsocietypublishing.org/letters/submit/royptb;365/1557/3413

Subject collections

Articles on similar topics can be found in the following collections
- cellular biology (87 articles)
- ecology (1798 articles)
- evolution (2062 articles)

Email alerting service

Receive free email alerts when new articles cite this article - sign up in the box at the top right-hand corner of the article or click here

To subscribe to *Phil. Trans. R. Soc. B* go to: http://rstb.royalsocietypublishing.org/subscriptions

This journal is © 2010 The Royal Society
Introduction

Dynamic energy budget theory restores coherence in biology

Tânia Sousa1,*, Tiago Domingos1, J.-C. Poggiale2 and S. A. L. M. Kooijman3

1Environment and Energy Section, DEM, Instituto Superior Técnico. Av. Rovisco Pais, 1, 1049-001 Lisboa, Portugal
2Centre d’Océanologie de Marseille, UMR CNRS 6117 Laboratoire de Microbiologie, de Géochimie et d’Ecologie Marines. Université de la Méditerranée, Campus de Luminy Case, 901, 13288 Marseille, Cedex 9, France
3Department of Theoretical Biology, Vrije Universiteit, Amsterdam, The Netherlands

We present the state of the art of the development of dynamic energy budget theory, and its expected developments in the near future within the molecular, physiological and ecological domains. The degree of formalization in the set-up of the theory, with its roots in chemistry, physics, thermodynamics, evolution and the consistent application of Occam’s razor, is discussed. We place the various contributions in the theme issue within this theoretical setting, and sketch the scope of actual and potential applications.

Keywords: dynamic energy budget theory; biology; metabolism; ecology; evolution; energetics

1. REACHING OUT FOR GENERALITY

In physics, there is a quest for a unified theory. Physical theories have a broad spectrum of application, a strong mathematical background and are subject to numerous empirical tests. By contrast, in biology, mathematical theory has played a secondary role because biology is frequently seen as a science of exceptions and particular cases, with little interest in abstraction and generalization. Exceptions are the research being done in the fields of theoretical biology and mathematical biology. However, theoretical and mathematical biology have frequently been carried out without a concern for empirical testing. When this concern appears, models are of narrow application, reducing their theoretical breadth. The dynamic energy budget (DEB) theory starts from the Dutch tradition of theoretical and mathematical biology, but couples it with a fundamental concern in producing general theory that is subjected to careful empirical testing.

DEB theory aims to capture the quantitative aspects of metabolism at the individual level for organisms of all species. It builds on the premise that the mechanisms that are responsible for the organization of metabolism are not species-specific (Kooijman 2001, 2010). This hope for generality is supported by (i) the universality of physics and evolution and (ii) the existence of widespread biological empirical patterns among organisms (Sousa et al. 2008). Table 1 synthesizes the essential criteria for any general model for the metabolism of individuals. We explore the links between DEB theory and each of the proposed criteria in the following paragraphs.

DEB theory is explicitly based on the conservation of mass, isotopes, energy and time, including the inherent degradation of energy associated with all processes. So it complies to criteria 1, table 1. For example, consider the most important difference between DEB models, the number of reserves (biomass components that fuel metabolism) and structures (biomass components that have maintenance needs) that are delineated. This depends on the degree of coupling of the various substrates an organism needs. Animals feed on other organisms, which couples uptake of the various substrates (proteins, carbohydrates, lipids, nutrients) tightly and explains why a single reserve and structure is appropriate for them. This does not hold for plants, for instance, where multiple reserves and structures (root, shoot) are required. This means that the applicability of a model can be judged a priori.

The structure of DEB theory is such that there is a smooth merging and splitting of reserves and structures, which is a key feature in response to evolutionary changes in acquisition strategies (Kooijman et al. 2003; Kooijman 2004, 2010;
Table 1. Criteria for general explanatory models for the energetics of individuals.

1. Consistency with other scientific knowledge: the models must be based on explicit assumptions that are consistent with thermodynamics, physics, (geo)chemistry and evolution.
2. Consistency with empirical data: the assumptions should be consistent with empirical patterns.
3. Life-cycle approach: the assumptions should cover the full life cycle of the individual, from initiation of development to death.
4. Occam's razor: the general model should be as simple as possible (and not more). The predictions should be testable in practice, which typically constrains its maximum complexity substantially (quantified in terms of number of variables and parameters).
5. Taxon-specific adaptations: restrictions that make a model applicable to particular taxa only, should: (a) be consistent with an explicit evolutionary scenario; (b) be explicit to allow the prediction that the model will apply to those species.

Kooijman & Hengeveld 2005; Troost et al. 2005; Kooijman & Troost 2007. It is possible to smoothly convert one DEB model into another, according to an evolutionary scenario that makes DEB species-specific models consistent with an evolutionary scenario (criterion 5, Table 1). This includes organisms that evolved from the merging of ancestors such as the mitochondria and chloroplasts that once had an independent existence, and many of the symbioses (e.g. corals) that exist today.

In an attempt to be explicit on consistency with empirical observations (criterion 2, Table 1), we organized generally observed patterns in empirical data on various aspects of energetics, life stages and stoichiometry in tables 2 and 3 (Sousa et al. 2006). DEB theory has an explanation for each of them. Many empirical models, such as Droop’s model for the nutrient limited growth of algae and Huggett and Widdas’ model for foetal growth, are special cases of DEB theory (Table 4). The large collection of empirical support for all these findings that accumulated in the literature and the bits of evidence that people working with DEB accumulated during the 30 years of DEB research makes DEB theory presently the best tested quantitative theory in biology (Kooijman 2010).

The pragmatic application of Occam’s razor (criterion 4, Table 1) in the construction of DEB theory privileged the smallest (possible) number of state variables, the smallest (possible) number of parameters, constant functions instead of linear and linear functions instead of nonlinear. For example, the variable stoichiometry of organisms, exposed to different food levels, is explained, in the DEB standard model, by describing biomass as two aggregated chemical compounds with constant chemical compositions and variable relative amounts.

Biomass is assumed to consist of one or more reserves and one or more structures. The dynamics of these metabolic pools is followed using five concepts of homeostasis, which are all meant for simplification and enhancing the testability of model predictions.

The various forms of homeostasis are linked to the principle (criterion 1, Table 1) that organisms have increased their control over metabolism during evolution allowing for some adaptation to environmental changes in short periods.

(a) Strong homeostasis

Metabolic pools do not change in composition and can be conceived as generalized compounds, i.e. mixtures of a large number of compounds of constant chemical composition and thermodynamic properties. The individual feeds on substrate (food, \( X \)) and produces products (faeces, water, carbon dioxide, ammonia, etc.), biomass (reserve \( E \) and structure \( V \)) and gametes (reserve allocated to reproduction). The standard DEB model (but not DEB models in general) assumes a fixed chemical composition for food. All (generalized) compounds have constant thermodynamic properties such as mass–energy couplers (chemical potentials) and mass–entropy couplers (specific entropies). Strong homeostasis imposes constant conversion coefficients on all aggregated chemical reactions occurring in the organism including assimilation, dissipation and growth, which comes with stoichiometric constraints. The combination of stoichiometric constraints and variations in the composition of biomass (reserve/structure ratio) leads to rather complex patterns at the various levels of organization.

(b) Weak homeostasis

The individual as a whole does not change in composition during growth in environments with constant food availability (possibly after an adaptation period). The composition (controlled by the ratio of reserve to structure) varies with the level of food availability. This implies constraints on the dynamics of reserve relative to structure.

(c) Structural homeostasis

The individual does not change the shape during growth, which controls how surface area relates to volume as they change in time. This simplifies the control of metabolism since some processes are proportional to surface area while others are proportional to volume. Transport processes, including food uptake, uptake and elimination of toxicants, osmosis and heat transfer, are proportional to surface area, which is compatible with the description of these processes in non-equilibrium thermodynamics (criterion 1, Table 1). By contrast, most maintenance costs are linked to (structural) mass (turnover), so to volume. The scaling of feeding relative to maintenance controls ultimate body size. Only the standard DEB model makes use of structural homeostasis, not the wider class of uni- and multivariate DEB models.

(d) Acquisition homeostasis

The individual eats what it needs (demand systems), rather than what is available (supply systems). Species can be ranked on the supply–demand spectrum; no species can follow the demand rules into the extreme (food must obviously be available at some minimum
level). All demand systems are animals that have typically a higher behavioural flexibility and a lower metabolic flexibility. Demand systems evolved from supply systems and most are endothermic.

(e) **Thermal homeostasis**

The individual (endotherms, mainly birds and mammals) heats the body to a constant temperature. This behaviour has an energetic cost, that might be

---

**Table 2.** Stylized facts and empirical evidence on feeding, growth, reproduction, respiration and death.

<table>
<thead>
<tr>
<th>stylized facts</th>
<th>empirical evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>feeding during starvation, organisms are able to reproduce</td>
<td>animals: Kjesbu et al. (1991), Hirche &amp; Kattner (1993) and Kirk (1997)</td>
</tr>
<tr>
<td>during starvation, organisms are able to survive for some time</td>
<td>animals: Stockhoff (1991) and Letcher et al. (1996)</td>
</tr>
<tr>
<td>the growth of isomorphic organisms at abundant food is well described by the von Bertalanffy growth curve Putter (1920) and von Bertalanffy (1938) for different constant food levels the inverse von Bertalanffy growth rate increases linearly with ultimate length Putter (1920) many species do not stop growing after reproduction has started, i.e. they exhibit indeterminate growth Heino &amp; Kaitala (1999) and Kozlowski (1996) holometabolic insects are an exception foetuses increase in weight approximately proportional to cubed time Huggett &amp; Widdas (1951)</td>
<td>animals: Frazer et al. (1990), Strum (1991), Chen et al. (1992), Schwartz &amp; Hundertmark (1993), Ferreira &amp; Russ (1994) and Ross et al. (1995)</td>
</tr>
<tr>
<td>the von Bertalanffy growth rate of different species corrected for a common body temperature decreases almost proportional to maximum body length</td>
<td>bacteria: Kooijman (2000, pp. 276–282)</td>
</tr>
<tr>
<td>reproduction increases with size intraspecifically but decreases with size interspecifically</td>
<td>animals: Peters (1983) and Kooijman (2010, pp. 69,323)</td>
</tr>
<tr>
<td>the use of dioxygen scales approximately with body weight raised to a power close to 0.75 Kleiber (1932) organisms show a transient increase in metabolic rate after ingesting food (heat increment of feeding)</td>
<td>animals: Richman (1958), Clarke &amp; Johnston (1999) and Savage et al. (2004)</td>
</tr>
<tr>
<td>ageing lifespan increases with size for endotherms, but is independent of size in ectotherms</td>
<td>animals: Jans &amp; Chappell (1995), Chappell et al. (1997), Hawkins et al. (1997), Rosen &amp; Trites (1997) and Nespolo et al. (2005)</td>
</tr>
</tbody>
</table>

**Table 3.** Stylized facts and empirical evidence on stoichiometry, energy dissipation and cells.

<table>
<thead>
<tr>
<th>stylized facts</th>
<th>empirical evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>stoichiometry chemical body composition of well- and poorly-fed organisms differ</td>
<td>animals: Hirche &amp; Kattner (1993), Du &amp; Mai (2004), Chen et al. (2005) and Molnar et al. (2006)</td>
</tr>
<tr>
<td>energy dissipation dissipation heat is a weighted sum of three mass flows: carbon dioxide, dioxygen and nitrogenous waste</td>
<td>yeasts: Hanegraaf et al. (2000)</td>
</tr>
<tr>
<td>cells cells in a tissue are metabolically very similar regardless the size of the organism Morowitz (1968)</td>
<td>animals: Huggett &amp; Widdas (1951) and Zonneveld &amp; Kooijman (1993)</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
</tbody>
</table>

**Phil. Trans. R. Soc. B** (2010)
significant under particular conditions, but it allows these species a higher independence over the environment since all metabolic rates depend on temperature.

The state variables of DEB theory are the structure(s), the reserve(s) and the level of maturity. The level of maturity controls life-stage transitions that cover the full life cycle of the organism (criterion 3, table 1).

Consistency with the evolutionary principle (criterion 1, table 1) that organisms inherit parents’ characteristics in a sloppy way allowing for some adaptation to environmental changes across generations makes the set of parameter values in DEB individual-specific. Selection leads to evolution characterized by a change in the species’ parameters (mean) values. The differences between the mean parameters values of different species are an evolutionary amplification of the differences between the parameters values of individuals.

Parameters of the standard DEB model can be classified into two classes: size-independent parameters that only depend on the very local physico-chemical sub-organismal (cell) conditions (but not on body size) and design parameters that depend on the maximum size of the individual. Size-independent parameters are assumed to be constant across species because cells are metabolically similar, regardless of the species or body size (table 3), which is consistent with Occam’s razor (criterion 4, table 1) and evolution (criterion 1, table 1). The DEB body size scaling relationships predict how these parameter values change as a function of the maximum size of the species (Sousa et al. 2008; Kooijman 2010).

The first focus of DEB theory is the individual level, but it has many implications for the sub- and supra-organismic levels (Nisbet et al. 2000; Kooijman 2001, 2010; Kooijman & Segel 2005). There is a direct coupling of sub-organismal processes to the individual level. For instance, hormonal regulation might stimulate growth and reproduction in particular situations, but it will not occur if substrate is not available. This testifies that our understanding of regulation processes must come from a multi-level analysis. There is also a direct coupling of the individual to the supra-individual level. For instance, the processes of food selection, feeding and product formation at the individual level directly link to interaction between individuals and species, in terms of competition, predation and syntrophy. These are key processes at the population level.

### Table 4. Empirical models that turn out to be special cases of DEB models, or very good numerical approximations to them.

<table>
<thead>
<tr>
<th>author</th>
<th>year</th>
<th>model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lavoisier</td>
<td>1780</td>
<td>multiple regression of heat against mineral fluxes</td>
</tr>
<tr>
<td>Gompertz</td>
<td>1825</td>
<td>survival probability for ageing (Gompertz 1825)</td>
</tr>
<tr>
<td>Arrhenius</td>
<td>1889</td>
<td>temperature dependence of physiological rates</td>
</tr>
<tr>
<td>Huxley</td>
<td>1891</td>
<td>allometric growth of body parts (Huxley 1932)</td>
</tr>
<tr>
<td>Henri</td>
<td>1902</td>
<td>Michaelis Menten kinetics</td>
</tr>
<tr>
<td>Blackman</td>
<td>1905</td>
<td>bilinear functional response (Blackman 1905)</td>
</tr>
<tr>
<td>Hill</td>
<td>1910</td>
<td>Hill’s functional response (Hill 1910)</td>
</tr>
<tr>
<td>Thornton</td>
<td>1917</td>
<td>heat dissipation (Thornton 1917)</td>
</tr>
<tr>
<td>Putter</td>
<td>1920</td>
<td>von Bertalanffy growth of individuals (Putter 1920)</td>
</tr>
<tr>
<td>Pearl</td>
<td>1927</td>
<td>logistic population growth (Pearl 1927)</td>
</tr>
<tr>
<td>Fisher and Tippitt</td>
<td>1928</td>
<td>Weibull aging (Fisher &amp; Tippitt 1928)</td>
</tr>
<tr>
<td>Kleiber</td>
<td>1932</td>
<td>respiration scales with body weight raised to 3/4</td>
</tr>
<tr>
<td>Mayneord</td>
<td>1932</td>
<td>cube root growth of tumours (Mayneord 1932)</td>
</tr>
<tr>
<td>Monod</td>
<td>1942</td>
<td>growth of bacterial populations (Monod 1942)</td>
</tr>
<tr>
<td>Emerson</td>
<td>1950</td>
<td>cube root growth of bacterial colonies (Emerson 1950)</td>
</tr>
<tr>
<td>Huggett and Widdas</td>
<td>1951</td>
<td>foetal growth (Huggett &amp; Widdas 1951)</td>
</tr>
<tr>
<td>Weibull</td>
<td>1951</td>
<td>survival probability for aging (Weibull 1951)</td>
</tr>
<tr>
<td>Best</td>
<td>1955</td>
<td>diffusion limitation of uptake (Best 1995)</td>
</tr>
<tr>
<td>Smith</td>
<td>1957</td>
<td>embryonic respiration (Smith 1957)</td>
</tr>
<tr>
<td>Leudeking and Pret</td>
<td>1959</td>
<td>microbial product formation (Leudeking &amp; Pret 1959)</td>
</tr>
<tr>
<td>Holling</td>
<td>1959</td>
<td>hyperbolic functional response (Holling 1959)</td>
</tr>
<tr>
<td>Marr and Pirt</td>
<td>1962</td>
<td>maintenance in yields of biomass (Marr et al. 1963)</td>
</tr>
<tr>
<td>Rahn and Ar</td>
<td>1974</td>
<td>water loss in bird eggs (Rahn &amp; Ar 1974)</td>
</tr>
<tr>
<td>Hungate</td>
<td>1975</td>
<td>digestion (Hungate 1975)</td>
</tr>
<tr>
<td>Beer and Anderson</td>
<td>1997</td>
<td>development of salmonid embryos (Beer &amp; Anderson 1997)</td>
</tr>
</tbody>
</table>
mass maintenance; we treat the volume-specific structural and the full circles are priority allocation rules.

The mass description allows us to deal with mass conservation and the entropy description (not presented). The mass description allows us to deal with mass conservation, the energy description with energy conservation, and necessitates the development of auxiliary theory for such more advanced applications. The software package DEBtool (http://www.bio.vu.nl/thb/deb/deblab) is developed specifically for this purpose.

Table 5. List of symbols of compounds and processes.

<table>
<thead>
<tr>
<th>compound</th>
<th>process</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>feeding</td>
</tr>
<tr>
<td>E</td>
<td>reserve</td>
</tr>
<tr>
<td>V</td>
<td>structure</td>
</tr>
<tr>
<td>P</td>
<td>products</td>
</tr>
<tr>
<td>M_i</td>
<td>compound i</td>
</tr>
<tr>
<td>T</td>
<td>somatic maintenance (surface related)</td>
</tr>
<tr>
<td>S</td>
<td>somatic maintenance (total)</td>
</tr>
<tr>
<td>J</td>
<td>maturity maintenance</td>
</tr>
<tr>
<td>G</td>
<td>growth</td>
</tr>
<tr>
<td>R</td>
<td>reproduction or maturation</td>
</tr>
</tbody>
</table>

Figure 1. Metabolism in a DEB individual. Circles are processes and rectangles are state variables. Arrows are flows of food $j_{XE}$, reserve $j_{ED}$, $j_{EM}$, feeding $j_{EF}$, energy $j_{EG}$, and structure $j_{EG}$. The full square is a fixed allocation rule (the $\kappa$ rule) and the full circles are priority allocation rules.

Table 6 lists the state variables of the standard DEB model while tables 5–7 summarize the notation. Table 6 lists the state variables of the standard DEB model; we here use time ($T$), length ($L$), mass ($M$) and energy ($E$) to present the standard DEB model. Mass can be quantified in grams (which allows for changes in chemical composition) or C-moles (which does not allow for changes in composition); we here use the latter quantification. Strong homeostasis allows for simple relationships between quantification in volume, mass and energy because specific densities, molecular weights and chemical potentials, are constant for all compounds. The length description allows us to deal with shape of the structure. The basic variable is volumetric structural length $L_f$, i.e. the cubic root of structural volume. We need surface areas $L_f^2$ for food uptake and volumes $V = L_f^3$ for maintenance; we treat the volume-specific structural mass $[M_f] = M_f/V$ as a constant (strong homeostasis). The mass description allows us to deal with mass conservation, the energy description with energy conservation and the entropy description (not presented in table 6) with irreversibilities (Sousa et al. 2006). Entropy balances can only be made when energy balances are known, which in their turn can only be made when mass balances are known.

The total biomass of the individual (in C-moles) has contributions from reserve, structure and the reproduction buffer: $M_R + M_V + M_E$. Maturity has no mass or energy, it is information that reflects the level of metabolic learning; stage transitions (from embryo to juvenile to adult) occur when maturity exceeds threshold values. We quantify maturity as cumulated mass of reserve invested in maturity, but this invested mass dissipates into the environment as products (carbon dioxide, water, ammonia, heat).

None of the state variables can be measured directly, only indirectly (a problem known as hidden variables). This complicates the practical testability, and necessitates the development of auxiliary theory apart from core theory (that deals with mechanisms) to link measurements to model predictions (Kooijman et al. 2008). The solution of the problem of hidden variables is that a set of measured variables is linked to a set of hidden variables. This involves the estimation of a set of parameters from several datasets simultaneously, simulating the development of appropriate statistical theory for such more advanced applications. The software package DEBtool (http://www.bio.vu.nl/thb/deb/deblab) is developed specifically for this purpose.

The auxiliary theory exploits the strong homeostasis assumption that is also used by the core theory, together with the rule that a well-chosen physical length measure $L_f$ (e.g. the head–body length excluding a tail) is proportional to the volumetric structural length $L_s$, i.e. the cubic root for structural volume: $L = \delta_M L_s$ where $\delta_M$ is the constant shape coefficient, see Kooijman et al. (2008).

Figure 1 presents an overview of the various processes that are delineated by the standard DEB model. In our description of the various processes below, we assume that temperature is constant. In
the standard DEB model, all rates depend on temperature in the same way to avoid that conversion efficiencies (from food to reserve, structure, offspring, products) become temperature-dependent; multiple-reserve systems are more flexible in this respect.

(a) Reserve dynamics drives metabolism

The core of DEB theory is that metabolism is fuelled by the mobilization of reserve \( J_{EC} \) during all life stages (embryo, juvenile and adult); reserve being replenished by assimilation \( J_{EA} \) after the maturity threshold for birth has been passed,

\[
\frac{dM_r}{dt} = J_{EA} - J_{EC},
\]

with \( J_{EA} = 0 \) for embryos (they do not feed), i.e. maturity \( M_H < M_{rT} \). This not only explains why embryos can grow (i.e. increase structure) without feeding, but also why starving individuals can for some time survive and pay maintenance costs without dying (Sousa et al. 2008).

---

**Table 6.** List of symbols of variables. Dimensions: — dimensionless; \( T \) time; \( L \) length; \# mass in moles or C-moles; symbols with \( \{ \} \) are per unit surface area, \([\] \) are per unit of structural volume and dots above are per unit time. \( \phi = A, G, S, T, M, J, G, R \) and \( \theta = X, E, V \).

<table>
<thead>
<tr>
<th>state variable ( {M_V} )</th>
<th>dimensions ( {e, L^3} )</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_V )</td>
<td>( {M_V} )</td>
<td>structure</td>
</tr>
<tr>
<td>( M_E )</td>
<td>( {e} )</td>
<td>non-allocated reserve</td>
</tr>
<tr>
<td>( M_{ER} )</td>
<td>( {e} )</td>
<td>reserve in reproduction buffer</td>
</tr>
<tr>
<td>( \dot{q} )</td>
<td>( T^{-2} )</td>
<td>cumulated reserve allocated to maturation</td>
</tr>
<tr>
<td>( h )</td>
<td>( T^{-1} )</td>
<td>hazard rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>variable ( {t} )</th>
<th>dimensions ( {T} )</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t )</td>
<td>( {L^{-3}} )</td>
<td>time</td>
</tr>
<tr>
<td>( X )</td>
<td>( {L^{-3}} )</td>
<td>substrate density</td>
</tr>
<tr>
<td>( m_E )</td>
<td>( {#} )</td>
<td>scaled reserve density</td>
</tr>
<tr>
<td>( e = m_E/m_{Em} )</td>
<td>( {#} )</td>
<td>scaled reserve density</td>
</tr>
<tr>
<td>( L )</td>
<td>( {L} )</td>
<td>volumetric structural length ( V^{1/3} )</td>
</tr>
<tr>
<td>( f )</td>
<td>( {#} )</td>
<td>scaled functional response</td>
</tr>
<tr>
<td>( \dot{J}_h )</td>
<td>( {#} )</td>
<td>mass flow associated with process ( \phi ) and compound ( \theta )</td>
</tr>
<tr>
<td>( R )</td>
<td>( {eggs T^{-1}} )</td>
<td>reproduction rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>parameter ( {\dot{J}_{Em}} )</th>
<th>dimensions ( {L^{-2} T^{-1}} )</th>
<th>interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( {\dot{J}_{Em}} )</td>
<td>( {L^{-2} T^{-1}} )</td>
<td>surface-specific maximum searching rate</td>
</tr>
<tr>
<td>( {\dot{J}_{Em}} )</td>
<td>( {L^{-3}} )</td>
<td>surface-specific maximum assimilation rate</td>
</tr>
<tr>
<td>( {M_{Em}} )</td>
<td>( {M_{Em}} )</td>
<td>maximum reserve density</td>
</tr>
<tr>
<td>( {M_{Em}} )</td>
<td>( {M_{Em}} )</td>
<td>maximum molar reserve density</td>
</tr>
<tr>
<td>( {\dot{J}_{EM}} )</td>
<td>( {\dot{J}_{EM}} )</td>
<td>volume-specific maintenance rate</td>
</tr>
<tr>
<td>( {\dot{J}_{ET}} )</td>
<td>( {\dot{J}_{ET}} )</td>
<td>surface-specific maintenance rate</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {\dot{J}_{VT}} )</td>
<td>yield of reserve on substrate (food)</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {#} )</td>
<td>yield of structure on reserve</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {L T^{-1}} )</td>
<td>energy conductance</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {#} )</td>
<td>fraction of mobilized reserve allocated to soma</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {#} )</td>
<td>fraction of reserve allocated to reproduction that is fixed in eggs investment ratio</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {T^{-1}} )</td>
<td>somatic maintenance rate coefficient</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {T^{-1}} )</td>
<td>maturity maintenance rate coefficient</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {L} )</td>
<td>maximum length</td>
</tr>
<tr>
<td>( {\dot{J}_{VT}} )</td>
<td>( {L} )</td>
<td>heating length</td>
</tr>
<tr>
<td>( M_{rT} )</td>
<td>( {#} )</td>
<td>threshold of maturity at birth</td>
</tr>
<tr>
<td>( M_{rT} )</td>
<td>( {#} )</td>
<td>threshold of maturity at puberty</td>
</tr>
</tbody>
</table>

---

Phil. Trans. R. Soc. B (2010)
The flux of mobilized reserve equals the sum of all metabolic activities, excluding feeding (and assimilation)

\[ \dot{J}_{EC} = \dot{J}_{ES} + \dot{J}_{EG} + \dot{J}_{EI} + \dot{J}_{ER} \tag{2.2} \]

i.e. somatic (S) and maturity maintenance (J), maturation (or reproduction, R) and growth (G).

In combination with the constraint that mobilization only depends on the amounts of reserve and structure, weak homeostasis implies that the mobilization rate is (see Sousa et al. 2008)

\[ \dot{J}_{EC} = M_E \left( \frac{\dot{v}}{L - \dot{r}} \right). \tag{2.3} \]

where specific growth rate \( \dot{r} = (1/V) dV/dt \) and structural length \( L \) can vary (equation 2.10), but energy conductance \( \dot{v} \) is constant. The residence time of ‘molecules’ in the reserve is \( t_w = M_E/\dot{J}_{EC} \), so the energy conductance for a fully grown individual (\( \dot{r} = 0 \) at \( L = L_{\infty} \)) equals \( \dot{v} = L_{\infty}/t_w \). This relationship provides a simple interpretation of energy conductance as a parameter. The independence of the reserve dynamics of food availability provides the individual with some protection against environmental fluctuations and some control over its own metabolism; \( \dot{J}_{EA} \) typically varies wildly, but \( \dot{J}_{EC} \) always varies slowly.

With the combination of equations 2.1 and 2.3 and the definition of surface-specific maximum assimilation rate \( \dot{J}_{EA} = \langle \dot{J}_{EA} \rangle V^{2/3} \), the dynamics of the reserve density \( m_E = M_E/M_V \) amounts to:

\[ \frac{dm_E}{dr} = \frac{1}{M_V} \frac{dM_E}{dr} - \dot{r} m_E = \left( \frac{\langle \dot{J}_{EA} \rangle}{\dot{v}[M_V]} - m_E \right) \tag{2.4} \]

which is independent of the specific growth rate \( \dot{r} \). If assimilation is at the maximum \( \langle \dot{J}_{EA} \rangle = \langle \dot{J}_{EAm} \rangle \) then the reserve density \( m_E \) goes to a maximum value. This

\[ m_{EAm} = \frac{\langle \dot{J}_{EAm} \rangle}{\dot{v}[M_V]} \tag{2.5} \]

is independent of the organism body size; only the embryo can exceed this maximum, because it obtained its reserve from the mother.

It turns out to be convenient to introduce the scaled reserve density \( e = m_E/m_{EAm} \); this dimensionless quantity varies between 0 and 1.

(b) Feeding and assimilation

Embryos do not feed; only juveniles and adults. Feeding only depends on substrate (food) density and amount of structure, not partaking in the other metabolic interactions. The heat increment of feeding suggests that there are processes associated with food processing only, i.e. that food goes through a set of chemical reactions that transform it into reserves (Sousa et al. 2008). This is the assimilation process that is characterized by a yield coefficient \( \eta_{EX} \) of reserve on food that is assumed to be constant (for a given type of food) owing to the strong homeostasis assumption.

Food uptake \( \dot{J}_{XA} \) at food density \( X \) is linked to assimilation \( \dot{J}_{EA} \) as

\[ \dot{J}_{XA} = \frac{\dot{J}_{EA}}{\eta_{EX}} = f(X) \frac{\dot{J}_{EAm}}{\eta_{EX}} \tag{2.6} \]

where the scaled functional response \( f(X) = X/(X+K) \) is a monotonous increasing function of food density \( X \) with \( 0 \leq f(X) \leq 1 \), with half saturation constant \( K = \langle \dot{J}_{EAm} \rangle/\langle F_m \rangle \), where \( \langle F_m \rangle \) is the maximum specific searching rate. The scaled functional response \( f \) results from the idea that the individual behaves as a synthesizing unit (SU; Kooijman 1998) with two sequential behavioural modes: searching and handling (including digestion and other metabolic work). Many extensions of this idea have been proposed. The original formulation of the behaviour SUs is stochastic but the standard DEB model only uses the mean feeding rate.

The maximum surface-specific assimilation rate \( \langle \dot{J}_{EAm} \rangle \) is assumed to be constant, which is explained by the fact that digestion and other food processing activities depend on mass transport processes that occur through surfaces.

At constant food density, the reserve density evolves to \( m_E = f \left( \langle \dot{J}_{EAm} \rangle / \langle M_V \rangle \right) \dot{v} = f m_{EAm} \) (equation 2.5), which is independent of size and proportional to the scaled functional response. The scaled reserve density \( e = m_E/m_{EAm} \) equals the scaled functional response \( f \) in equilibrium.

(c) Allocation

DEB’s \( \kappa \)-rule for the allocation of mobilized reserve states that there is a constant fraction \( \kappa \), with \( 0 \leq \kappa \leq 1 \), of mobilized reserve that is allocated to the soma (somatic maintenance and growth), i.e.

\[ \dot{J}_{ES} + \dot{J}_{EG} = \kappa \dot{J}_{EC} \]

and

\[ \dot{J}_{EI} + \dot{J}_{ER} = (1 - \kappa) \dot{J}_{EC}. \tag{2.7} \]

Somatic maintenance has priority over growth (i.e. increase in structure) and maturity maintenance has priority over maturation or reproduction. The ultimate size an individual can reach directly results from the competition between somatic maintenance and growth. Reproduction and growth do not compete directly with each other, which explains why they can occur simultaneously; as listed in the stylized empirical facts in table 2.

Static and dynamic generalizations of the \( \kappa \)-rule allow for the accurate description of the growth of body parts (including tumours), and the relationship with energetics. In particular fields, such as in fisheries research, it is standard to let growth directly compete with reproduction dynamically. This can be done by allowing \( \kappa \) to be a function of structure. The partitionability requirement for reserve dynamics (which is implied by weak homeostasis) allows this dependence (Kooijman 2000; Sousa et al. 2008). However, the dependence of \( \kappa \) on structure makes \( \kappa \) a design parameter implying that the maximum surface area specific assimilation rate can no longer be proportional to maximum length (Sousa et al. 2008). The
consequence is that scaling relationships such as the interspecific Kleiber’s Law would be lost as implied properties of the model. Moreover, if $\kappa$ would depend on size, the von Bertalanffy growth curve no longer applies at constant food density. This empirical evidence together with the fact that the inverse von Bertalanffy growth rate increases linearly in the ultimate length (table 2) is strong support of the assumption that $\kappa$ is generally constant.

(d) Somatic and maturity maintenance

The need to allocate energy to maintenance is intimately related with the second law of thermodynamics because the level of maturity, i.e. complexity of the organism, would decrease in the absence of energy spent on its maintenance.

Somatic maintenance is the use of reserve to fuel the set of processes that keep the organism alive, where $J_E$ and $J_M$ are the reserve flows allocated to volume, e.g. protein turnover, and to surface maintenance costs, e.g. heating in endotherms:

$$J_{ES} = J_E + J_M = [J_E]L^3 + [J_M]L^2. \quad (2.8)$$

The volume-specific somatic maintenance costs $[J_E]$ are assumed to be constant; the turnover of structure comprises a big proportion of these costs, but they also include activity, for instance. The surface-specific somatic maintenance costs $[J_M]$ are only positive for particular taxa (endotherms and osmotic work for freshwater species). It is convenient to introduce the heating length $L_T = [J_E]/[J_M]$. This turns out to be the reduction in ultimate length owing to surface-linked somatic maintenance.

Ultimate length $L_\infty$ (when $\dot{r} = 0$) follows from the balance between assimilation and maintenance and does not depend on growth. Growth ceases if $k_{EC} = J_{ES}$ (cf. equation 2.7). Using equation 2.3 and 2.5, the result is $L_\infty = \mu_{m} - L_T$ with maximum length $L_m = \kappa [J_{EM}]/[J_E]$. Reserve is assumed to require no maintenance, as empirically supported by the fact that freshly produced eggs almost exclusively consist of reserve and hardly respire (see Sousa et al. (2008) for a detailed explanation). Reserves do not need turnover; they have a limited residence time owing to assimilation and mobilization. In fully grown individuals, the residence time amounts to $t_E = L_\infty/\dot{v}_E$, but it is shorter in smaller individuals. This explains why babies need to feed more frequently than adults.

Maturity maintenance is the use of reserve to maintain the complexity of the structure where $J_M$ is the reserve flow allocated to this process and $k_M$ is the maturity maintenance rate coefficient:

$$J_M = \dot{k}_M M_M. \quad (2.9)$$

The $J_M$ is constant in adults since for them maturity constant, $M_M = M_M^0$. 

(e) Growth

Growth is the increase of structure; the specific growth rate $\dot{r}$ follows from the reserve dynamics (equation 2.3), the $\kappa$-rule (equation 2.7) and the somatic maintenance costs (equation 2.8). The result is

$$\frac{dM_V}{dr} = J_{VG} = J_{EC} Y_{VE} = \dot{r} M_V \quad \text{with}$$

$$\dot{r} = \frac{\dot{v} e L - (1 + L_T/L)/L_m}{e + g}, \quad (2.10)$$

where $J_{EC}$ is the mobilized reserve allocated to growth and $Y_{VE}$ is the yield of structure on reserve. Maximum length $L_m$, heating length $L_T$ investment ratio $g$ are all given in table 7. Now the specific growth rate $\dot{r}$ is specified, the mobilization rate $J_{EC}$ in equation 2.3 is specified as well, so is the residence time $t_E$ of ‘molecules’ in the reserve during ontogeny.

For any constant food level, the scaled reserve density $e$ settles at the level of the scaled functional response $e = f$ and the dynamics of structural length $L = v^{\alpha/3} = (M_V/M_M)^{1/3}$ simplifies to von Bertalanffy growth for juveniles and adults:

$$\frac{dL}{dt} = r_E (L_\infty - L) \quad \text{with}$$

$$r_E = \dot{k}_M / 3 = \frac{1}{3/\dot{v}_M + 3/L_m/\dot{v}_E}, \quad (2.11)$$

where the somatic maintenance rate coefficient $k_M$ is given in table 7. The inverse of the von Bertalanffy growth rate $r_E$ is thus linearly increasing with ultimate length, as listed in the stylized empirical facts in table 2.

If allocation of reserve to soma is not sufficient to pay the somatic maintenance costs, structure can shrink:

$$\frac{dM_V}{dr} = - J_{VS} \left(1 - \frac{\min(J_{ES}, J_{EC})}{J_{ES}} \right) \quad \text{with}$$

$$J_{VS} = [J_{VM}]L^3 + [J_{VT}]L^2, \quad (2.12)$$

where the somatic maintenance costs $J_{VS}$, if paid from structure, have the same set-up as those paid from reserve (equation 2.8). A natural simplification is to assume that $[J_{VM}]/[J_E] = (J_{VT})/(J_{ET})$, but this ratio should be larger than one for thermodynamic reasons. Death by starvation occurs if structure, relative to the maximum the individual once had, decreases below a minimum. This minimum fraction is for supply systems typically smaller than for demand systems, but even for demand systems, empirical support for shrinking exists (Genoud 1988). Most species seem to avoid shrinking, e.g. by using the reproduction buffer to cover the somatic maintenance costs.

In extreme cases, species can sport suicide reproduction, and convert part of their structure to gametes before dying.

(f) Maturation and reproduction and initial state of the individual

Maturation is the use of reserve $J_{ER}$ to increase the level of maturity, $M_M$. This level controls qualitative changes in metabolism (life-stage events). The initiation of feeding occurs at birth when $M_M = M_M^b$. The initiation of allocation to reproduction occurs at puberty when $M_M = M_M^p$ it is coupled to the ceasing
of maturation. Other life-history events, such as cell division, metamorphosis or other stage transitions (e.g. to the pupal stage) occur also at threshold values for $M_H$.

Multicellular organisms typically have three life stages: embryo, juvenile, adult. At the start of development, age $a$ is set to zero, structure $M_I$ and maturity $M_H$ are zero, $M_E^0 = 0$ and $M_H^0 = 0$, and the initial amount of reserve $M_R^0$ is such that the reserve density $m_g$ at birth equals that of the mother at egg formation; the maternal effect as listed in the stylized empirical facts in Table 2. This fully specifies $M_E^0$ for an efficient algorithm to obtain $M_E^t$, see Kooijman (2009). Dividing unicellulars are treated as juveniles; division of maturity follows that of structure and division occurs if maturation exceeds a threshold.

The allocation to maturity (in embryos and juveniles) or reproduction (in adults) is

$$\dot{J}_{ER} = (1 - \kappa) \dot{J}_{EC} - k_J M_H.$$  \hspace{1cm} (2.13)

The change in maturity (in embryos and juveniles) is given by

$$\frac{d}{dt} M_H = \dot{J}_{ER} \frac{k_J}{k_I} \text{ if } M_H < M_H^0.$$  \hspace{1cm} (2.14)

where $k_J^t = k_J$ if $\dot{J}_{ER} > 0$, but for shrinking maturity (rejuvenation), it is a free constant parameter. Empirical evidence for rejuvenation induced by starvation is presented in Thomas & Ikeda (1987). The hazard rate owing to starvation is proportional to the difference between the maximum maturity level that the individual has reached and the actual level.

The reproduction buffer fills at rate $\dot{J}_{ER}$ for $M_H = M_H^0$. The details of the conversion of the reproduction buffer of females to a number of eggs are rather species-specific, typically including requirements on temperature and filling of the buffer; the conversion of the reproduction buffer of males to sperm is typically linked to female reproductive behaviour. The simplest buffer handling rule is to produce an egg as soon as the reproduction buffer allows; this rule involves no new parameters. The conversion of the content of the reproduction buffer to one or more eggs involves an overhead cost of the reproduction process, i.e. a fraction $(1 - \kappa G)$ of the converted buffer (and so of the reserve allocated to reproduction $\dot{J}_{ER}$) dissipates, and a fraction $\kappa G$ is fixed into eggs. The cost per egg equals the initial amount of reserve $M_R^0$.

The reproduction rate in terms of numbers of eggs per time, is a delta function of time. Ignoring the effect of the reproduction buffer, and treating reproduction as a continuous process, the reproduction rate would amount to $R = \kappa G \dot{J}_{ER} / M_R^0$.

Foetal development is a variation on egg production, where the mother does not fill a reproduction buffer, but directly adds to the reserve of the foetus, bypassing its digestive system. This process can, therefore, not be seen as a feeding process from the foetal perspective.

(g) Three organizing fluxes in metabolism

An implication of strong homeostasis is that the different types of aggregated chemical reactions occurring in the organism have constant stoichiometries. These reactions are assimilation ($X \rightarrow E + P$), growth ($E \rightarrow V + P$) and dissipation ($E \rightarrow P$), where dissipation is defined as

$$\dot{J}_{ED} = \dot{J}_{ES} + \dot{J}_{EJ} + (1 - \kappa_R) \dot{J}_{ER}.$$  \hspace{1cm} (2.15)

and $\kappa_R = 0$ for the embryo and juvenile stages. Thus, metabolic transformation has three degrees of freedom; the flow of any compound (e.g. dioxygen), produced or consumed, in the organism is a weighted sum of these three organizing flows. The method of indirect calorimetry (Table 3) is a particular case: the flow of heat is a weighted average of the fluxes of carbon dioxide, dioxygen and nitrogenous waste. Since, reserve is key to the ability to delineate these three fluxes (without reserve we would have two), the empirical success of the method of indirect calorimetry gives strong support to the topology of the standard DEB model.

(h) Ageing

The hazard rate $h$, i.e. the probability of dying, owing to ageing, is taken to be proportional to the density of damage compounds (e.g. modified proteins):

$$\frac{d}{dt} h = \bar{q} - \bar{r} h,$$  \hspace{1cm} (2.16)

where $\bar{r} h$ is the dilution by growth and $\bar{q}$ the change in ageing acceleration, which is proportional to the density of damage-inducing compounds (e.g. changed mitochondrial DNA). Damage compounds are generated by damage-inducing compounds at a rate proportional to the metabolic activity measured by the reserve mobilization rate (equation 2.3). The production of damage-inducing compounds is again taken to be proportional to the reserve mobilization rate (as quantifier for the respiration rate, excluding contributions from assimilation, which are supposed to have local effects only).

The change in ageing acceleration is given by

$$\frac{d}{dt} \bar{q} = \left( \frac{L^3}{L_m^{15/9}} \bar{q} + h_a \right) \left( \frac{\bar{q}}{L} - \bar{r} \right) - \bar{r} \bar{q},$$  \hspace{1cm} (2.17)

where $\bar{r} \bar{q}$ is the dilution by growth and the factor $e(\bar{q}/L - \bar{r})$ is proportional to the mobilization rate, cf. equation 2.3. The proportionality factor $(\bar{q}/L^3/L_m^{15/9}) h_a$ increases linearly with $\bar{q}$ because damage-inducing compounds promote their own production. This expression involves two new parameters, the Weibull ageing acceleration $h_a$ and the Gompertz stress coefficient $h_a$. The latter parameter is close to zero for most ectotherms, but for endotherms it is typically positive. It can be shown that if the growth period is short relative to the life span, both the Weibull and the Gompertz ageing models result, Table 4. For further details on ageing see van Leeuwen et al. (2010).
(i) Parameters

Each individual is characterized in DEB theory by a set of primary parameters: the surface-area specific searching rate (feeding) \( \{F_m\} \), the surface-area specific maximum assimilation rate (assimilation) \( \{J_{Edm}\} \), the yield of reserve on food (digestion) \( \nu_{Ed} \) and of structure on reserve (growth) \( \nu_{SM} \), the energy conductance (mobilization of reserve) \( \nu \), surface and volume-specific somatic maintenance costs \( \{L_E\} \) and \( \{L_SM\} \), the specific maturity maintenance \( \kappa_M \), the fraction of mobilized reserve allocated to soma \( \kappa_s \), the reproduction efficiency \( \eta_M \), the maturity threshold levels that trigger the onset of feeding and reproduction \( M^g \) and \( M^p \), the Weibull ageing acceleration \( h_a \) and the Gompertz stress coefficient \( g \). This amounts to 14 primary parameters including the two ageing parameters and excluding parameters for species-specific handling rules for the reproduction buffer. The details of death by starvation involve another four parameters; these can be avoided by letting the individual die upon shrinking or starvation-induced rejuvenation. One can argue about the status of the mass–volume coupler \( [M^j] \); this parameter relates measurements with no impact on processes.

The standard DEB model is meant to be the simplest in the DEB family that still has all essential features, a canonical form. Many applications need extensions of various types; the specification of respiration, for instance, requires the elemental composition of various compounds (food, faeces, reserve, structure). We agree with Nisbet et al. (2010) that some other applications, such as in population and ecosystem dynamics, require simplifications.

Most applications allow setting \( \kappa_M = 1 \) and \( \kappa_s = \kappa_M^s \); the latter equality implies that maturity density, \( M^g \) and \( M^p \) remains constant and metabolic switches occur at fixed amounts of structure. This means that the maturity thresholds can be replaced by structure thresholds and maturity can be avoided as state variable. If reproduction occurs with one offspring at a time, the reproduction buffer can be avoided as state variable. If investment in heating (or osmosis) is small, we have \( \{J_{ET}\} = 0 \) and \( \nu_M = 0 \). Ageing is not always an important cause of death in field situations; the ageing acceleration and the hazard rate can be avoided as state variables under those conditions, and the two ageing parameters are lost. All simplifications together reduce the standard DEB model to two state variables (reserve and structure) and nine parameters, while it still covers a full specification of feeding, digestion, maintenance, development, growth and reproduction over the full life cycle of the individual. This amounts to 9/6 = 1.5 parameter per process; we think a remarkable simplicity. Typical applications involve only a subset of these parameters because they do not involve all processes.

(j) Covariation of parameter values

A rough estimation of DEB parameters for each species can be made with the scaling relationships, i.e., based only on the species maximum size and a reference species; the accuracy of this estimation increases with the similarity between the species. The design parameters are \( \{J_{Edm}\} \) and \( \bar{a}_s \), which scale with maximum length, and \( M^g_H \) and \( M^p_H \), which scale with maximum volume. All the other primary parameters are independent of size, so also independent of the maximum size of a species. The body-size scaling relationships can also be used partially, making optimal use of all data at hand (Kooijman et al. 2008; Kooijman 2010) to detect species-specific deviations from the general trend.

These rules also determine how properties that can be written as functions of the primary parameters depend on maximum length. An example is the respiration rate (i.e. the use of dioxygen). It works out to be approximately proportional to weight to the power 3/4, both inter- and intraspecifically (see the list of stylized facts in table 2), but for very different reasons. The weight-specific respiration rate decreases intraspecifically because growth decreases (and so the contribution of overhead costs of growth to respiration); it decreases interspecifically in fully grown adults because reserve density increases with the maximum size of a species and somatic maintenance is only paid for structure. The explanation offered by DEB theory also allows to understand taxon-specific variations in the scaling of respiration, since quite a few parameters contribute to the result and evolutionary adaptations cause deviations of parameters from the pattern. Many alternative attempts to explain the scaling of respiration fail to distinguish between intra- and interspecific comparisons, probably owing to the similarity of the numerical behaviour.

Many scaling relationships work out differently for intra- and interspecific comparisons. Feeding scales with surface intraspecifically, but with volume interspecifically, while maximum reproduction increases with size intraspecifically but decreases with size interspecifically. The remarkable prediction for life span is that it scales with length if the Gompertz stress coefficient is positive (as expected for endotherms), but hardly scales with length if it is zero (as expected for ectotherms), which is consistent with empirical data (table 2).

3. CURRENT RESEARCH TOPICS IN DYNAMIC ENERGY BUDGET THEORY

(a) The sub-individual level

The links that DEB theory establishes between the sub- and supra-individual levels together with the high amount of throughput data becoming available at the sub-individual level have allowed the use, test and development of DEB models at this organization level (van Leeuwen et al. 2010; Peccquerie et al. 2010; Vinga et al. 2010).

Vinga et al. (2010) use DEB theory for a top-down approach to understand the dynamic behaviour of metabolites. The individual level controls the size of the reserve fluxes that are associated with assimilation, dissipation and growth: \( J_{ET}, J_{EG} \) and \( J_{EG} \). These fluxes impose constraints on the rates of aggregated chemical reactions and on the overall amount of each aggregated chemical compound at the sub-individual level. Vinga et al. compare the DEB approach with biochemical
systems theory (BST) in modelling \textit{in vivo} data of lactic acid bacteria under various conditions. In contrast with DEB theory, BST is a bottom-up approach that models each chemical compound and each chemical reaction explicitly. The complementarity between the two approaches is important in bringing new insights to unsolved problems that link the sub-individual and the individual levels such as the mechanisms underlying gene expression or the mechanisms underlying ageing (van Leeuwen et al. 2010).

Pecquerie et al. (2010) develop DEB theory further to provide a framework for stable isotope dynamics. The fundamental processes of the standard DEB model—assimilation, dissipation and growth—are further detailed into their anabolic and catabolic transformations to account for the mass balance of stable isotopes. Isotope dynamics reveal features that remain hidden in aggregate mass dynamics such as turnover rate of structure. This turnover process has a catabolic as well as an anabolic component. Since turnover has a substantial contribution to somatic maintenance, it is also of importance to energetics. This DEB module on isotope dynamics will allow for the correct interpretation of the increasing amount of data becoming available on isotope ratios contributing to the identification of trophic web structures, the reconstruction of individual life histories, and the tracking of the flow of elements through ecosystems.

van Leeuwen et al. (2010) review the DEB-based approaches to ageing and link them to current research at the molecular/cellular level. The authors link alternative ageing DEB-based modelling approaches with different cellular senescence processes. This is a first step towards a fundamental understanding of the link between mechanisms of cellular senescence and ageing, at the individual level.

(b) The individual level

The mass, energy and entropy description of all fluxes, provided by DEB theory, are most useful to study the internal concentration of specific compounds such as isotopes (Pecquerie et al. 2010), reactive oxygen species (van Leeuwen et al. 2010) and toxins (Jager & Klok 2010; Ducrot et al. 2010) that affect the performance of organisms at the individual level. The mode of action of a compound is, in the context of DEB theory, defined by the parameters that are affected. When the internal concentration increases, more and more parameters become affected, but at a sufficiently low concentration only a single parameter is significantly affected, but the consequences might be complex, involving feeding, growth and reproduction. For example, an increase in the specific maintenance rate, $\dot{F}_{EM}$, leads directly to a decrease in growth, and ultimately also to a smaller adult that reproduces less while the decrease in the efficiency of reproduction, $\kappa_R$, leads to a decrease in the rate of reproduction (Jager et al. 2004), but does not affect growth or feeding. Jager and Klok compare several DEB approaches for analysing the toxicity of copper in the earthworm \textit{Dendrobaena octaedra}: the Kooijman–Metz formulation (Kooijman & Metz 1984; which has no reserve or maturity), the DEBtox approach (Kooijman & Bedaux 1996; which has no explicit maturity) and the DEB3 approach (Kooijman 2010). Results on mortality and growth rate for the DEBtox and the DEB3 approach were similar. Also, they compare DEB based and empirical approaches concluding that only the former allow extrapolation for field relevant conditions. Ducrot et al. use the DEBtox model to assess the toxicity data of diquat on the gastropod \textit{Lymnaea stagnalis}, where they include data on embryo development, making full use of the life-cycle features of DEB theory in variable environments, which is crucial for environmental risk assessment.

(c) The population level

The step from the individual to the population level requires extra rules for the interaction between individuals and for the transport of resources in the environment. The simplest interaction rule for the standard DEB model is that individuals only interact via competition for resources. The standard bookkeeping technique to follow the performance of populations as collections of individuals is (hyperbolic) partial differential equations (PDEs). Diekmann & Metz (2010) present a wider mathematical framework that removes some of the shortcomings of PDEs in this context. The standard DEB model has some features, however, that still cause mathematical problems including the existence of metabolic events (birth, puberty) that lead to singularities in the equations, the fact that eggs are not infinitesimally small and last but not least, DEB is deterministic (apart from the survival module). Some of these problems can be removed. For example, SU-dynamics (which is used to specify feeding) is stochastic by nature and differences between individuals can be implemented using different parameter values.

Other problems, however, are rather fundamental and call for individual-based approaches or a fully stochastic framework. For example, when feeding on a single resource in homogeneous space, the DEB rules imply that small (young) individuals can rather easily outcompete the large (old) individuals, perhaps to an extent that is not very realistic. There are several DEB solutions to this problem. Kearney et al. (2010) consider that food quality required by the individual depends on size. Large individuals mainly have to cover their maintenance cost (reproduction is low at the carrying capacity, where competition is strongest) while small individuals need to grow. Energy for maintenance can be supplied by carbohydrates and lipids while energy for growth must be supplied by a protein-rich resource. This set-up favours the large (old) individuals. Nisbet et al. (2010) consider survival to be maturity-dependent; in juveniles, the probability of dying is much more dependent on the food level, which favours adults. With this feature they have been able to understand the occurrence of daphnid population oscillations under particular conditions, using a reduced version of the DEB individual dynamics. They beautifully illustrate that not all details are important under all conditions; if
food density is rather constant and different food levels are not compared, reserve and maturity typically play a minor role.

The step from the individual to the population levels can be done using a variety of schemes. Jager & Klok (2010) use DEB-structured individuals in a matrix and continuous Euler–Lotka population models to extrapolate toxic effects from individuals to populations. Kooi & van der Meer (2010) use a physiological-structured population model to describe the dynamics of a population in a semi-chemostat environment where reproduction is a discrete event process. In the case of organisms that reproduce by division, the transition from the individual to the population is simpler, because organisms can be considered as V1-morphs, i.e. individuals that change in shape during growth such that their surface area is proportional to volume. In this case, a population of a few big individuals behaves identically to that of many small ones if the sum of their masses match: the individual level completely drops out of the equations. This is the case in Lorena et al. (2010) and Poggiale et al. (2010), where population performance hence directly links to sub-individual physiology.

(d) Variable environments

Since DEB theory specifies the interaction between the individual and its environment dynamically, it has no problems at all with variable environments. This volume has a nice collection of examples, where these variations are explicitly used to study the underlying organization. Ducrot et al. (2010) use this feature when analysing the effects of a weed control agent when the concentrations vary in time; they show how DEB-based models can capture observed survival patterns where typical models fail. Lorena et al. (2010) model microalgal populations in a chemostat with a variable light regime and study how the biochemical composition of microalgae depends on light. More specifically they discuss the relationship between chlorophyll, biomass and the production of exopolymers: substances; these are key features in the interpretation of remote sensing data. Pecquerie et al. (2010) evaluate how variations in isotope concentrations in the environment work out for the organism. They do not make the common assumption that isotope dynamics is at equilibrium, and include the full metabolism in their analysis. Troost et al. (2010) adjust DEB individual models for cockles and mussels to a specific site by adjusting the functional response. With this model they detected food preferences in cockles and mussels, inferring the importance of detritus and intraspecific competition under field situations.

Kearney et al. (2010) position DEB theory in a wider ecological setting, linking it to the theories of biophysical ecology and the Geometric Framework for Nutrition. The combination of these fields stimulates the development of models at their interfaces that can shed more light on the detailed interaction of organisms and their environment. Biophysical ecology provides a framework for the climatic niche of an organism (distribution limits as constrained by heat and water balances) making use of spatial environmental data while the Geometric Framework for Nutrition provides a way to determine the nutritional niche of an organism (distribution limits as constrained by dietary needs) making use of information on the availability of food. These theories presently make use of allometrically derived static mass and energy budgets. By using DEB theory, these theories can model physiological rates across the life cycle under variable food and climatic environments and establish links between individuals and their functional traits and population. This is an essential step towards the goal of building predictive niche models that can tackle questions such as the impact of climate change on a species distribution.

(e) SUs: combining multiple mass and energy flows

The kinetics of SUs (Kooijman 1998, 2010) is an essential building block for the dynamics of multiple reserve and/or multiple structure organisms. SUs can be conceived as generalized enzymes that link the product fluxes to the arrival fluxes of substrate at the enzyme. Poggiale et al. (2010) interpret the different types of co-limitation in an SU context and demonstrate that these different types are of importance at the ecosystem level. Lorena et al. (2010) use SU-dynamics to model the co-limitation of photosynthesis by light and carbon dioxide and the co-limitation of growth by a carbon and nitrogen reserves. While Kearney et al. (2010) use SU-dynamics to transform food into separate nutrient reserve pools, and then regulate the assignment of reserves mobilized from each pool into maintenance, structure, maturity maintenance and reproductive output.

(f) Parameters

The generality of DEB theory allows the use of more parsimonious models (fewer parameters) to accurately describe experimental data under different environmental conditions. The DEB model for glycolysis in Lactococcus lactis uses much less parameters than a comparable BST model Vinga et al. (2010), while it better catches the differences between growth under aerobic and anaerobic conditions.

DEB rate parameters depend on temperature. Freitas et al. (2010) compare the temperature tolerance (this is the set of temperatures for which the Arrhenius relationship applies) and temperature sensitivity for a variety of marine species. Their results suggest that the width of the temperature tolerance range increases with the optimal growth temperature. Differences in life-history strategies of related species translated nicely in differences in parameter values: high-optimal growth temperatures, large tolerance ranges and high sensitivities are linked to low-specific assimilation rates and low-specific maintenance costs.

Although DEB theory does not use any optimization argument, it remains thought-provoking to study to what extent parameter values, or life-history traits, are optimal, or at least could be seen as the outcome of an evolutionary optimization process. The

Phil. Trans. R. Soc. B (2010)
theory of adaptive dynamics is ideal for this, because there is no need to specify any explicit optimization criterion. Moreover, it includes the interactions between organisms and their environment and the long-term consequences of changes in traits in a natural way; the outcome depends on the realism of the ecosystem model. Kooi & van der Meer (2010) study the handling rules of the reproduction buffer of _Macoma_ under seasonal forcing. They successfully capture the observed spawning behaviour of this iteroparous species, which spawns once a year; although, they were not able to correctly predict the timing of spawning. Application of adaptive dynamics with seasonal forcing is an impressive tour de force; more research is required to understand why _Macoma_ spawns in spring. Moreover, Kooi & van der Meer (2010) demonstrate that the techniques that they use are a special case of bifurcation theory (Troost et al. 2007), which can lead to cross-fertilization.

Although a lot has been done already, the development of DEB theory has only started. Almost all contributions in this volume illustrate this in different ways. Future developments should include extensions in the sub-individual level, e.g. at the molecular level, and in the supra-individual level, e.g. the development of DEB-based biogeochemical climate models (Kooijman 2004). There are also open issues at the individual level such as the development of behavioural extensions, the formalization of multiple reserve and multiple structure organisms and DEB models for plants. The behavioural timescale including food searching, food selection, sleeping, social interaction, parental care, etc., is particularly important for animals and humans. Quite a few behavioural extensions have already been proposed using SU-dynamics but they have to be explored more systematically. Also, DEB models for plants have been proposed (Kooijman 2010), but they need to be tested against data; which has not been easy because detailed studies in plant biology are not available.

4. CONCLUDING REMARKS

Given the richness of biodiversity on Earth, general explanatory models have to be lean, capturing taxon-specific phenomena in modules that extend the non-taxon-specific core. For particular applications (e.g. in ecosystem dynamics), the standard DEB model will be too complex, for other applications (e.g. in medicine and molecular biology) not detailed enough. This directly relates to the timescales of interest. Simplifications as well as extensions should be done, respecting a natural order in timescales, where the standard DEB model deals only with the slowest processes at the individual level. It makes little sense to include very fast processes if slower processes are not included. Extensions should be consistent with the existing assumptions.

Balancing realism at a detailed level against simplicity (in terms of numbers of parameters and variables) depends on subjective judgement and context. Although the standard DEB model is simple relative to the complexity of biological reality, estimating its parameters on the basis of published data is a challenge. Extensions make this problem worse, not easier, and we believe that obtaining accurate estimates for the primary parameters should generally have priority over extensions. We made a systematic start in the add_my_pet program (http://www.bio.vu.nl/thb/deb/deblab) and hope that the collection extends rapidly and improves in quality. We hope that a new generation of scientists will collect data in the light of DEB theory that allows the accurate estimation of its parameters and further critical testing of the underlying assumptions.

We thank the authors of this issue for the valuable feedback that we received on drafts of this paper. This research was supported by FCT through Grant no. PPCDT/AMB/55701/2004.

REFERENCES


Russell, N. R. & Wootton, R. J. 1992 Appetite and growth compensation in the European minnow, Phoxinus phoxinus (Cyprinidae), following short periods of food

Phyl. Trans. R. Soc. B (2010)
Chapter 18:
Joana Mendonça and Manuel Heitor, “Industrialization, geography and policy: the changing pattern of industrial production”
Industrialization, Geography and Policy:
The changing patterns of industrial production

Joana Mendonça and Manuel Heitor

Center for Innovation, Technology and Policy Research, IN+
Instituto Superior Técnico, Technical University of Lisbon

Abstract

Changing patterns of industrial production at a world level are discussed in terms of shaping factors for technical change and related impacts on skilled employment and advanced training of human resources. This is because industrialization has been the main driver behind rapid productivity growth achievement and social well-being improvements in different world regions in the last 200 years. Despite this, there is a clear difference in industrialization levels across and within different regions, with the world’s geographical patterns of production changing over the last decades.

It is well known that the weight of manufacturing in the economy has been decreasing substantially in many countries, which has become service-based economies, with the notable exception of Germany, Korea and China. This has happened together with an increasing democratization of the access to higher education and the opening-up of the science base in many world regions. On the other hand, the geographical concentration of industrial production have facilitated many of those regions to increasingly loose their productive ability, leading to changes in employment structure and the mobility of skilled people. This raises new concerns for science and industrial policies, which require to be discussed in terms of new forms of stakeholder engagement. Our analysis focus on patterns of knowledge production and diffusion, looking specifically at the role of new technology based firms towards revisited industrial policies.

1. Introduction: Innovation and socio-economic resilience

It is clear that technoglobalism and the globalization of trade and supply chains led to the emergence of increasingly competitive global markets and to facilitated access to new suppliers, independently of their geographic location (Berger, 2005). This has allowed countries and regions with strong technological and industrial bases to profit on the lowering of trade barriers to access new markets, while the majority of firms located in other regions remained confined to local markets.

In addition, the analysis of the overall trend on moving towards knowledge intensive services and its relation with job creation and economic growth requires some pragmatism. This is because, in parallel
to technoglobalism came post-industrialism, promoting services, and at the same time the new developed countries’ economic growth overtook manufacturing industries. Captivated by the prospects of accelerated and cost-effective economic growth, many countries started shifting their focus from manufacturing industries to knowledge-intensive services (Hepburn, 2011; Ghani and Kharas, 2010).

The result is emerging with many regions worldwide lagging behind. In fact, evidence shows that, when compared to knowledge-intensive services, hard industries have higher labor productivities, a more balanced income distribution, higher income growth rates and the ability to generate exports, which are negligible in the case of services (Nairn, 2002; Fingleton, 1999).

Looking at the US, together with other most developed economies (including Germany), we can identify some common factors, that need to be understood in international comparative terms: strong industrial bases, diversified economy, and supply chain and knowledge networks’ complexity (Amsden, 2001; Hidalgo and Hausmann, 2009).

It should be clear that a new generation of industries will drive the economic recovery over the next decade, fuelled by long-term changes in technology, society and geopolitics. The recession has not been only a point of change, and many argue that it has acted as a catalyst for growth. As the business landscape alters, we will see the emergence of new ways of doing business in an increasingly interconnected world.

This paper presents an analysis of the geography of production in the world, and parallel evolution of scientific and technological bases. In addition, it looks at the structure and availability of human resources and competences, as the basis for industrial activity.

The analysis presented in this paper is the starting point of a broader research framework to be developed focusing on industrialization. It is our understanding that under this framework, research should look at how industry is distributed, and how production is done across countries and regions, and how capabilities are being used in different countries and regions. In fact, this framework is leading to several research programs in centers and universities around the world, including the “Production in the Innovation Economy” project at the MIT (Berger, 2005), and leading to a significant body of research in Brazil on industrialization, namely through regulation and local content policies (Apolinário e Silva, 2011; Cassiolato et al., 2008; Salles-Filho et al., 2010).

This paper starts by looking and industrialization and employment in section two, where we show the evolution of industrial activity in the world. Section three presents the background on the role of knowledge capabilities and R&D in the development of industries. In section four shifts the analysis towards people and capabilities, followed by section 5 where we address the role of new firms in the development of new industrial activity. The last section presents a summary.

2. Industrialization and employment

Industrialization has been the main driver behind rapid productivity growth achievement and social well-being improvements in different countries in the last 200 years (Murphy et al., 1989). Over the
last century, the world’s industrial landscape has been changing showing a shift towards emerging Asian and Latin American economies in the last decades. Industries have emerged, evolved and migrated to new countries. This transformation reflects the emergence of China, South Korea, India, Brazil, Mexico and other developing countries as major economic powers (Hepburn, 2011). The Asian shift is not new, as more than 50% of the world’s industrial trade was produced in China and India since 1750. This growth of emerging countries is affecting the world’s geography of industry and pattern of industry employment.

As a consequence, there is an increasing concern that deindustrialization is hampering growth and undermining the competitiveness of developed economies leading to loss of jobs (Wessner and Wolff, 2012). In many developed regions recent deindustrialization trends resulted not only from the mentioned emergence of new and fast-expanding markets, the surge of new pools of cost-efficient workforces, but also from the emerging countries’ aggressive development policies.

In many economies the weight of manufacturing in the economy has been decreasing substantially over the last decades, followed by an increase in services, and production has been concentrating in certain regions, while others have been increasingly losing their productive ability. In agreement, policies in the last 20 years have been favoring the development of knowledge intensive services oriented towards short-term growth and the financial markets (Johnson, 2002). Parallel to this process of “technoglobalism”, post-industrialism advocates have promoted in the last decades services as a new panacea for rapid economic growth in developed - and developing - countries, overtaking the role of manufacturing industries as drivers of growth (Bell, 1973; Kelly, 1998; Naisbitt, 1982). In an increasing manner, growing emerging economies are shifting to higher added value components of international supply chains, while their physical, human and institutional capital strengthens, and begin to compete directly with the more developed ones (Spence and Hlatshwayo, 2011).

This trend can be observed looking for example at China’s exports manufacturing structure, which has been shifting in a similar way to Japan and Korea did, going from labor-intensive products, to more capital intensive products, eventually to move to human-capital intensive activities. Figure 1 shows the evolution of exports for China and Korea for the period 1995-2009. China is in a catching-up phase, and it is making a clear move out of apparel, textiles, footwear, and toys and a move into electrical machinery, telecom, office machines, and to a lesser extent metals, leading to a very different product export distribution from the early 1990s (Amiti and Freund, 2007). Korea’s pattern is already different, as it had been catching up before, by developing their industrial capacity first, and later their science and technology (Hollanders and Soete, 2010).
A similar pattern is observed on the evolution of high-tech exports, which are represented in figure 2 for China, Brazil, Japan, China, Korea, European Union and the United States. As the figure illustrates, high tech sectors are gaining weight in exports in China and Korea, decreasing significantly in the US, slightly in Japan, and maintaining in other regions, as signs of delocalization of production, and change in the productive structure of the countries.

Figure 2 - Evolution of high tech exports- 1988-2011

Source: Data from the World Bank
The shift in production illustrated should have clear impacts on the employment structure of countries and regions. In the US, the weight of manufacturing in the job distribution has decrease around 40% since the 1980’s, as the number of U.S. manufacturing jobs declined from 20 million in 1979 to about 12 million 30 years later (OECD, 2010 and Tassey, 2010). The loss of employment in manufacturing in the US, illustrated in figure 3, was caused by out-migration of functions in global supply chains, associated with lower value added per job (Spence and Hlatshwayo, 2011). One of the issues resulting of the changes in geographic location is that, as the emerging markets grow, they also start to compete for more sophisticated functions, allowing the out-migration of more valuable parts of the supply chain and increase competition for higher paid jobs. In fact, firms from these emerging economies are growing and moving upstream in the value chain (Hollanders and Soete, 2010).

Several authors argue that the evolution in Germany has been different, as they have been relying more on manufacturing, with a lower impact on job distribution. In fact, Germany experience a lower reduction of people employed, but still showing a 30% decrease in the number of people employed in manufacturing. But naturally, the evolution of Germany has to be seen in the context of the European Union, where employment in manufacturing decreased slightly in a 10-year period.

Figure 3 - Evolution of job distribution by sector in the US (1970-2010) and the EU 27 (2000-2011)

At the same time, the total employment grew substantially in the US, based mainly on the non-tradable sectors, with government and health leading this generation of new jobs (Spence and Hlatshwayo, 2011). This was a result of “policy shifts” from manufacturing industries to knowledge-intensive services in developed regions worldwide, including the US, but with the notable exception of countries such as Germany (Ghani and Kharas, 2010 and Hepburn, 2011).
Evidence shows that an over-reliance on services, without an adequate balance with industrial activity, damages export performance, reduces labor productivities and favors inequality (Sirkin et al., 2011), and that consequently, deindustrialization may lead to increases in unemployment and export capacity, becoming a social economic risk. In fact, considering that jobs in services are not likely to continue its increasing pace, and given the increased global competition for employment in manufacturing, this (de)equilibrium between sectors may lead to major employment problems (Spence and Hlatshwayo, 2011).

Cowie (2001) has shown that the end of deindustrialization is usually social turmoil, triggering other systemic risks related to Social Unrest (Jovanovic et al., 2012). These two risks have a reciprocal causal relation between them since deindustrialization leads to social unrest and the latter leads to deindustrialization while socially instability hinders attracting and retaining industrial activities. This is an issue not only for developed and industrialized countries, such as the US, which is losing ground to other economies, or European peripheral countries which are increasingly focused in services; but this is an issue also for developing countries such as Brazil, for example, where there are few world-class industrial “islands” disconnected from the rest of the country.

This shift in localization of production raises concern to increase exports and substitute domestic production for imports, which is leading to increase interest and focus in industrial policy, expressed in Obama’s administration efforts to raise lending, defend the auto sector, and boost innovation in selected sectors (particularly energy, medical, pharmaceuticals, IT) (Wade, 2011).

Nevertheless, industry production in the world has increased. Evidence shows that, when compared to services, manufacturing activities are associated with higher productivity levels, and more balanced income distributions, higher income growth rates, and more ability to generate exports (Fingleton, 1999; Sirkin et al., 2011). In addition, manufacturing activities today require the existence of a group of services provided to manufacturing, which leads to the generation of more jobs and a higher diversification on job requirements. Given the changes of the nature and complexity of manufacturing and localization of production, and the ability to react to exiting economic changes, the future of manufacturing is developing in an environment of far greater risk and uncertainty than ever (Manyika et al., 2012).

3. Knowledge capabilities

Endogenous growth theories treat R&D investment and education as endogenous variables in the long run of economic expansion, placing strong emphasis on the role played by knowledge creation and commercialization in promoting economic growth (Romer, 1990; Aghion and Howitt, 1998). The commercialization of knowledge depends on knowledge generation by universities and public R&D labs, as well as on R&D activities by firms, (Jaffe, 1989; Cohen and Levinthal, 1989). Also, formal training in science and engineering has become essential to gain command over industrial technology (Ostry and Nelson, 1995). In this context, the development of the scientific knowledge with the ability of valuing it economically is particularly important, allowing the use and advantage of the spillover effects, essential for economic growth.
Spillovers occur whenever a firm shares knowledge with other institutions performing R&D, such as other firms, universities and government institutions (Griliches 1992). If information about new technologies, goods and processes flows locally more easily than over great distances, then establishing direct contact with entities that can produce knowledge which is valuable for a firm’s activity should be one of the main driving forces leading to the geographic concentration of both production and innovative activities. As new knowledge can spill over, it is possible that one person may discover an opportunity and different person may use the opportunity to exploit it. Such knowledge may be more than just about products and processes, and it can include also organizational forms, management procedures, or other industry trends (Anselin et al., 2000; Gilbert and Kusar, 2006). In addition, Murphy et al. (1989) argue that there are spillover effects of industrialization, that is, that industrialization in one sector can increase spending in other manufacturing sectors by altering the composition of demand.

Processes related with diversification and industrial specialization are complex and essentially related with learning processes and incorporation of knowledge and technology in people and organizations. Research and development has become the primary means by which firms gain understanding over new products and processes, which allow them at least to keep up with the technology evolution (Ostry and Nelson, 1995). Consequently, the competitiveness of industries lies in the ability to produce and use knowledge and technologies mostly developed in a wide network of institutions consisting on a distributed knowledge base (Conceição and Heitor, 2005).

Important factors to manufacturing industries include access to low-cost and/or high-skill labor; proximity to demand; efficient transportation and logistics infrastructure; availability of inputs such as natural resources or inexpensive energy; and proximity to centers of innovation (Manyika et al., 2012). Employment and its skills are paramount to ensure the sustainable development of industries and the socioeconomic impact of industrialization, and this is also true at firm-level. However, as reported by Sirkin et al. (2011), firms tend to focus on short-term results, not providing appropriate training of their own employees, not using their workforce inputs, and not applying technology to develop or improve products. According to Sirkin et al. (2011) “Japanese manufacturers were far more likely than U.S. companies to apply the concept of continuous improvement, seeking to make frequent, incremental changes to their products and production lines.”

Certainly, innovation and ultimately competitiveness require knowledge boundaries to be much wider than the production ones (Brusoni et al., 2001), exploring historical perceptions of the utility of scientific knowledge, labor division and the shape and emergence of contemporary institutions of knowledge and innovation (Baark, 2007). The production, acquisition, absorption, reproduction and dissemination of knowledge are crucial in determining competitiveness and fostering innovative activity (Audretsch et al., 2004). The most decisive input in the knowledge production function is new economic knowledge; however, the generation of new economic knowledge requires a high degree of human capital, a skilled labor force and a high presence of scientists and engineers (Audretsch, 2002).

In addition, the success of this knowledge absorption requires qualified human resources able to, acquire the knowledge, use it, and diffuse it. As established by Cohen and Levinthal (1989), firms differ in their ability to absorb the pool of new knowledge resulting from research that becomes accessible, independently of the degree or nature of its development. This is relevant not only to firms, but to
other forms of organizations and contexts, such as regions and countries. Previous research also shows that local access to knowledge and human capital significantly influence entry of knowledge-based firms into regions (Baptista and Mendonça, 2010).

Also, Borras and Domingo (2007) argue that innovation in a region depends on its own R&D efforts, its innovative tradition and its human capital endowments, and that the composition of economic activity has a positive effect on innovation, and the more specialized regions are, the more innovative activity they have. Accordingly, Acs and Armington (2004) found evidence that cities with higher average levels of human capital and with more knowledgeable people grow faster, confirming the need to look at (de)industrialization processes with a regional focus, without losing sight of the context and taking into account knowledge and innovation capabilities of countries and regions.

Despite the evident benefits from industrialization and the success of many countries in achieving it, numerous other countries remain unindustrialized and poor, and processes allowing some countries to industrialize but not others are not completely understood (Murphy et al, 1989). Part of the explanation is the ability to create and produce knowledge, and to invest in knowledge in a sustained way. Several authors have argued that there is a relationship between the R&D investment and economic growth, and Ostry and Nelson (1995) discuss that the high R&D investment explained the American lead in high-technology products in the 1960s.

Naturally, the evolution of the industrial capacity built successfully in Korea and Japan was a long process, which required many decades accumulation of knowledge and technological ability, through the sustained development of the education system, which allowed for the continuous improvement of productivity levels and international competitiveness (REF). Other Asian economies, such as Malaysia, have developed production capacity without building internal capacity to design and innovate, thus having built a manufacturing sector totally dependent on MNC (Wade, 2011). In contrast, China’s strategy has developed differently, including backward links from MNC operations, and domestic innovation capacity, thus gaining competitiveness against other Southeast Asian economies. In addition to developing internal innovative capacity, Chinese firms are re-concentrating within China the production value chains (Wade, 2011).

Figure 4 shows the evolution of R&D expenditure in several countries since 1991 to 2009. In this time period, the share of GERD in GDP increased substantially in Japan (from 2.93 to 3.36 %), more than doubled in China (from 0.73 to 1.70), and doubled in Korea (1.80 to 3.56%), the last overtaking the investment of UK, US, Japan, France and Germany. For the same time period, the investment in US and Germany was stabilized, and France and the UK showed a slight decrease.
As emerging economies increase their R&D investment and their knowledge capabilities, multinational companies will tend to decentralize their research activities also to these locations, with clear benefits in reducing labor costs and increasing access to markets, local human capital and knowledge in all their locations. In addition, firms from emerging countries are acquiring firms in developed regions, gaining their knowledge capital in the process, also contributing to the global changes on the distribution of R&D (Hollanders and Soete, 2010). As R&D investment start to shift towards Asia, more conditions will exist for the decentralization of research activities by companies, and of higher value parts of the value chain.

Hausman and Hidalgo (2009) suggest that changes in countries’ productive structure can be seen as a combination of the process by which countries find new products as unexplored combinations of the capabilities they already possess, and the process by which countries accumulate new capabilities and combine them with other previously available capabilities to develop yet more products. China and Korea for instance, have successfully changed their industrial structure, with a significant increase in investment in knowledge creation (figure 3).

According to them, cross-country income disparities can be explained by differences in economic complexity, as measured by the diversity of capabilities. Therefore, diversity of capabilities can explain differences in economic performance of countries and regions, and the productivity of a country resides in the diversity of its available nontradable “capabilities” (Hausman and Hidalgo, 2009). Thus, diversification is associated with socioeconomic resilience, or the capacity of firms, regions or countries, to recover following a disturbance or disruption, coming from a crisis, for example. Investment in R&D may provide firms, and regions, the capabilities needed for diversification and the
to becomes more resilient. If this is true, changes in R&D performance should affect the ability to compete. However, additional research is needed to clarify how this relationship truly develops.

4. Geography and industry: looking at people and competences

For many developed regions deindustrialisation trends resulted from the emergence of new and fast-expanding markets; the surge of new pools of cost-efficient workforces, as well as emerging countries aggressive development policies. The consequences for developed economies’ competitiveness and social well-being have been studied by historians, economists and engineers (e.g. Nairn (2002), Fingleton (1999)) and the over-reliance on services has been shown to damage export performance and favor ill-balanced social development, among other claims. Following this literature, we propose integrating a strategy following different complementary directions. On the one hand, acknowledging the lack of existent indicators allowing for an understanding of the industrialization phenomena in a multi-scale perspective, leads us to the need for research analyzing existent data and understanding of the best measures and indicators to be used in this context. Another gap that needs to be addressed is related to the inadequacy of the existing framework to capture the complexities of current supply chains and competitive environments.

Following this argument, we propose an approach focused on people and competences, not common in the existent research, and aim to focus on specific sectors. The role of new technology based firms should be integrated, as part of the analysis on the emergence on new industries. In the end, we wish to build on visualization and communication methods to allow comprehension of this issue and public debate.

It has been long acknowledged that human capital is central to our understanding of technological change and economic growth (Lucas, 1988) and human capital investments, namely expenditure in formal education and training, can explain the production advantage of the most advanced countries (Schultz, 1961). Knowledge, embodied in individuals, can build the sources of competitive advantage by firms and regions (Teece, 1998). In fact, one possible explanation for differences in GDP between countries is that some of the individual activities that arise from the division of labor, such as property rights, regulation, infrastructure, specific labor skills, cannot be imported, which implies that countries need to have them locally available in order to produce. Therefore, countries’ productivity resides in the diversity of its available nontradable “capabilities,” and therefore, cross-country differences in income can be explained by differences in economic complexity, as measured by the diversity of capabilities present in a country and their interactions (Hausman and Hidalgo, 2009).

The issue is how regions and economies develop processes of transforming human capital into competitive advantages in different sectors, are able to contribute to knowledge generation and commercialize technology and from there acquire socioeconomic resilience. Countries differ significantly in their human capital levels, and in the availability of qualified human resources, which determine the capabilities of each economy. In figure 4 we can observe these differences on the level of people with a tertiary education in the labor force, where we observe a substantial growth in the
levels of educated people in Korea, which has surpassed the levels of the European Union, and the high levels in Japan, the UK and the United States (where data is only available for 2001).

Regions and countries with more human capital will tend to be more innovative and entrepreneurial, thereby growing faster. Also, according to Hausman and Hidalgo (2009) more diversified countries produce more complex products, in the sense that they require a wider combination of human capabilities, and higher complexity is associated with higher economic growth. In addition to the presence of strong industrial bases, most of the well performing economies share other common features. One is diversification of activities, meaning that the economic output and that of industrial manufacturing in particular is distributed across a broad set of activities, which allows for mitigating risks but also for enhancing the capacity of the innovation system and the growth potential of the economy as a whole. This diversity can be found in countries such as Korea or Taiwan (Amsden, 2001; Berger, 2005).

Other common feature is supply chain and knowledge networks’ complexity, related to the interactions between people and organizations measured in terms of the activities and job positions needed to produce and certain good within an economy. In fact, one possible explanation for differences in GDP between countries is that some of the individual activities that arise from the division of labor, such as property rights, regulation, infrastructure, specific labor skills, cannot be
imported, which implies that countries need to have them locally available in order to produce. Therefore, countries' productivity resides in the diversity of its available nontradable “capabilities,” and therefore, cross-country differences in income can be explained by differences in economic complexity, as measured by the diversity of capabilities present in a country and their interactions (Hausman and Hidalgo, 2009).

Figure 5 illustrates the number of researchers in the working population of several countries, showing growth in the number of researcher for most countries, except a stable number in Japan, which already had a high share of researcher in the labor force. Again, this figure shows the increase effort of Korea in R&D, and the more recent growth of China, which doubled its proportion of researcher in the population.

![Figure 5 – Researchers per thousand workforce (%)](image)

The issue is certainly how far we all take advantage of opportunities that arise with the increasingly dynamic and globally distributed geography of innovation, as well as how it fosters a new global order and help others to use similar advantages at local levels.

This is because one must take up the challenge of probing deeper into the relationships between knowledge and the development of our societies at a global scale. Our inspiration comes from, among others, the seminal work of Lundvall and Johnson (1994), who challenge the commonplace by introducing the simple, but powerful, idea of learning. Lundvall and Johnson (1984) speak of a “learning economy”, not of a “knowledge economy”. The fundamental difference has to do with a dynamic perspective. In their view, some knowledge does indeed become more important, but some also becomes less important. There is both knowledge creation and knowledge destruction. By forcing
us to look at the process, rather than the mere accumulation of knowledge, they add a dimension that makes the discussion more complex and more uncertain, but also more interesting and intellectually fertile in an international context.

This view follows closely the lessons Eric von Hippel, at MIT, has provided in recent years based on the American experience that user-centered innovation is a powerful and general phenomenon (von Hippel, 1988). It is based on the fact that users of products and services - both firms and individual consumers - are increasingly able to innovate for themselves. It is clear that this is growing rapidly due to continuing advances in computing and communication technologies and is becoming both an important rival to and an important feedstock for manufacturer-centered innovation in many fields.

Eric von Hippel has also shown that the trend toward democratization of innovation applies to information products such as software and also to physical products, and is being driven by two related technical trends: first, the steadily improving design capabilities (i.e., innovation toolkits) that advances in computer hardware and software give to users; and second, the steadily improving ability of individual users to combine and coordinate their innovation-related efforts via new communication media such as the Internet. More educated people have a higher ability to generate and create their own products, as users. So, the focus on people and competences needs to take into account the existence of users with the ability to create and commercialize products, thus engaging directly in the production processes, and contributing to new industrialization processes.

In other words, beyond suitable technical infrastructure, the process of “democratization of innovation” at a global scale requires people with the ability to engage in knowledge-based networks without borders. It is about people and knowledge beyond national borders, and this constant interaction has gained particular importance in recent years.

5. New technology based firms and the new generation of industries

Technology-based entrepreneurship is increasingly seen as a key element of regional competitiveness and that has been taken as “the model” for many other regions and countries worldwide. Knowledge creation and knowledge spillovers are essential elements in stimulating economic development, and technology based firms are seen as an important source of new employment and important promoters of technological change (Feldman and Audretsch, 1999). Silicon Valley and Route 128 in the Boston-Cambridge area, the most dynamic regions in the world today in terms of growth and innovation, were propelled mainly by new technology and the creation of startups - Apple, HP, Google, and Intel, to name a few. Naturally, the formation of Silicon Valley cluster did not occur suddenly; it was a process that evolved over 40 years and it was a process that was only possible due to a set of condition that came together, such as new ideas, creative people and a culture of risk taking (Venkataraman, 2004).

Start-up companies are also becoming global enterprises and engage in services, manufacturing, and research throughout the world, with strong links to universities and research groups. Others are going beyond their borders to procure products and services at lower prices, often from new companies or subsidiaries in countries like China, India and Brazil. Well-trained engineers and computer scientists from Bangalore and Shanghai are competing for jobs that traditionally went to their counterparts in
Europe and the US. At the same time, research universities worldwide are attempting to “emulate” their US counter parts and foster a range of technology transfer offices and commercialization activities, together with industrial liaison programs, mostly intended to foster entrepreneurial environments and the launching of technology-based start-ups. Bringing ideas to the market is their main goal.

Notably, beyond the concentration of people and skills in a number of regions, a key issue that has differentiated North America from many other countries and regions is the availability of a mix of public and private funding sources, in a quite diversified pattern and, most of them, of easy access to SME’s. It is in this context that a few countries have tried to emulate the SBIR program (“Small Business Innovation Research”), which remains unique in many of its characteristics. Although many difficulties have been found in the public support to continue SBIR (as well as that of TIP at NIST), its enormous success and impact should be further acknowledge. This is a program of the utmost importance and relevance that has helped American innovative firms to growth. In addition, many other schemes to fund and support new technology-based firms have been used in America in quite original ways, namely through public procurement through the Defense and Energy Departments.

It is clear that the issue of industrialization cannot be addressed without having into account the role of new firm creation on the development of new industries, namely those coming from the academic environment and built from highly qualified individuals. Nevertheless, their role in the building of new industries and their potential contribution to industrialization processes are yet to be understood.

Summary

The literature suggests that industrialization leads to more well paid jobs, better economic performance, and consequently higher socioeconomic resilience, thus implying that more industrialized counties and regions have a higher ability to adapt to changes and shocks in competitive market, technological, policy and related conditions (Simmie and Martin, 2010). In this paper, we propose the development of a new research agenda addressing deindustrialization as a social and economic risk, as it leads to loss of jobs and lower economic growth, and consequently lower socioeconomic resilience. As Spence and Hlastshwayo (2011) “Expanding employment in the tradable sector almost certainly has to be part of the solution. Otherwise, the US will have a longer-term employment problem”, and this is true for other developed regions in the world. This also applies to more emergent economies, as the advantages of location are not straightforward, and depend greatly of the way countries and regions incorporate knowledge and people in the region, and have the capability to integrate value in the supply chain of the industry (Amaral Filho, 2001).

Even though even though we live in a global economy, innovation and industry policies are developed on a national base, that is, even though companies are global, universities are global, we still have industrial national policies not integrated in international networks. Thus, it is clear that the emerging patterns of innovation require new perspectives for public policies, which have in the past, in the US and in the EU, relied on supporting manufacturers and their intellectual property. Certainly we need to move on from those days and consider better ways to integrate policies, as well as to diversify them at a global scale to better consider “win-all” approaches. A potential way to achieve this is to avoid
overemphasizing current rivals sectors and competitive strategies, but rather to look at science, education and innovation policies towards new challenges that require a strong collaborative and pre-competitive approach.

References


Amsden, AH 2001, The Rise of “the Rest” – Challenges to the West from Late-Industrializing economies, Oxford University Press, Oxford.


Berger, S 2005, How we compete – what companies around the world are doing to make it in today’s global economy, Doubleday press, New York.


Chapter 19:

Manuel Heitor, “How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?”
How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?

Manuel Heitor
mheitor@ist.utl.pt

Center for Innovation, Technology and Policy Research, IN+
Instituto Superior Tecnico, Technical University of Lisbon, Portugal

Paper submitted for publication in “Science and Public Policy”
A version of this paper was presented at the ASHE International Forum 2012, November 2012

November 2012
How far university global partnerships may facilitate a new era of international affairs and foster political and economic relations?\(^1\)

Manuel Heitor

Abstract

A new paradigm of structured international university relationships is emerging as shaped by a new era of international affairs. It is driven by political and economic interests, but also by an increased perception of the growing perceived evidence of the potential benefits resulting from the economic appropriation of the results and methods of science by society.

This paper argues that those relationships act as *agents of change* and a new narrative in university-government-industry relationships. They consider activities that are fundamentally different from the traditional role of universities, involving, most of the times, capacity building. They also require understanding the nature of international cooperation beyond the exporting/importing of “academic services” in all the institutions involved. In addition, they clearly break traditional boundaries of “national systems” and bring new challenges in terms of the necessary institutional integrity the partnering universities need to preserve and foster.

1. Introduction

I argue in this paper that our societies are entering critical times that require the creation of conditions able to strengthen institutions aimed to foster change through patterns of international cooperation. This is well beyond the boundaries of “national systems” and requires people trained to act in quite diversified and global environments. Universities may play that role if their internationalization and specialization path is understood as a key element in a new era of international affairs, where governments and industry intervene through knowledge.

This is because it has become a common place to argue that we need to foster the internationalization of universities, namely by promoting student mobility and university

\(^1\) This paper was prepared during a sabbatical leave at Harvard University, in the “Science, Technology and Public Policy Program”, STPP, at the Kennedy School of Government, co-funded by the Luso-American Foundation, FLAD
networks\textsuperscript{3}, to foster attractive and competitive research and learning environments and to attract and train highly qualified human resources\textsuperscript{4}. But the key issue to understand is why universities need to go global and promote global agendas of teaching and research?...and why governments need to fund universities beyond national borders?

This paper addresses these issues, leading to a new narrative in the relation between universities, governments and industry. It claims for the need for national policies to go beyond “national systems of innovation”. In addition, it calls for the emerging perception worldwide about the “academic divide”, as well as for the impact of the increasingly dynamic and globally distributed geography of innovation on higher education.

The evidence and new narrative provided in this paper is based on lessons learned with international partnerships in science, technology and higher education, as established historically, but with emphasis over the last decade. Four main approaches are used. First, structured partnerships established with the Massachusetts Institute of Technology (MIT) are briefly discussed to help reflecting why so many governments and universities worldwide want to cooperate with MIT. Second, other evolving partnerships in the last three decades are discussed, with emphasis in European consortia built with the goal to foster student mobility and, more recently, to help accelerating innovation. Third, the program of structured joint ventures established between Portuguese universities and US counterparts since 2006\textsuperscript{4} is addressed, as a sample example of a government strategy towards change through international cooperation in higher education. Last, but not least, I briefly discuss details associated with one of those programs, as represented by the MIT Portugal joint venture established in 2006.

In summary, the paper is aimed to address the new conditions for international scientific and academic cooperation and development, to identify their main actors, and to discuss their impact on the emergence of new social realities in many countries and their potential as factors of economic and social development on a global scale. The paper is, therefore, a new contribution about the way international affairs may shape universities and their positioning in increasing globalized societies and economies.

\textsuperscript{2} Bhandari, R and Blumenthal (2011), International Students and Global Mobility in Higher Education: national trends and new directions, Palgrave Macmillan.

\textsuperscript{3} See, for example, the discussion on the emerging “meta-university” by Vest, C.M. (2007), “The American Research University – from World War II to World Wide Web: Governments, the Private Sector and the emerging Meta-University”, University of California Press.

\textsuperscript{4} Heitor, M., and Bravo, M., (2010), “Portugal at the crossroads of change, facing the shock of the new: People, knowledge and ideas fostering the social fabric to facilitate the concentration of knowledge integrated communities”, Technological Forecasting and Social Change, 77, 218-247.
2. Towards a new narrative in university-government-industry relationships

This section considers new conditions fostering a new era of international affairs. They include the need to think and act beyond “national systems of innovation”, the emerging perception of the “academic divide” at world level and, last but not least, the increasingly dynamic and globally distributed geography of innovation.

2.1 Beyond national systems of innovation

Any modern narrative on the evolution of national policies requires the analyses of, at least, the last decades and the seminal work of Sylvia Ostry and Dick Nelson (1995)\(^5\), among many others for the last twenty years, has called for our attention of the relationship between the globalism of firms and the nationalism of governments, as well as the related interplay of cooperation and competition that characterizes high technology and knowledge-based environments.

It should be noted that the Brookings Institution’s project\(^6\) of the early 90’s has promoted this debate, although in a different international context, and clearly shown that tensions about deeper integration arise from three broad sources: cross-border spillovers, diminished national autonomy, and challenges to political sovereignty. As a result, the technoglobalism of the 80’s gave rise to national policies designed to help high-tech industries become more innovative and, consequently, the emergence of technonationalism.

It is under this context that the concept of “national systems of innovation” emerged in academia, mainly through economists and related schools of thought, to explain and explore how and why the systems have evolved differently in the major industrial nations, mainly US, Japan, UK, Germany and France. It was clear by then that the increasing international tensions were largely a result of the attempt of governments to impose national technology and innovation policies on a world in which business and technology are increasingly transnational.

The concept of “national systems of innovation” has evolved during the last two decades, first in association with the need to fight against “market failures”, then against “system failures”. And it helped building new nationalistic policies all over the world, but just as business and science are becoming increasingly transnational. The end result has been a

---


\(^6\) Integrating National Economies, Brookings Institution.
frustration of national policies, on one hand, and a further move toward the multi-nationalization of business, on the other.

This requires many observations and, certainly, deepening the debate in relation to the current economic and social situation in the US and EU, as compared to those in newly industrialized regions and the, so-called, BRICS. First, the myth of “national” high tech industries and related policies to protect them requires to be better understood, if analyzed in terms of the increasing unemployment rates. Second, the debate itself on “national innovation policies” is in any case naïve. No country, even in non-democratic regimes, ever seems to have had a broad, well coordinated one, mainly because of the complex structures associated with any “innovation ecosystem”.

Looking at the last two decades, the picture that is emerging at a global level is not very much different from that discussed by Sylvia Ostry and Dick Nelson in the early 90’s. In other words, it is one of increasing internationalization of private business strategies, while government innovation polices and science funding agencies remain overwhelmingly national. This is leading to new dilemmas for policymaking and to new sources of international friction, although with new boundaries and new players. The key issues to answer include what are the implications of increasing technoglobalism for national and international innovation policies, namely US and EU innovation policies? And, also, what new approaches are required to reduce international frictions and where do public policies need wider integration?

Analysis in the literature has clearly shown that China’s capacity to innovate is still quite limited as compared especially to the capacity of the US. A similar comment could be raised about Brazil, India or Russia and, therefore, there is a large scope to better discuss US and EU innovation policies in a broad international context, well beyond national borders. It is under this context that national innovation policies should help fostering a better understand of future international collaborative paths in education, science and innovation. Ultimately, this will become a key issue for competitiveness everywhere, as discussed below.

2.2 On the emerging perception of the “academic divide” at world level

The recent explosion in demand for higher education by millions of young people around the world, associated with a growing perceived evidence of the potential benefits resulting from the economic appropriation of the results and methods of science by society, have changed the perception of the “academic divide” or “scientific divide” at world level.
Academic institutions from many regions worldwide are now operating internationally, addressing not only potential students individually (this was the traditional paradigm), but increasingly addressing foreign universities, foreign local authorities and governments, in order to develop new types of institutional arrangements. These include helping creating, monitoring or evaluating emerging institutions in other countries, transferring organizational skills, operating training programs for teachers and researchers, contributing to higher education and research capacity building abroad and to the marketing of its benefits for economic and social progress in other societies. Such new arrangements may also include the coaching and steering of research programs in emerging and developing regions, their early inclusion in international networks, and the affiliation of private companies to academic and research programs.

On the other hand, many emerging regions and developing countries are now facing the need and the opportunity of large investments in science, technology and higher education (public and private), aiming at responding to the explosive social demand for higher education and to the vast social and political transformations already induced by new waves of educated youth. These investments not only seek new skills and but also the certification of quality that may be expected from working along together with well established academic and scientific institutions from developed countries. For these institutions, such institutional arrangements provide new forms of expansion, as they tend to help securing new financial or human resources, and to challenge their own traditional competences and agendas.

However, this new paradigm in international academic cooperation does not appear to match the usual model for exporting services. Franchising, for instance, may seem attractive at short notice but its glamour fades away under increasing academic and political criticism. It seems that a new reality is emerging, in which the export of services is intimately associated with the development of national institutional capacities deriving their strengths from the much needed accumulation of qualified human resources, as well as from institutional participation in and recognition from international academic and research networks.

### 2.3 Time and space in higher education

The issue is certainly how far we all take advantage of opportunities that arise with the increasingly dynamic and globally distributed geography of innovation, as well as how it
fosters a new global order and help others to use similar advantages at local levels.

This is because one must take up the challenge of probing deeper into the relationships between knowledge and the development of our societies at a global scale. Our inspiration comes from, among others, the seminal work of Lundvall and Johnson\(^7\), who challenge the commonplace by introducing the simple, but powerful, idea of learning. Lundvall and Johnson speak of a “learning economy”, not of a “knowledge economy”. The fundamental difference is to do with a dynamic perspective. In their view, some knowledge does indeed become more important, but some also becomes less important. There is both knowledge creation and knowledge destruction. By forcing us to look at the process, rather than the mere accumulation of knowledge, they add a dimension that makes the discussion more complex and more uncertain, but also more interesting and intellectually fertile in an international context\(^8\).

The richness of the concept of the learning economy has been demonstrated in recent years throughout the world, by both leading scholars and policy makers. It has been recently addressed beyond Europe and it is at the center of the debate in China, India and Brazil. For example, MGK Menon, former Indian Minister of S&T and Member of Parliament and current President of the India International Center in New Delhi, has recently written about the conditions necessary for innovation to thrive, which require specific local action through a process of “communitization”.

This closely follows the lessons Eric von Hippel, a well-known professor at MIT, has provided in recent years based on the American experience that user-centered innovation is a powerful and general phenomenon\(^9\). It is based on the fact that users of products and services - both firms and individual consumers - are increasingly able to innovate for themselves. It is clear that this is growing rapidly due to continuing advances in computing and communication technologies and is becoming both an important rival to and an important feedstock for manufacturer-centered innovation in many fields\(^10\).

Eric von Hippel has also shown that the trend toward democratization of innovation applies to information products such as software and also to physical products, and is being driven by two related technical trends: first, the steadily improving design capabilities (i.e., innovation toolkits) that advances in computer hardware and software give to users; and

second, the steadily improving ability of individual users to combine and coordinate their innovation-related efforts via new communication media such as the Internet.

In other words, beyond suitable technical infrastructure, the process of “democratization of innovation” at a global scale requires people with the ability to engage in knowledge-based networks without borders. It is about people and knowledge beyond national borders, and this constant interaction has gained particular importance in recent years.\(^\text{11}\)

It is clear that the emerging patterns of innovation require new perspectives for public policies, which in the US and other developed countries have in the past relied on supporting manufacturers and their intellectual property. Certainly we need to move on from those days and consider better ways to integrate policies, as well as to diversify them at a global scale to better consider “win-all” approaches. A potential way to achieve this is to avoid overemphasizing current rivals sectors and competitive strategies, but rather to look at science, education and innovation policies towards new challenges that require a strong collaborative and pre-competitive approach.

Long-term challenges, namely those with current direct implications for firms (large and small), researchers and universities include the emerging opportunities associated with the democratization of human genome sequencing and the emergence of personalized medicine throughout the world, as well as the increasing convergence between health sciences, physical sciences and engineering. But also sustainable energy systems worldwide should be a subject of priority for innovation policies with a great potential for global impact.

The question that does arise is how far can we help transforming R&D and human capital into productivity gains everywhere?

It is not a trivial matter to understand the processes that enable investments in R&D and human capital to be transformed into productivity gains everywhere, at a global scale. Actually, there is a widespread view among economists in many world regions that this kind of investment is too costly for the economic efficiency gains it provides.

This however is a too naïve and superficial approach. Viewed from a wider perspective, in the longer term R&D and human capital investments do matter and are probably the most important factor in explaining economic growth. However, the naïve view has a point: the transition of human capital to growth is not automatic. Specific policies and actions are needed to make this transition happen successfully.

As mentioned above, this challenge is particularly true in what concerns small and transition economies worldwide, but also developing regions in large countries without knowledge-intensive critical masses.

We argue that “international knowledge networks” oriented for S&T policy purposes help in enabling investments in R&D and human capital to be transformed into productivity gains if oriented towards exports and lead markets worldwide. A possible approach is that developed through direct government and diplomatic actions, such as “Swissnex - Switzerland’s Knowledge Network”, “GAIN – German Academic International Network”, “GIAN - The Geneva International Academic Network” and, more recently, “ISTP Canada - International Science and Technology Partnerships Canada Inc.”. These initiatives cover a wide range of approaches, with different scopes and methodologies, and have been, in general, a major tool for national S&T policy action.

Another emerging approach, is the establishment of a new paradigm of international academic and scientific cooperation, as reported in this paper and described below.

3. Our evidence: a new paradigm of international academic and scientific cooperation

The focus of this paper is on the establishment of well organized and structured international academic and scientific cooperation, which may include:

• to set up and maintain a dense network of contacts with universities, research institutions, companies and other organizations worldwide, as well as to support national/regional scientists and entrepreneurs;
• to strengthen the emergence of the different nations and/or regions as a location for science, technology and innovation in close international cooperation;
• to structure, strengthen and promote the interests of national research institutions, universities and leading corporations;
• to support the internationalization efforts of national institutions worldwide, strengthening the development of scientific and technological exchange;
• to help structure, implement and extend bilateral and international research and advance training cooperation programs;
• to facilitate opening-up national universities and research institutions to emerging regions and countries worldwide;
• to facilitate the access of national companies to emerging markets worldwide,
making use of research and knowledge networks with leading researchers and academic institutions worldwide.

Looking at the present and tentatively forecasting the future, we argue in this paper that this new paradigm of international academic, scientific and technological cooperation is emerging as shaped by a new era of international affairs, as driven by political and economic interests. At the same time, it may result as a major shaping factor for development at an unprecedented level. Strengthening the internationalization of universities is recognized as a way to stimulate the integration of national institutions in emerging scientific and economic-oriented networks at an international level. To build our evidence, four main approaches are used in the following paragraphs.

3.1 Partnering with MIT

It is well known that leading American research universities are playing a key role in the process of internationalizing higher education worldwide and, overall, America is gaining from that role. It should also be clear that this is not a new issue. For example, Morgan (1979) describes the role US universities played in helping to build and indigenous S&T base in developing countries until de 70’s and how far American Universities, and the US overall, has gained from that process.

A sample example of such arrangements includes the creation of the Brazilian Institute of Aeronautic Technology, ITA, in S Jose dos Campos in the vicinity of São Paulo, as founded with MIT’s help in 1950. It ranks today among Brazil’s top technical universities, with a particular close development to Embraer, the main Latin American aeronautical manufacturer.

Some thirty years ago, Morgan recommended universities and policy makers about the future involvement on four main areas: institutional building, cooperative R&D, resource base development, and education and training. By that time, he already identified the negative impact of short-term approaches based on franchising academic activities and called for the need to better develop the capacity of the “supplier” university to promote effective institutional building in the “receiving” institution and country.

More recently this theme has been subject of various books and papers in the technical

literature and, for example, the analysis of Bruce Johnstone, Altbach et al., as well as that of Knight, shows an active participation of US universities in indigenous and local development practices, indicating related major advantages, as well as challenges for them and the US innovation policies in the near future. They mostly rely on student mobility frameworks and in the capacity of US universities to attract thousands of immigrants, although they launch a few concerns for the need to foster global research agendas. A recent report by the Royal Society further emphasizes these aspects in terms of scientific collaboration.

In the specific case of the Massachusetts Institute of Technology, MIT, a unique set of international collaborations with governments worldwide has been developed for a number of years based on advanced training initiatives, but integrating most of the times thematic R&D networks, research agendas with local impact, and, in a few cases, industrial affiliation programs.

The Institute has long held a unique position in research and education and has had a remarkable capability to attract students worldwide. Today, though, MIT has become increasingly involved in larger collective experiments with international research and education, most of which are funded by foreign governments, as described in Table 1. These partnerships rely on the success that MIT has had pioneering new frontiers of knowledge through scientific and technical research, as well as working with industry in North America and as a “local” leader in innovation and entrepreneurship. These traits are what attract other universities, companies, as well as countries and their politicians, through partnerships aimed to “emulate” MIT’s success and the environment that has created its success through capacity building.

But the reality is that this is very difficult to enact. People underestimate the social, political and economic challenges, beyond the more basic cultural differences. Creating an exact replica of an environment such as the MIT innovation ecosystem is extraordinarily difficult, if not impossible, no matter where the attempt might be made (MIT’s own failed effort to replicate its world-famous Media Lab in Ireland is one example). Besides, it is the wrong approach and certainly not the appropriate goal of any international partnership. Processes

---

for diffusing knowledge, and by extension higher education, are context-sensitive. Technological systems are, in part, social constructs. In other words, a nation, region, or even a university, anywhere in the world, has to learn its own way and create its own development path.

Although a relatively large institution in international terms, it is interesting to note that most of the partnerships of Table 1 have involved a rather restricted number of MIT faculty and very much concentrated in a nucleus of senior faculty involved in multidisciplinary themes associated with engineering systems. Most notably, the MIT’s “Engineering Systems Division” brings together most of the faculty involved in international ventures. Traditional disciplinary departments and related faculty have resisted engaging in international cooperation abroad.

Table 1. Sample MIT joint ventures in research and higher education – the last decade

<table>
<thead>
<tr>
<th>Strategic partnership</th>
<th>Period</th>
<th>Brief description and evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentine (Mendoza) MIT</td>
<td>1994-1998 (Only one single phase)</td>
<td>The Province of Mendoza in Argentina sponsored this program to develop technologies for its socio-economic development to expedite the transfer and adaptation of technology and management skills to public and private sectors. To meet these objectives, the Fundación Centro de Innovación Tecnológica (CIT) was established within the state university in Mendoza in 1995 and its focus was to: i) Improve the quality of research and graduate education; ii) Begin a new research tradition imbedded in regional needs; iii) Create a research environment attractive to young, ambitious faculty and to government and/or industry, and iv) Establish a favorable context for developing a rich reciprocity between academia, industry, and government. Joint research projects were developed involving faculty from MIT, Universidad Nacional de Cuyo, and other regional and national research and academic institutions.</td>
</tr>
<tr>
<td>Singapore-MIT Alliance</td>
<td>since 1998, (currently under 3rd phase)</td>
<td>1st phase (1998-2005): Singapore-MIT Alliance (SMA) was founded in 1998 as an initiative to develop research talents who can contribute locally to the economy. Born out of a collaboration between the Massachusetts Institute of Technology, the National University of Singapore and the Nanyang Technological University, it started by offering five postgraduate programs: Computational Engineering (CE); Computation and Systems Biology (CSB); Manufacturing Systems and Technology (MST9, Advanced Materials for Micro- and Nano-Systems (AMM&amp;NS); Chemical and Pharmaceutical Engineering (CPE). 2nd phase (2007-2010): The Singapore-MIT Alliance for Research and Technology (SMART) Center was designed and developed as a major research enterprise in Singapore established by the Massachusetts Institute of Technology and the National Research Foundation of Singapore. The SMART Center brought together faculty, researchers, and graduate students from MIT with academic and industry researchers in Singapore and Asia to collaborate in exciting new areas of science and technology. Interdisciplinary Research Groups (IRGs) at the SMART Center included: BioSystems and Micromechanics (BioSyM); Center for Environmental Sensing and Modeling (CENSAM); Future Urban Mobility (FUM); Infectious Disease (ID) 3rd phase (2010-..): The Singapore University of Technology and Design (SUTD) is developed in collaboration with MIT to nurture technically-grounded leaders and innovators to serve societal needs. The collaboration with MIT is multifaceted – the development and offering of curriculum, establishment of a major co-located research center, and recruitment and professional development of SUTD's university leadership team and faculty.</td>
</tr>
<tr>
<td>Cambridge-MIT</td>
<td>2000-2006. (Only one single phase)</td>
<td>The Cambridge-MIT Institute (CMI) was established in 2000 to explore how academics, industrialists and educators might work together to stimulate competitiveness, productivity and entrepreneurship. It was proposed by former British Chancellor of the Exchequer Gordon Brown in the summer of 1998, who wanted to bring the entrepreneurial spirit of</td>
</tr>
</tbody>
</table>

| Malaysia MIT Program | The Cooperative Program between MIT and MUST Ehsan Foundation for assistance in the establishment of the Malaysia University of Science and Technology (MUST) was conceived and planned as part of Malaysia's Second Industrial Master Plan in response to the Prime Minister's call for a creation of a world-class science and technology sector. The major objective of the program was to produce top quality graduate engineers who would understand and participate effectively in Malaysia's program for balanced social and industrial development. The program was sponsored by the Ministry of Science and Technology in Malaysia and the Motorola Foundation and Motorola GTSS and GSG at MIT. MIT participation ended November 30, 2004 due to withdrawal of Motorola commitment to MIT because of economic circumstances. MIT provided advice and assistance to MUST regarding faculty recruitment, IT infrastructure, course delivery, as well as setting up a library, academic media production services, and other critical offices needed. By the end of 2004, MUST had hired 24 faculty members who were graduates from leading universities. It offered seven Master of Science degree programs, which comprised of approximately 50 courses developed with MIT’s assistance. MIT and MUST completed nine joint research projects. |
| Malaysia University of Science and Technology (MUST) | 2002-2004 (Only one single phase) |
| MITPortugal Program | A post-graduate education network of intense and wide ranging collaboration between Portuguese Universities, research institutions, companies, and the Massachusetts Institute of Technology (MIT), has been funded by the Portuguese Science Foundation (FCT) for the period 2006-2011. The network offers Portugal a truly international education program serving as a model for the intersection of engineering education, research, innovation and entrepreneurship. The program has built a research platform for cutting-edge concepts in three promising areas of science and technology: novel biomedical therapies and devices; sustainable energy and transportation systems; and integrated product design. |
| since 2006 | |
| MIT - Abu Dhabi: Masdar Institute of Science and Technology | The Masdar Institute started recruiting students in September 2009, with five 24-month Master of Science programs. It involves about 170 students from 32 countries in spring 2011. The establishment of Masdar Institute is part of a resource policy for the Emirate of Abu Dhabi. Abu Dhabi’s leadership views research and education in alternative energy as a keystone for the future development of the emirate and have expressed their commitment through the establishment of Masdar Initiative, Masdar City and the Zayed Future Energy Prize. The Masdar Institute is a private, not-for-profit, research-driven university, governed by an independent Board of Trustees. MIT’s role with the Masdar Institute is diverse and evolving, but currently is focused on four main areas: 1) development and management of joint collaborative research; 2) assistance in development of degree programs; 3) outreach that encourages industrial participation in Institute research and development activities; and 4) support for capacity building at the Institute in terms of its organizational and administrative structure as well as scholarly assessment of potential faculty candidates. The Masdar Institute faculty, once chosen and appointed, spend up to one year working closely with faculty at MIT in Cambridge, Massachusetts in joint research projects on topics of relevance globally and to Abu Dhabi. The faculty also spend considerable time auditing the graduate-level classes they will eventually teach at the Masdar Institute. |
| 2007-... | |
| Skolkovo -MIT | A three-year collaboration between the Skolkovo Foundation, Skolkovo Tech and MIT to develop a new graduate research university - the Skolkovo Institute of Science and Technology (Skolkovo Tech) in Skolkovo, Russia. The new institution aims to break new ground in bringing together Russian, US and global research and technology – and in integrating teaching, research, innovation and entrepreneurship. Education and research at Skolkovo Tech will be organized around multidisciplinary technological challenges, rather than traditional academic disciplines. The new institution will focus on the following programs: energy science and technology; biomedical science and technology; information science and technology; space science and technology; nuclear science and technology. Master's and doctoral degree programs will be organized under these programs, with focused degree tracks in specialized research areas within each program. Research centers... |
under the Skolkovo Tech organizational umbrella will be multidisciplinary and multi-institutional. In each center, faculty, researchers and students from one or more Russian universities will collaborate with faculty, researchers and students from one or more universities outside Russia. A defining component of Skolkovo Tech will be its Center for Entrepreneurship and Innovation (CEI), which will integrate education, research and practice in entrepreneurship and innovation, as applied to the research results of the Skolkovo Tech research centers. MIT will assist in creating the CEI organization and education program.

What, then do, MIT global partnerships offer? They are an opportunity to work closely with a very successful institution to adapt and improve lessons learned in the building of an environment that promotes innovation. Overall, they are unique opportunities for political and strategic actions fostering change. But it does not occur automatically and requires building a dynamic process of “transformation” in both MIT and the partner institutions, involving the continuous learning and assessing “what works?” and “who effectively cooperates?”.

Among them, it is clear that student and faculty exchange schemes together with “teaching the teacher” programs have certainly been very successful, mainly when concentrated in Cambridge, Massachusetts. The most problematic and complex activities rely in “in-house” developments in the partner institutions and in building-up research agendas with local impact. Setting-up “test beds”, as experienced through the MIT Portugal joint venture, has become a relevant tool and it is further discussed below in this paper.

Time is usually underestimated and it is interesting to note that only one single venture of Table 1 has lasted for more than three consecutive periods (i.e., Singapore). This is because establishing leading institutes of science and technology and an overall research and advanced educational infrastructure is a very complex process and dependent upon many factors, including: the host country’s determination, the economy, the political stability, the resources available and long-term commitment that is required for maturity of such institutions.

In addition, financial agreements with students (scholarships) are very important to the success of these projects. Student recruitment standards must remain high and government support is key over long periods of time, in a way as much independent as possible from political cycles.

Analysis of the various sample examples provided in Table 1 also shows that emerging partnering ventures with governments represent important socio-political shifts. In the past, many institutions (including MIT) have attempted to “export” services or to franchise their
brands, so that, for instance, a country could have a local “branch” of a certain foreign university. In addition, most universities were conceived with a strong “national approach” to their work, oriented toward highly localized needs and locally related research. In many countries, very few (if any) leading universities have adapted to a genuinely international research agenda.

On the other hand, our analysis shows the emergence of a new model, rather complex and time and resources consuming, which is intimately associated with the development of national institutional capacities that derive their strengths from the accumulation of qualified human resources and from institutional participation and their recognition from international academic and research networks. This approach does not appear to match the usual model for exporting services, but it is rather a “learning process” for both MIT staff and those in the partner institutions. The temptation for easy adoption of “academic services” may seem attractive at short notice but its glamour fades away under increasing academic and political criticism.

It should also be noted that this new model of academic cooperation, that includes but does not seem to be a hostage of the traditional forms of services’ international commerce, may derive its uniqueness from the very nature of academic communities and from the strong meritocratic and universalistic ideals that prevail in science on an international scale. In addition, they are also driven by the flow of students and researchers, and by the citizen sense of being part of a “mission” for scientific and social development that motivates some of the best professionals in academic institutions worldwide.

3.2 Other evolving partnerships

At this stage it should be clear that the internationalization of universities has significantly emerged over the last 30 years and Table 2 lists sample initiatives with emphasis in promoting student mobility making use of university networks, most of them with a strong European flavor. They refer to large international partnerships that have been able to attract undergraduate students, mainly those with economic capacity to support additional costs of living abroad.

These partnerships have facilitated a major change in the internationalization of a new generation of people in Europe, together with provoking entrepreneurial attitudes among them. They have also helped strengthening institutional links at the highest international level, mainly within Europe, although their impact in establishing cross-country institutions
is far from being acceptable. Compared to the joint ventures listed in Table 1, they largely represent very-low cost initiatives, student-oriented and, most of the times, without any serious alteration of institutional paths and/or the creation of new dual degrees and/or joint diplomas.

Table 2. Sample examples of typical collective actions among different universities developed in EU since the 80’s

<table>
<thead>
<tr>
<th>Sample example</th>
<th>Main characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utrecht Network</td>
<td>The Utrecht Network represents 31 European universities in 29 countries, co-operating in the area of internationalization in the broadest sense of the word. The group is particularly committed to such areas as student and staff mobility, summer schools, the internationalization of curricula, joint curricula and double/joint degrees (mainly at Master level). The Utrecht Network started off as a small group that was active in student exchanges since the early 1980s. At the start of the ERASMUS program, the existing activities were adopted as the first framework ICP, followed by a steady expansion in the number of partners and exchange students. Initially, this quick start formed the great advantage of the institution-wide approach taken by the Utrecht Network. At present, however, the fact that this construction makes practically every student, from each discipline and from all affiliated universities, a potential ERASMUS student, is regarded as the most important advantage. Each academic year student mobility within the Utrecht network in Europe totals about 1200 students. Approximately 100 students are exchanged each year to the American Universities and 50 to Australian Universities.</td>
</tr>
<tr>
<td>Cluster Network</td>
<td>CLUSTER has grown to a community of 13 research-intensive universities, with emphasis on engineering, science and technology, collaborating in areas of research and advanced education, including joint master and PhD programs.</td>
</tr>
<tr>
<td>Universitas 21</td>
<td>From its origins in Melbourne in 1997, Universitas 21 has grown to a community of 24 research-intensive universities who collaborate in areas of common interest and application to students and faculty. It considers a framework for collaborative research, including undergraduate research opportunities, and joint PhD programs.</td>
</tr>
<tr>
<td>IDEA League</td>
<td>The IDEA League is a strategic alliance between five leading European universities of technology, which was launched in October 1999. It has created a special grant scheme for students that is open to all levels: bachelor, master and doctorates. In addition, it runs one joint master program and it is in the process of launching a doctoral school.</td>
</tr>
<tr>
<td>EIT, European Institute of Innovation and Technology</td>
<td>EIT is a top-down initiative launched through the European Commission with the goal to strengthen innovation in the European area. It has been promoted since 2010 through three “knowledge and Innovation Communities”, which link higher education, research and business institutions to one another in the following topics with high societal impact: Climate change mitigation (Climate-KIC), Information and Communication Technologies (EIT ICT Labs), Sustainable Energies (KIC InnoEnergy).</td>
</tr>
</tbody>
</table>
The analysis of Table 2 shows a clear trend at institutional level, towards the gradual implementation of “Graduate Schools” with a cooperative nature among quite restricted networks, which have been developed progressively worldwide over the past decade in diversified ways (e.g., doctoral schools under IDEA league and EIT). They range from interdisciplinary structures, based in a single university (thus, closely resembling the US model), to subject-specific inter-university structures, as documented in Table 3. In general they aim to provide a better link between research training and research strengths and, in a few cases, have provided flexible structures to attract and hire researchers and graduate students far beyond the traditional university departments.

Sample examples of academic institutions from developed countries operating internationally and developing new types of institutional arrangements include the British University of Dubai and the Sino-Danish Center for Education and Research, in Beijing. These are two of the leading initiatives recently established in developing countries bringing together a network of well-established European institutions trying to access new and emerging economies. The joint ventures between Portuguese and leading American Universities represent a complementary type of action, with a strong capacity building nature.

In this respect, and following some of the issues raised by John Ziman\(^\text{18}\) many years ago and also noted by Nobel Laureate Richard Ernst (2003)\(^\text{19}\), one critically important and emerging institutional issue refers to the training of students and young scientists in order to provide them with core competencies that help them to become successful researchers and prepare them with the adequate “transferable skills” for the job market outside research and academia.

In addition, it is also clear from the analysis of Table 3 that strengthening experimentation in social networks does necessarily involve flows of people. It is the organized cooperation among networks of knowledge workers, together with different arrays of users that will help diffuse innovation and the design of products and services. But establishing these innovation communities requires the systematic development of routines of collaboration on the basis of formal education programs, sophisticated research projects, and a diversified and non-structured array of informal processes of networking.

---
Table 3. Sample examples of recent and emerging collective actions among different universities, bringing together governments and/or industry at a world level, as launched in the 2000’s.

<table>
<thead>
<tr>
<th>Sample example</th>
<th>Main characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>British University of Dubai</td>
<td>Established in 2004 with five British universities to facilitate access to world-class education, training and research in the Middle East. It is the Middle East region’s first, research based, postgraduate university. It is organized around specialized Institutes, each linked to a leading UK partner university. Each Institute offers a distinctive discipline based on their excellence in research and teaching: <strong>University of Edinburgh</strong> Faculty of Engineering and IT (MSc Informatics (Knowledge and Data Management) and MSc in IT Management); <strong>Carnegie Mellon University</strong> Faculty of Engineering and IT (MSc in the Sustainable Design of the Built Environment and MSc in Intelligent Buildings Design and Automation); <strong>King’s College London</strong> Faculty of Business (MSc in Construction Law and Dispute Resolution)</td>
</tr>
<tr>
<td>Portugal-US universities (MIT; Harvard Medical School, Carnegie Mellon; Univ. Texas at Austin)</td>
<td>Established in 2006 through the Portuguese Science and Technology Foundation to facilitate thematic networks in world-class research and advanced education across Portuguese universities. The leading American Universities served as catalysts of the networks formed among Portuguese universities, bringing the necessary leadership to guarantee the success and operation of the networks, as well as their international recognition. Partnerships were continuously open to all Portuguese Universities, but included the following: <strong>MIT-Portugal</strong>: 6 Portuguese Universities, with 8 schools providing joint degrees. <strong>Carnegie Mellon-Portugal</strong>: 8 Portuguese Universities, with 10 schools providing dual degrees. <strong>UTAustin-Portugal</strong>: 2 Portuguese Universities, with 4 schools providing joint degrees. Participating institutions: University of Edinburgh</td>
</tr>
<tr>
<td>Sino-Danish Center for Educ. &amp; Res., Beijing</td>
<td>A joint project on education and research between the eight Danish universities, the Danish Ministry of Science, Technology and Innovation, the Graduate University of the Chinese Academy of Sciences (GUCAS) and the Chinese Academy of Sciences (CAS). The Center will be located at GUCAS’ future Yanhu Campus. The Sino-Danish Center will be fully operational in March 2013. It will accommodate 100 researchers from both countries. Moreover, the Center will offer high quality master programs to 300 master students as well as PhD training programs to 75 PhD students. Participating institutions: University of Copenhagen</td>
</tr>
<tr>
<td>Songdo Global Univ. Campus, South Korea</td>
<td>The Songdo Global University is aimed to have 10 different foreign universities operating on a single campus. It is under construction on land reclaimed from the Yellow Sea in the Incheon Free Economic Zone, which aims to be an educational and high-tech hub. Participating foreign universities have each received a $1 million planning grant to study the feasibility of opening a campus in Songdo, and generous subsidies to support a campus in its first five years of operation. The <strong>State University of New York at Stony Brook</strong> was the first – and so far only – university to move in. Stony Brook’s Songdo location began operating in March 2012 with an enrollment of 35 students in four master’s and Ph.D. programs in two fields – computer science and technology and society. <strong>George Mason University</strong>’s Board of Visitors has authorized the university to move forward in establishing a campus in Songdo in October 2012. This will be the Virginia university’s second attempt to establish an overseas branch: its first, in the United Arab Emirates, ended in failure. The university devoted three years to developing a degree-granting campus in the Ras-Al-Khaimah province only to withdraw in 2009 due to slow enrollment growth, funding difficulties, and disagreements with the U.A.E. government body that was financing the campus. Other universities that are moving ahead with planned campuses in Songdo are <strong>Ghent University</strong>, in Belgium, and the <strong>University of Utah</strong>, which is currently conducting a feasibility study. A number of other American universities that were originally interested in opening a campus in Songdo have dropped out, including North Carolina State University.</td>
</tr>
</tbody>
</table>
Table 4 summarizes the various forms of joint ventures in research and higher education at a world level, providing examples of the different emphasis given. They range from attempts to build new universities, to research collaborations and offering of degrees in association, as well as bilateral agreements among institutions aimed to promote joint academic degrees. But they also consider, with an increasing emphasis worldwide, the development of consortia oriented towards a new paradigm of technology commercialization through international academic and scientific cooperation. It refers to the capacity to turn science-based inventions into commercially viable innovations and related new potential factors of progress on a global scale, in association with a growing perceived evidence of the potential benefits resulting from economic appropriation of the results and methods of science by society. The approach is on sustained growth in emerging and developing regions, which can occur only with the continuous introduction of truly new goods and services, namely in the form of radical technological innovations that disrupt markets and create new industries.

I argue that the accumulation of knowledge by skilled people and institutions in the area of technology-based entrepreneurship requires a specific learning process that takes place together with the building-up of the necessary critical masses in the research community, but needs to be oriented to external and emerging markets worldwide. Making-off local knowledge intensive communities, which are associated with local and specific institutional and university contexts, able to operate in global and very much sophisticated markets requires organized networks fostering new competences in international technology commercialization and diffusion. For example, UTEN Portugal, as established through the Portuguese Science and Technology Foundation in cooperation with Portuguese technology transfer offices and the University of Texas at Austin, focuses on stimulating competences in technology commercialization in a way to help fostering access of technology-based start-ups to emerging markets worldwide. The related experience in implementing country- and regional-wide “university technology enterprise networks” calls for the need to better understand endogenous growth through the accumulation of human capital, beyond the need to access capital and markets.
Table 4. Sample forms of joint ventures in research and higher education at a world level, 2011

<table>
<thead>
<tr>
<th>Main focus</th>
<th>Sample example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation of campuses abroad</td>
<td>Sino-Danish Center for Education &amp; Research, Beijing</td>
</tr>
<tr>
<td></td>
<td>European University Centre at Peking University</td>
</tr>
<tr>
<td></td>
<td>Songdo Global University, South Korea</td>
</tr>
<tr>
<td>Collaboration in the creation of a new university</td>
<td>Singapore University of Technology and Design (SUTD, with MIT)</td>
</tr>
<tr>
<td>and campus</td>
<td>Masdar Institute of Science and Technology (with MIT)</td>
</tr>
<tr>
<td></td>
<td>Skolkovo Institute of Science and Technology (Skolkovo Tech with MIT)</td>
</tr>
<tr>
<td>Research collaboration and offering of degrees in</td>
<td>British University of Dubai</td>
</tr>
<tr>
<td>association</td>
<td>Utrecht Network</td>
</tr>
<tr>
<td></td>
<td>Portugal-US universities (MIT; Harvard Med. School, Carnegie Mellon; Univ. Texas Austin)</td>
</tr>
<tr>
<td>Collaboration and mobility in R&amp;D programs</td>
<td>IARU Alliance</td>
</tr>
<tr>
<td></td>
<td>Worldwide Universities Network</td>
</tr>
<tr>
<td></td>
<td>Masariki Network of Universities</td>
</tr>
<tr>
<td></td>
<td>British Universities Iraq Consortium</td>
</tr>
<tr>
<td>Bilateral agreements among institutions – joint</td>
<td>Cluster</td>
</tr>
<tr>
<td>degrees</td>
<td>Universitas 21</td>
</tr>
<tr>
<td></td>
<td>IDEA League</td>
</tr>
<tr>
<td>Collaboration oriented towards</td>
<td>University Technology Enterprise Network, UTEN - Portugal</td>
</tr>
<tr>
<td>technology commercialization</td>
<td>Skolkovo Institute of Science and Technology (Skolkovo Tech with MIT)</td>
</tr>
</tbody>
</table>
3.3 Case study: the Portuguese initiative of international joint ventures in research and higher education

International partnerships, as we have already noted, were introduced in Portugal in 2006 through thematic networks among Portuguese and a sample of leading American universities, as described in table 5. They introduced a new slant on institutional development, very specifically intended to offset the disadvantages of scale, which limited size imposes on some research units. Multiplying science-based networks stimulates the generation and diffusion of new knowledge. It drives scientific development forward at a time of constant change when the internationalization of the science base is itself a phenomenon of constant flux.

This was a bold step on Portugal’s part, looking outside of Europe to link up with leading universities in the United States. That boldness, though, is not what makes that set of programs unique, though. Rather, it is that the programs’ objectives were specifically, and deliberately, outlined as part of the national political priority given to scientific and technological development, with elements that go well beyond the traditional cross-national collaborations between higher education institutions. Furthermore, the programs were not initiated by the Portuguese universities, but at the level of the national government, in a way to call for “collective action” of all universities and research laboratories. These programs were funded by the Portuguese Science Foundation, and have since enjoyed the imprimatur of the Portuguese people, and are indeed known throughout the country.

The designing of this strategy gained from other experiences in Europe and elsewhere, namely that established by the British Government in 2000 involving a single university partner in Britain, the University of Cambridge, and a single American university, MIT. It was understood at the very initial stage the need to engage not only one, but several universities in Portugal, as well as to work independently with several leading universities in America. This has facilitated engaging a larger number of academic research groups in Portugal, as well as the specialization of the programs towards well defined thematic areas, with those working with MIT focused on engineering systems, those with Carnegie Mellon focusing on information and communication technologies, those with the University of Texas at Austin focusing on digital media and technology commercialization, and those with Harvard Medical School on translational biomedical research. But, above all, it has also facilitated a

---

rather interesting “natural competitive” and dynamic environment among the various thematic networks established throughout the years, which has been particularly stimulating for the success of the overall initiative.

Table 5. Main projects involved in the Portuguese initiative of international joint ventures in research and higher education (2006-2011)

<table>
<thead>
<tr>
<th>Strategic partnership</th>
<th>Launched</th>
<th>Brief description and evolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT Portugal</td>
<td>October 2006</td>
<td>Focused on the field of “engineering systems”, with special emphasis to the complex processes associated with industrial production, sustainable energy, bio-engineering and transport systems. Three main thematic areas for R&amp;D in close cooperation with an industrial affiliation program were established in sustainable energy and transportation systems, stem cell engineering for novel therapies in regenerative medicine, and materials and design-inspired products with specific applications in electric mobility and new medical devices. Overall, the program involved over 340 master and doctorate students at the beginning of its third year in September 2009.</td>
</tr>
<tr>
<td>Carnegie Mellon-Portugal</td>
<td>October 2006</td>
<td>Through the joint program with MIT, co-operation with the Sloan School of Management was strengthened through an international MBA program, “Lisbon MBA”. This involves co-funding from seven major Portuguese companies and banks in a way that will stimulate new research and the quality of education in management sciences in Portugal. Focused on information and communication technologies, in particular the so called Future Internet technologies and services, and involving dual professional masters and PhD programs by Portuguese institutions and Carnegie Mellon University. The areas covered include new generation networks, software engineering, cyber-physical systems for ambient intelligence, human-centric computing (including language technology), public policy and entrepreneurship research, and applied mathematics. Overall, the program involved about 170 master and doctorate students at the start of its third year in September 2009. Three new innovation networks were launched at a later stage, whose goal is to consolidate and expand the successful cooperation among all partner institutions and industrial affiliates: 1) Security and Critical Infrastructure Protection (NET-SCIP); 2) Future Internet Services and Technologies (NET-FIT); and 3) Services and Technologies for Interactive Media (NET-STIM).</td>
</tr>
<tr>
<td>UTAustin-Portugal</td>
<td>March 2007</td>
<td>An “International Collaboratory for Emerging Technologies, CoLab” was established with emphasis on collaborative research in advanced interactive digital media and integrating advanced computing and applied mathematics. Overall, the program involved about 70 doctorate students at the start of its third year in September 2009. Under the joint collaboration with the University of Texas in Austin, a “University Technology Enterprise Network, UTEN” was established in 2007 and oriented towards international technology commercialization and the professionalization of university technology managers.</td>
</tr>
<tr>
<td>Fraunhofer Portugal Research Association</td>
<td>May 2008</td>
<td>Establishment in Portugal of the first Fraunhofer Institute in Europe outside Germany. This project focuses on emerging information and communication technologies, such as “Ambient Assisted Living”, to be complemented by the establishment of R&amp;D consortia and co-operative projects involving several Portuguese institutions and Fraunhofer institutes in Germany.</td>
</tr>
<tr>
<td>Harvard Medical School-Portugal</td>
<td>May 2009</td>
<td>Focus on translational research and information fostering translational and clinical research programs and the development of a new infrastructure for delivering medical information produced by medical schools to medical students across the academic institutions, to health practitioners and to the general public, thus contributing to strengthen the relationships of medical schools and health science institutions with their main constituencies.</td>
</tr>
<tr>
<td>International Iberian Nanotechnology Laboratory</td>
<td>July 2009</td>
<td>It is the first research laboratory set up under international law in the Iberian Peninsula and it is the first such institution worldwide explicitly focused in nanotechnology. It is expected to achieve a reputation as an international institution of excellence in application areas of food and water quality, environmental monitoring and nanomedicine, conceived for about 200 researchers from all over the world, a total of 400 people, and an annual investment and operational budget of around 30 million Euros that is being funded equally by both countries. It is expected that this Laboratory will develop strong links with industry and will attract the membership of more European countries and countries of other continents.</td>
</tr>
</tbody>
</table>
Strengthening the international dimensions in higher education and in S&T is a well-established way to integrate national institutions in science networks as they emerge at the international level. From this it follows that internationalization should be a central component in most, if not all, science and education oriented projects. Internationalization spurs on the mobility of academics, research staff and students. The benefits are considerable. Early mobility in a research career is highly important in determining the work that will be carried out in the future just as it is in forging international ties as part of academia’s Invisible College. With such considerations in mind, each program is tied in with an international partner, strategically and carefully selected in the light of those specific and equally strategic objectives that identify and differentiate each program from others (see Table 5).

In addition, projects directed towards the internationalization of Portuguese industry were also started. Thus, the synergy generated by cross national partnerships within academia extends onward to industrially linked programs – to stem cell engineering for regenerative medicine, automotive engineering, low-energy systems through the MIT Portugal joint venture, telecommunications and information systems with the Carnegie Mellon-Portugal partnership, the Fraunhofer-Portugal association and the UT Austin-Portugal Program. A network of technology transfer offices supporting the development and internationalization of technology-based entrepreneurial ventures was stimulated through the University Technology Enterprise Network (UTEN).

3.4 A case study: The MIT Portugal joint venture

Looking backwards to the 2005 Portuguese political campaign for prime minister, Jose Socrates made the enhancement of Portuguese science and technology a major part of his strategy. The population at large embraced a “Technology Plan” as part of a new political movement to strengthen Portugal’s march to modernity. At the center of the Technology Plan, is the MIT Portugal joint venture, an international partnership that focuses on improving economic and societal development in Portugal. As noted before, it was developed together with a few other joint ventures in a way to guarantee a “natural competitive” and dynamic environment among various joint ventures, but it represents a unique initiative. It goes far beyond the traditional cooperative ventures that research universities in the United States have undertaken with institutions in other countries. As such, it has lessons not only for countries and regions like Portugal, but also for America’s leading schools.
The MIT Portugal program has been promoted as a post-graduate education network of intense and wide ranging collaboration between Portuguese Universities, research institutions, companies, and the Massachusetts Institute of Technology (MIT), funded by the Portuguese Science and technology Foundation (FCT), as described in Table 6. The network offers Portugal a truly international advanced education platform, serving as a model for the intersection of engineering education, research, innovation and entrepreneurship.

A total of 6 Portuguese universities, 28 Portuguese research centers and national laboratories, together with 25 MIT departments, and all 5 Schools within MIT are involved in this ongoing partnership. Seven Doctoral, Master’s of Business Engineering and Master’s of Science programs have been created in the areas of Bioengineering, Sustainable Energy and Transportation Systems and Engineering Design and Advanced Manufacturing.

Table 6. Sample data about MIT Portugal partnership on “Engineering Systems” (2006-2011)

<table>
<thead>
<tr>
<th>Areas</th>
<th>Bioengineering, Sustainable Energy and Transportation Systems, Engineering Design and Advanced Manufacturing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partner institutions and affiliation of private firms</td>
<td>Portugal: 6 Universities, with 8 schools providing joint degrees, Research groups in 13 Universities, 9 Associate Labs and 1 State Lab. involved in R&amp;D projects, 59 firms involved in R&amp;D projects and advanced education programs, as affiliated companies</td>
</tr>
<tr>
<td>Faculty</td>
<td>Portugal: Doctorate researchers contracted and funded (100%): 23 University professors involved: 270 Faculty exchange program at MIT: 28</td>
</tr>
<tr>
<td>MIT: University professors involved: 70</td>
<td></td>
</tr>
<tr>
<td>Advanced Education</td>
<td>4 PhD programs: Bioengineering Systems, Sustainable Energy Systems, Transportation Systems; Engineering Design and Advanced Manufacturing, 3 Master programs: Sustainable Energy Systems, Transportation Systems; Engineering Design and Advanced Manufacturing</td>
</tr>
<tr>
<td>Students</td>
<td>Total number: 631 (414 PhD students, 217 master students), Degrees awarded (November 2012): 214 (171 master degrees; 43 PhD degrees) Student fellowships funded: 234 Fraction of Portuguese students: 70% Students with long periods at MIT (one year): 120 Date of first PhD degree awarded: 25 July 2011</td>
</tr>
<tr>
<td>R&amp;D Projects</td>
<td>R&amp;D projects contracted and funded: 20 Projects 3 years (out of 72 applications), 36 firms involved</td>
</tr>
<tr>
<td>Assessment</td>
<td>Yearly based, by external review committee</td>
</tr>
</tbody>
</table>

Over the past six years, the program has built a research platform for cutting-edge concepts in three promising areas of science and technology: novel biomedical therapies and devices; sustainable energy and transportation systems; and integrated product design. This has been attempted through scalable living laboratories, which were designed to take the form of “test beds”. They include development and demonstration programs oriented to design,
test and implement systems, new products and modeling capabilities for markets worldwide.

In addition, a high-visibility Venture Competition has helped identify and reward global projects that are at an embryonic stage – projects with high-level technological content, or products or services that are able to demonstrate a highly innovative approach. Finalists from other national Portuguese competitions have been encouraged to participate. The Venture Competition has included four tracks: i) Sustainable Energy & Transportation systems; ii) Life Sciences; iii) Information Technology and the Web; iv) Products and Services.

The MIT Portugal joint venture was designed to be implemented along the medium term and three main challenges are identified today, by the end of the initial six years of the program. First, the need to continuously promoting change in the patterns of teaching and learning. This considers promoting student active work and fostering student-centered education schemes, together with stimulating entrepreneurial attitudes. Several design studios have been created or adapted for the MIT Portugal program in several participant universities (e.g. IST-Lisbon, FEUP-Porto). Although, significant improvements have been achieved in some classrooms, there is still a great room for improvement in most of them, namely in what regards classroom physical environment and in the use of interactive technologies in teaching (e.g. course materials and assignments). Main current challenges include forms of "design thinking", “systems thinking”, and establishing routines of an entrepreneurial culture.

Second, establishing stable “test beds” for collaborative research and industry-science relationships. Significant research outcomes and new patterns for collaborative research have been produced by 2011, but there is still room for improvement in stakeholder engagement, together with deepening a basic research infrastructure at the university level. Current examples of on-going test beds for collaborative research include:

- Green Island, at Azores archipelago (Portugal), leading to new sustainability approaches for islands worldwide21.
- New therapies in regenerative medicine, involving hospitals22.
- Urban metabolism, fostering comparative urban studies at international level23.

• Foz Tua\textsuperscript{24}: revisiting regional development in the Tua Valley, in a remote and isolated area of north-eastern Portugal, involving comparative studies at international level.

Third, promoting modern industrial strategies and policies, through the continuous involvement of large and modern industries, supporting the integration of new technology based firms (i.e., “University based”) in large and international industry value chains. Although many activities have been launched and promoted in terms of stimulating entrepreneurial actions by young researchers, including a national competition, there is still a great room for improvement in accessing international markets and in linking them with medium and large companies worldwide and, above all, to promote manufacturing. Typical emphasis on “Knowledge-based services” should continue, but also foster forms of “industrial innovation” and this represents a major challenge worldwide, with particular implications in peripheral and small economies, such as Portugal. Current program activities that may be used to foster new opportunities include engineering design and lean manufacturing in the aerospace and the oil&gas sectors. Also, new therapies, such as those deriving from tissues engineering, leading to bio-industries.

4. Discussion

A new paradigm of organized international university relationships is emerging to help accelerating knowledge diffusion and exchange in many regions worldwide, considering activities that are fundamentally different from the traditional role of universities. They are shaped by a new era of international affairs and involve, most of the times, capacity building and institution building.

Analysis shows that those relationships have to consider accommodating new configurations of knowledge production by establishing alliances with an increasingly large range of “knowledgeable” institutions (Nowotny et al., 2001\textsuperscript{25}). Also, they need to secure and promote a sufficiently stable environment to train and supply talented people, including researchers for that increasingly large range of “knowledgeable” institutions. I argue this leads to the need, more relevant than never before, for systems and related public policies promoting effective institutional autonomy and integrity of modern universities (as defined


Two main emerging issues drive the rationale for this paper. First, the emerging explosion in demand for higher education worldwide, associated with the increasing economic appropriation of the results and methods of science by society, have changed the perception of the “academic divide” or “scientific divide” at world level (Ajakaiye and Kimenyi, 2011). Many developing regions and countries are now facing the need and the opportunity of large investments in science, technology and higher education (public and private), aiming at responding to the explosive social demand for higher education and to the vast social and political transformations already induced by new waves of educated youth (Chen, 2004). This is certainly the case of the recent investments in a few hundred new campi in China (e.g., the new campus in Macao), Russia (e.g., Skolkovo, in the vicinity of Moscow), Latin America (e.g. Northern Brazil, Ecuador, or Colombia), and Africa (e.g, Arusha in Tanzania and other developments in South Africa, Rwanda, Turkey).

Second, it is well known that academic institutions from developed countries are now operating internationally, monitoring or evaluating emerging institutions in other countries, transferring organizational skills, operating training programs for teachers and researchers, contributing to higher education and research capacity abroad and to the marketing of its benefits for economic and social progress in other societies. For example, Heitor and Bravo (2010) refer to such new arrangements, including their early inclusion in international networks, and the affiliation of private companies to academic and research programs, in the case of Portugal and related partnerships with leading American universities.

But I argue in this paper that the key issue is the creation of “organized forms” of international partnerships able to strengthen institutions and the necessary critical masses to compete at an international level and, at the same time, guarantee the adequate level of

28 M. Heitor, “Global knowledge networks for inclusive growth or national systems of innovation?...is there a choice?”, Technological Forecasting and Social Change, submitted for publication.
31 M. Heitor and M. Bravo (2010), “Portugal on the crosstalk of change, facing the shock of the new: People, knowledge and ideas fostering the social fabric to facilitate the concentration of knowledge integrated communities”, Technological Forecasting and Social Change, 77, pp. 218-247.
institutional integrity of universities in emerging and developing regions (Marginson et al., 2002). These networks may have an important impact in advanced education and research, also helping to attract students, as well as to training their future teaching staff in times when higher education systems at those regions are becoming increasingly relevant.

Understanding the new paradigm of international partnerships in higher education will gain from our increasing knowledge of the operational advantages and shortcomings of large international research consortia and organizations. It also requires the understanding of the local characteristics of the processes of technical change and of their specific regulatory and institutional constraints and it calls upon our knowledge of the social construction of technological systems.

This new model of academic cooperation, that includes but does not seem to be a hostage of the traditional forms of services’ international commerce, may derive its uniqueness from the very nature of academic communities and from the strong meritocratic and universalistic ideals that prevail in science on an international scale, as well as by the flow of students and researchers, and by the citizen sense of being part of a “mission” for scientific and social development that motivates some of the best professionals in academic institutions worldwide. However, under which conditions is such a model sustainable?

To answer this question, Table 7 summarizes major lessons learned from the Portuguese experience in setting-up international research networks. It considers three major steps, including: i) training the trainees, through co-hiring of young researchers and exchange programs for faculty; ii) institutional building, by promoting the role of scientific institutions in society, their links with the private sector and adopting policies that foster the creation of critical mass, including those oriented towards fostering R&D consortia; and iii) test beds and thematic R&D networks, facilitating the integration of researchers and scientific institutions in international thematic networks with local relevance. Test beds are “living laboratories” for the production and dissemination of knowledge and facilitating ideas for markets worldwide. They may be assembled and integrated in international collaborative programs in a way to boost local companies’ capacity to export and access new markets.

---

Table 7 - Potential guidelines to foster international research networks

<table>
<thead>
<tr>
<th>Major objectives and policy instruments</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>People</strong></td>
<td></td>
</tr>
</tbody>
</table>
| Train, attract and co-hire researchers, fostering their exchange and the training of a teaching body | • Sustain excellence and internationalization in doctoral programmes  
• Foster and systematize the hiring of researchers with PhDs |
| **Institutions**                       |               |
| Reinforce and promote the role of scientific institutions in society, and their links with the private sector (promoting R&D in business enterprises) | • Reinforce institutional evaluation mechanisms, in order to improve systemic and organizational efficiencies  
• Adopt policies that foster the creation of critical mass, including policies oriented towards fostering R&D consortia. Promote the training of a new generation of technicians and other human resources to support R&D activities |
| **Test beds and thematic R&D networks** |               |
| Facilitate the integration of researchers and scientific institutions in international networks focused on “test beds”, as living laboratories for the production and dissemination of knowledge with local relevance and facilitating ideas for markets worldwide | • Reinforce international partnerships and foster participation in international knowledge-based networks as a way to improve scientific quality and the employability of researchers  
• Foster S&T thematic networks in terms of test beds and living laboratories that can boost companies’ capacity to export and access emerging markets. |

5. **Summary**

A new paradigm of international academic, scientific and technological relationships is emerging as shaped by a new era of international affairs. They consider activities that are fundamentally different from the traditional role of universities, involving, most of the times, capacity building and institution building.

This paper argues that those relationships act as a new narrative in university-government-industry relationships and claims for the need of national policies to go beyond “national systems of innovation”. The new model of academic cooperation, that includes but does not seem to be a hostage of traditional forms of services’ international commerce, may derive its uniqueness from the very nature of the academic communities.

In addition, it is influenced by the strong meritocratic and universalistic ideals that prevail in science at an international scale, as well as by the flow of students and researchers, and by the citizen sense of being part of a “mission” for scientific and social development that motivates some of the best professionals in academic institutions worldwide. As a result, this paper addresses new conditions for international scientific and academic cooperation and discusses their emergence as *agents of change*, as well as their potential impact on new social realities in many countries.

The approach considered in this paper follows four main lines of thought. First, it looks at partnerships established with MIT and reflects why so many governments and universities worldwide want to cooperate with MIT. It is argued that people underestimate the social,
political, and economic challenges associated with establishing leading institutes of science and technology. Any research and advanced educational infrastructure is a very complex process and dependent upon many factors, including: the host country’s determination, the economy, the resources available and long-term commitment that is required for maturity of such institutions.

Second, the paper discusses the evolution of emerging partnerships worldwide, namely those involving the collective action of different universities, and argues that large networks have been very interesting and relevant, but are not effective in promoting change. Network competitiveness depends on many factors, requiring increasingly focused partnerships.

Third, it discusses the case of the Portuguese program of joint international ventures, established in 2006, as representing a new slant on institutional development, very specifically intended to offset the disadvantages of small scale. Multiplying science-based networks stimulates the generation and diffusion of new knowledge. It drives scientific development forward at a time of constant change when the internationalization of the science base is itself a phenomenon of permanent flux.

Fourth, the paper analyses the main “case study” of the joint MIT Portugal partnership established in 2006. The program has built a research platform for cutting-edge concepts in three promising areas of science and technology: novel biomedical therapies and devices; sustainable energy and transportation systems; and integrated product design. The MIT Portugal Program positions Portugal as a scalable living laboratory, making use of test bed developments and demonstrations to help designing and testing systems, new products and modeling capabilities for markets worldwide.

Overall, the paper discusses the way international affairs may shape universities and their positioning in increasing globalized societies and economies.