On technology emergence and pattern detection in aggregate innovative activity

Zur Erklängung des akademischen Grades eines

Doktors der Wirtschaftswissenschaften
(Dr. rer. pol.)

von der Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)
genehmigte

DISSERTATION

von Dipl.-Ing. M.Sc. Vladimir V. Korzinov

Referentin: Prof. Dr. Ingrid Ott
Korreferent: Prof. Dr. Robin Cowan

Mündliche Prüfung: 15.01.2018
Karlsruhe

Abstract

Theories of General Purpose Technologies (GPT) introduced heterogeneity in the world of technological change in order to explain fluctuations in economic growth, productivity paradoxes, or technological clusters. The first part of this thesis addresses in detail a GPT's emergence, presenting a new theoretical model that focuses on pervasiveness character of a GPT emphasizing knowledge network characteristics. Each new product in the economy is a result of a combination of technologies forming a complex network structure of technological inter-dependencies, where a general knowledge emerges as a result of knowledge spillovers, coordination of innovative efforts among economic agents, dynamics of expected profits, and the structure of knowledge base. The model demonstrates similar network characteristics when compared to empirical networks of products and technologies as well as explains clustering of innovations in time, change of technological paradigms and mechanism leading to a technological lock-in.

It is argued that robotics and especially new developments in service robotics can be considered as a potential GPT. Employing a machine learning technique, namely Support Vector Machine, the second part of this thesis introduces a methodology for identification of service robotics patents within databases. The result is a novel possibility to allocate patents which reduces expert bias regarding vested interests on lexical query methods, avoids problems with citational approaches, and facilitates evolutionary changes. Resulting patents are geographically localized and analyzed, being a proxy for knowledge production in service robotics.

The last part of the thesis focuses on a general detection of emergent patterns in micro data. Firstly, a method for statistical identification of clusters of innovative activity is applied to service robotics patents and all German R&D data. A micro-geographic approach identifies spatial localization or dispersion by comparing observable spatial distance patterns between R&D establishments to counter-factual simulations. Findings demonstrate the localization of the knowledge production in service robotics as well as the share of localized German industries in production being higher than in services. Secondly, employing a new methodology based on Markov chain simulations it is assessed whether the number of sustained superior employment growth performing firms in four European economies is different from what would be expected by chance. A mixed evidence of presence and absence of factors determining firm dynamics is found.

Acknowledgements

I would like to thank professor Dr. Ingrid Ott for her expert advice and encouragement throughout this work, as well as professor Dr. Robin Cowan and professor Dr. Patrick Llerena for their brilliant comments. I appreciate the support of my collaborators - Dr. Ivan Savin, Dr. Andrea Hammer, Dr. Florian Kreuchauff and Dr. Stefano Bianchini - for creating fruitful and productive atmosphere, as well as thank Dr. Moritz Müller for discussions at the beginning of my PhD. This work won't be possible without support of my family and Alena Kalyakina.

The thesis was written based on the work carried out within the project "Value creation in and beyond technology" with a financial support from the Helmholtz Association (HIRG-0069) and Projex CSES, Initiative d'Excellence, Université de Strasbourg. Different parts of it benefited from presentations at workshops in Strasbourg, Karlsruhe, Essen as well as from CEF conference in Bordeaux and ISS2016 congress in Montreal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The following thesis is devoted to a various aspects of studying technological innovations as well as the development and application of methods for pattern detection in economics' data. Due to the broad nature of the work this chapter provides a brief introduction and motivation for each project. A more comprehensive and thorough discussion on the embedding of each project into a current stream of literature is provided within each chapter. For the same reason the conclusion of this thesis highlights major contributions to the literature streams, while results sections in each chapter offer more in-depth discussions. Figure 1.1 graphically summarizes all content and will serve as a guide throughout this chapter.

To a large extend this work concerns with structural technologies (left hand side of the Figure 1.1). Economic interest in studying technologies has a very solid foundation. As pointed out by Lipsey and Carlaw technological change "is a necessary condition for sustained economic growth since growth based on capital accumulation with constant technology would sooner or later come to a halt." (Carlaw and Lipsey 2011, p. 564). Developed and developing countries have institutionalized innovation and the creation of new knowledge which brought them on a track of an immense growth in welfare and quality of life. Despite recent discussion on a slowdown in the growth of total factor productivity spurred by the book of R. Gordon (Gordon 2016) many so called "techno optimists" share the opinion that new wave of technologies will bring significant advancements (Brynjolfsson and McAfee 2011).

Technological waves are thoroughly studied in economics and management theories highlighting the heterogeneity of innovation nature and focusing on those that have a drastic, and irreversible impact (Lipsey et al. 2005). Understanding the mechanisms of those innovations has been a subject of the research in economics (Solow

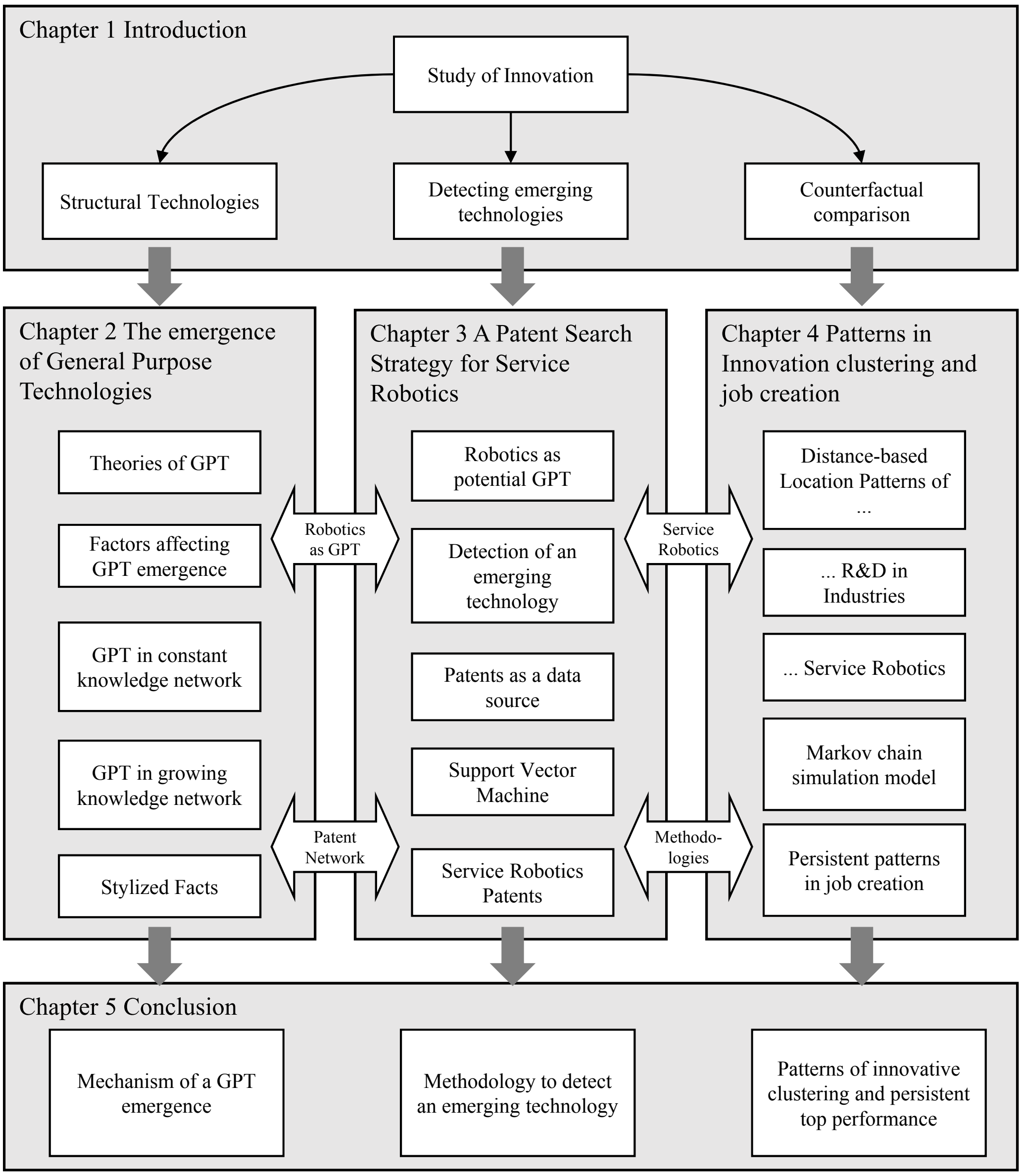

Chapter 1 Introduction
Study of Innovation
Structural Technologies
Detecting emerging technologies
Counterfactual comparison
Chapter 2 The emergence of General Purpose Technologies
Theories of GPT
Factors affecting GPT emergence
GPT in constant knowledge network
GPT in growing knowledge network
Stylized Facts
Robotics as GPT
Patent Network
Chapter 3 A Patent Search Strategy for Service Robotics
Robotics as potential GPT
Detection of an emerging technology
Patents as a data source
Support Vector Machine
Service Robotics Patents
Service Robotics
Methodo-logies
Chapter 4 Patterns in Innovation clustering and job creation
Distance-based Location Patterns of ...
... R&D in Industries
... Service Robotics
Markov chain simulation model
Persistent patterns in job creation
Chapter 5 Conclusion
Mechanism of a GPT emergence
Methodology to detect an emerging technology
Patterns of innovative clustering and persistent top performance

Figure 1.1: The structure of the dissertation.

1956, 1957, Aghion and Howitt 1998, Helpman 1998, Rosenberg and Trajtenberg 2004, Petsas 2003, Cantner and Vannuccini 2012, Ott et al. 2009, Menz and Ott 2011, Strohmaier and Rainer 2016), and management (Youtie et al. 2008). One of the predecessors of theoretical foundations to study the phenomena of technological change and innovation heterogeneity is a concept as techno-economic paradigm. This concept includes a collection of related technologies and the associated economic structure, that is represented by systemic relationships among products, processes and institutions coordinating economic activity. This view on technologies is complemented by the concept of micro and macro inventions, where the former is a series of small incremental improvements while the latter is a set of radical new ideas that arrive as idiosyncratic shocks. Recently these theories evolved into the concept of structural technologies, in the center of which is the concept of general purpose technology presented in a seminal book of Helpman (1998), and continued by Bresnahan and Trajtenberg (1995), Bresnahan (2010), Lipsey et al. (2005) and other scholars (Ott et al. 2009). The name is given for technologies that allowed the mankind to brake important limitations in mastering the forces of nature and drive economic growth and prosperity. Think of a steam engine that allowed to produce greater power than water and wind energy or electricity that broke another limitation allowing the power to be produced in one place and be consumed in the other (Rosenberg and Trajtenberg 2004, Crafts 2004, Moser and Nicholas 2004). Other famous examples are three-masted sailing ship, information and communication technologies)(Brynjolfsson 1993, Vuijlsteke et al. 2007) and recently bio or nanotechnologies (Shea et al. 2011, Lipsey et al. 2005) including modern developments in robotics and artificial intelligence. All these examples have an immense influence on production capabilities and leave a lasting impact, which has been shown by historical studies (Lipsey et al. 2005). These technologies pushed productivity to higher levels and changed organizational and managerial structures opening new opportunities. Literature draws attention to the long lasting effect of these GPTs on the economy, productivity slowdowns, as well as analyses from a historical point of view. However, there has been no attempt to build a model of general knowledge discovery keeping factors that influence the emergence of GPT innovations hidden.

Chapter 2 (see Figure 1.1) of the following thesis narrows this gap, suggesting a model of the endogenous formation of a GPT. In order to reveal mechanisms of GPT emergence, one has to look at the knowledge itself and the process of its creation. In economics literature knowledge is seen, first of all, as a factor of production which properties are rather special. It is cumulative and produced using the existing *stock of knowledge*. It is also non-rival in supply meaning that it can be exploited by many agents simultaneously without decreasing its value for each of them (Grossman and Helpman 1991). Finally, the knowledge is only partially excludable, making it impossible for its producers to enjoy full returns (Grossman and Helpman 1991), which means that knowledge created and applied in a certain context will also create a value for other contexts, introducing *knowledge spillovers*. Due to these properties knowledge is treated specially in economic models using functions with increasing returns to scale (Romer 1986, 1990a) and traditionally is incorporated as a homogeneous variable that can be accumulated. This treatment of knowledge, however, does not allow to capture its heterogeneity which is essential when modeling different types of technologies such as GPT and non-GPT.

Instead of a view on knowledge stock as an aggregate factor of production, this thesis offers a discrete view seeing it as a network of interconnected pieces each representing a fraction of the whole stock. The building block of this idea is an assumption that knowledge is heterogeneous. A similar idea in relation to physical products was offered in works of Hausmann and Hidalgo where authors introduce the concept of 'technological capabilities' that are needed to produce every product (Hausmann and Hidalgo 2011). Thus, the view on the concept of GPT in this thesis is located between techno-economic paradigm and Mokyr's macro inventions. It allows to balance between a very inclusive former concept and the latter one binary dividing innovations on incremental and radical. Assembling an argumentation line around network concepts, a more smooth transition from incremental to radical innovations is introduced and the role of four factors in the process of general purpose technology adoption is demonstrated for a simulated economy. Those factors are knowledge diffusion, coordination of agent's innovative efforts, dynamics in expected returns on innovation and density of agents' knowledge network. The mechanisms of influence are revealed by concentrating on the pervasive nature of GPTs and introducing its emergence as a continuous process of technology adoption studying the mechanisms fostering technological convergence.

The results of this work demonstrate that knowledge diffusion is a key prerequisite for the emergence of a GPT since being discovered once the knowledge spills over to many other applications benefiting most those technologies, which have a potential to be used in many distinct products and industries. The structure of our knowledge should have a sufficient density for a GPT to become pervasive, where by structure the interconnectedness of innovators' ideas is meant and by density – interchangeability of our knowledge among various applications. With the novel metric (*Multiplicity Index*) it is demonstrated how to measure that density given the presence of knowledge spillovers. In addition to these factors coordination of Research and Development (R&D) efforts (concentrating on technological trajectories with more accumulated knowledge) favors GPT in a short term, but changes the influence to an inverted U-shape form in a long run illustrating the famous exploitation vs. exploration trade-off. For the same reason, volatility in the rank of expected returns on products has a negative effect on GPT's emergence in a long run. In addition, the model replicates some known stylized facts (see left-hand side of the Figure 1.1) as S-shaped curve of technology adoption, temporal clustering of innovations in time and some distinct features of networks of the product and technology relatedness discussed by Hidalgo and Hausmann (2009) as well as Boschma et al. (2014).

The focus on the knowledge discovery process brings empirical challenges when attempting to measure knowledge. A proven standard here are bibliometric studies of publication and patent databases. While a publication network might serve as a proxy of a knowledge network however patents seem to be a better measure given that patentability requires an indication of the intended commercial implementation. Despite all the difficulties that arise in the use and interpretation of patents, they are widely accepted as an indicator for innovative activity (Griliches 1990, Hall et al. 2005). After theorizing about the emergence of technologies a methodology is developed to detect and monitor the developments of a potential GPT in the existing databases. The middle of the Figure 1.1 depicts a major content of the Chapter 3 that complements the theoretical work in this thesis with an empirical methodology to detect a general purpose technology within a patent database.

The field of service robotics has been chosen as a part of a broader technology - robotics that has a potential to become a future GPT of our time. Rapid developments recently observed in artificial intelligence, microelectronics, sensors and other related technologies (Brynjolfsson and McAfee 2011, Graetz and Michaels 2015, Ford 2016) indicate that robots might potentially disrupt current trends and significantly contribute to economic growth. Beyond its potential productivity effects, service robotics is believed to induce visible changes in employment structures (Autor et al. 2003, Ott 2012, Frey and Osborne 2013, Graetz and Michaels 2015). It has a potential to change an organization of processes in firms and everyday life of people by the diffusion of at least semi-autonomous physical systems out of industrial fabrication and into service economies. Using the advances of modern digital economy robotics can move from a professional use to a more private use. In order to understand service robotics one needs to identify its scope and detect it within various databases.

The process of detection is non-trivial due to the fact that there is no widely agreed-upon definition of emerging technologies (Halaweh 2013). The initial lack of common knowledge, standards, and specifications entails uncertainties along various dimensions (Stahl 2011). Future costs and benefits, relevant actors, technological adoption, and potential socio-economic implications such as creative destruction are highly unclear (Srinivasan 2008). Given these inputs a methodology is developed that limits expert bias with respect to a technology definition. The method is based on classification of patents and has several advantages over usual technology classification tasks. First, experts do not choose the terms and keywords should be added to or excluded from the primal search. Hence, the typical lexical bias towards preferred subfields is limited. Speaking of lexical versus citationist approaches, the method also avoids a major drawback of citational methods which circle around a core dataset and rely on future works explicitly referring to this prior art. Since citations in patents are generally rare, for young emerging technologies the citation lag decreases the expected number of citations for any given document to a negligible amount. Second, the procedure offers strong portability, so that it can easily be applied to scientific publications. Moreover, the classification method developed in Chapter 3 can be applied to any emerging technology - not only those that arise as an initially small subset consisting of niche applications like emerging service robotics out of robotics. For example nanotechnology would have been hard to detach from some well-defined mother technology, or Industry 4.0, which is a superordinate concept describing digitally cross-linked production systems and, thus, enveloping various heterogeneous sub-technologies that are hardly classifiable.

The methodological development started in the Chapter 3 of this thesis is continued in Chapter 4 (see right-hand side of the Figure 1.1). Here the broader perspective is taken considering the problem of technology detection as a part of a broader set of problems in identifying micro trends and patterns in macro level data. Macro- and microeconomics, unfortunately, largely remain separated from each other with one being concerned with aggregate economy trends and the other being focused on single markets and people behavioral patterns. Due to the fact that social systems are shaped by humans whose behavior still remains to a large extend a mystery, it is hard to predict macro trends from microeconomics data. However, recent trends in machine learning, artificial intelligence, big data and the growing volume of information, due to the developed ICT structures, are already helping to bridge these two fields together. Modern techniques and methodologies, developed also in other science domains, make it possible to consider in calculations enough micro patterns to generalize them on a macro level. Thus, as much as technologies drive economic growth they themselves will drive the economics as social science helping it to overcome its challenges. Chapter 4 targets the question of how not to be misled by chance while observing micro data on a macro level.

Firstly, a geographical perspective on technological innovations is taken into account including service robotics on the example of Germany. For the first time a distance-based "dartboard approach" (Duranton and Overman 2008) to the new R&D data from Germany is applied, showing whether clusters of innovative activities can be significantly distinguished from the ones expected by a random process. A simulation technique applied allows the detection of deviations from a normal pattern that point to an existence of agglomeration forces. In particular, it is shown that service robotics knowledge production is significantly clustered in southern regions of Germany. Analyzing the overall industry location patterns of R&D on a 3-digit level reveals that 40.8% of industries deviate significantly from random spatial location patterns. In general, the share of localized industries in production is higher than in service industries. Thus, knowledge creation in production industries tends to be more localized than in services. In service industries dispersion occurs more often than localization. Interestingly, especially research-intensive service industries exhibit strong cross-distance indices of dispersion. Overall, the evidence on industry-specific spatial concentration of R&D is relatively weak. The results indicate that localization of both R&D establishments and researchers, if it occurs, mainly is observable for production industries over relatively long distances. However, these results do not contradict with the notion of R&D itself being concentrated. They rather indicate that clustering of R&D establishments or researchers at short distances is not or only weakly connected to the 3-digit industries, where innovative activities are performed.

Secondly, with a methodology using Markov chain simulations it is demonstrated whether randomness can be ruled out when observing sustained superior job creation in Spain, United Kingdom, France and Italy. It is shown that the observed number of firms can not be explained with a simple process modeled through a first order Markov chain, demonstrating that it is not enough to assume that the employment growth of the firm in the next period depends solely on the growth in the current period. This pattern can be seen regardless of the confidence level and definitions of superiority of growth. The research strongly advises for a presence of drivers enabling sustained high-growth performance in the economy. Economic theories explain such behavior with an idiosyncratic shock that helps those firms with higher relative efficiency experience a reduction in prices, which allows them to expand at the expenses of less efficient units. Together with higher profitability and sounder financial conditions more productive firms access the resources needed to invest and fuel additional growth. In accordance with managerial literature this drivers might as well be firm's dynamic capabilities and resources that are unique, durable, create value on the market, and generate returns which are appropriated inducing competitive advantages. All these factors combined lead to a sustained superior performance of a firm. Accumulation of these capabilities overtime allows firms to build various routines that help them to grow. Altogether the research grants encouragements to the economic and management theories seeking for factors of a persistent high-growth performance. It also provides a positive sign to policy-makers indicating that if such factors exist they could be targeted by a specific policies spurring employment.

As can be seen the following thesis concerns with a broad set of research questions that can be united by the interest in technological change and the development of new methodologies applicable in economics research. Chapter 2 addresses the question of how do general purpose technologies emerge and what factors contribute to this process. Chapter 3 continues with the question of how an emerging technology of service robotics can be detected within databases using modern techniques such as machine learning. Section 4.2 in Chapter 4 assess the significance of spatial clustering activity of innovative centers in Germany including service robotics, while Section 4.3 demonstrates a methodology to rule out chance in observing sustained superior job creation in four European countries. Finally Chapter 5 highlights major contributions to the various strands of literature.

Chapter 2

General Purpose Technologies as an emergent property[1]

2.1 Theories of general purpose technologies

Innovations are vital for the process of economic growth (Solow 1956, Romer 1990a, Aghion and Howitt 1992, 1998, Helpman 1998). Throughout the history concepts such as techno-economic paradigms, technological trajectories and revolutions, Mokyr's macro inventions, or 'enabling technologies' were introduced in order to distinguish within innovations and highlight those that have a drastic and irreversible impact on a society (Lipsey et al. 2005). A specific type of those drastic innovations called General Purpose Technologies were introduced as one of the forces to explain growth process and its cyclicality (Bresnahan and Trajtenberg 1995, Bresnahan 2010). Ever since their wide acknowledgment in the book of Helpman (1998), these technologies are seen as engines of economic development of countries (Ott et al. 2009) or industries (Strohmaier and Rainer 2016). Despite some disagreements on what technologies shall be considered as GPTs, this concept stays relevant up till now and is proved to be important during the first and the second industrial revolutions as well as for an information age (Bresnahan 2012, p. 612).

A formal definition of GPT put by Lipsey et al. (2005, p. 98) says a "... GPT is a single generic technology, recognizable as such over its whole lifetime, that initially has much scope for improvement and eventually comes to be widely used, to have

[1] Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104.

many uses, and to have many spillover effects". The literature claims that in order to be classified as a GPT an innovation has to possess three major characteristics. The first one *pervasiveness* implies that a technology or its principle is used in vast amount of products throughout an economy and in various applications (as, e.g., electricity is used from heating and lighting our houses to powering trains). The second *technological dynamism* postulates that these technologies experience significant improvement in their efficiency and effectiveness throughout their lifetime (one example is the 'Moore's Law' well-known in the semiconductor industry). Finally, *innovation complementarity* (also called a 'dual inducement mechanism') means that these improvements induce innovations in application sectors of this technology (e.g., the evolution of semiconductors has led to the introduction of numerous portable devices) and *vice versa* (Helpman 1998).

The majority of historical GPTs' studies focused on their impact such as, for example, the revolution in manufacturing brought with the introduction of electricity. In the early formal models of GPT, the emphasis was on the attempt to account for a "residual" in aggregate production functions of mainstream neo-classical models (Bresnahan and Trajtenberg 1995, Helpman 1998) and explain the famous 'productivity paradox' (Brynjolfsson 1993). In these models the new transforming technology appears periodically and exogenously and induces changes in economic structures (like in Bresnahan and Trajtenberg (1995), where a switch to a new production regime using a GPT happens after a certain number of the new intermediates becoming producible, while agents realize their ability to produce these intermediates at a pre-specified moment). In later models authors followed a so called "structuralist-evolutionary approach" (Lipsey et al. 2005), where technologies “evolve under a stream of innovations” and the effect of a newly arrived GPT on the economy is determined endogenously, but the moment of arrival is still exogenous (see also Carlaw and Lipsey (2006)). The work of Lipsey and Carlaw names GPTs as a part of 'structural technologies' with eleven key characteristics incorporating them in a sequential model with simultaneous GPTs (Carlaw and Lipsey 2011). Similar to others this model uses the concept of an aggregate production function, which does not allow to reveal the heterogeneity of knowledge stock out of which GPTs emerge. More recent models on GPT focus on a "dual inducement mechanism" between GPT and its application sectors (Bresnahan 2012) assuming one in a pair of complementary technologies to have generality of purpose. These works also elaborate on different types of knowledge or 'growth bottlenecks' (Bresnahan 2012), but their arguments take a GPT introduction for granted. Thus, the literature has long been focusing on explaining the effect, which GPTs have on economy, but so far none of the models tried to address the process of GPT formation, or, as pointed by Cantner and Vannuccini (2012, p. 74), in all current models a GPT "arrives from the outside of the system". Therefore, factors and properties of the economic system, that influence the GPT emergence, remain hidden.

The present study focuses on the pervasive nature of a GPT identifying factors that foster it's inclusion as an input into newly discovered products and proposing mechanisms how these factors work. Thus, we consider the emergence of a GPT not as a binary outcome but as a continuous, where certain technologies may exhibit the pervasive property to different extents, and the larger this extent, the more likely the technology will be classified as a GPT. For the same reason, this work is not meant to answer the question, in which exact moment a GPT can be considered as an emergent. Instead, we look for forces boosting the process of "technological convergence" coined by Rosenberg (1976), where economy utilizes the same technologies for different purposes and consumer products become related through similar technologies. We offer a novel perspective on the knowledge discovery process as a network growth, where nodes are single technologies (knowledge pieces), and each new connection (link) represents a new knowledge being discovered (technology combination resulting in value added); each technology allows to produce a certain intermediate input, while fully connected groups of those nodes (cliques) stand for producible final goods.

This work builds on the literature started by Schumpeter (1934, p. 65) defining innovations as "new combinations" of new or existing knowledge, and continued by theories of architectural innovation (Henderson and Clark 1990), recombinant growth (Weitzman 1998), combinatorial technology models (Arthur and Polak 2006) and works on technological capabilities (Hidalgo and Hausmann 2009), considering knowledge as a collection of heterogeneous pieces, being interconnectable with each other in one or another way. In other words, technologies are assumed to have a hierarchical structure and be interrelated (Lipsey et al. 2005).[2]

Thus, the process of GPT formation transforms into inclusion of a single technology

[2] While in reality a complex technology can consist of sub-technologies, which in their turn consist of sub-sub-technologies and so on, we simplify this modular structure implying final goods to be producible out of a large group of interconnectable but single technologies. Note that this is done without loss of generality since those complex technologies can be seen as interconnected groups of intermediates, which in their turn have to be all connected to further technological inputs to invent new final goods.

(*potential GPT*) in as many as possible final goods. To consider this, we model the technology to have the *potential* to be included in all final goods, but without *certainty* to do so, which is achieved by allowing multiple competitive ways of producing the same good. The models counts only first discovered combination of technologies as production inputs for a product, which is done as a simplification to concentrate on the process of product discovery and not further competition between substitute goods over production costs,[3] and shall reflect the fact that technologies included in early product discoveries have a lead time advantage over future competitors (Arthur 1989). The more often the potential GPT enters those early product discoveries (in other words, fulfills its potential), the easier it should become to identify an emergence of a GPT.

Thus, the aim of the present work is to reveal the factors that may foster or hamper inclusion of the potential GPT as an input for production. Among the usual suspects we outline the process of knowledge diffusion, the structure of the technological network, the choice over technological trajectories to follow and the pressure from the demand side (and, in particular, its variation over time) in discovering new final goods. The knowledge diffusion is considered because of the famous public good property of knowledge (Arrow 1962) and the resulting possibility to create "complementarities among trajectories" (Dosi 1982, p. 154). The extent of this effect, however, is contingent on the exact network structure of knowledge considered, since the complex interrelationships between technologies can result in some technological links being present in numerous products (as was the case, e.g., for a steam engine combined with a wheel) or very few only. Another rationale to consider the knowledge network is that the potential GPT is not necessarily the only technology having large scope of applications, but that all technologies have a different potential degree of pervasiveness, thus, affecting each other chances to become included as an input in final goods. The mechanism behind choosing among technological trajectories, in its turn, is important due to the competition among the aforementioned alternative technological combinations in becoming first to satisfy each consumer need. Since the innovation process is seen as search in complex technology spaces (Silverberg and Verspagen 2005, p. 226) and characterized by a strong path dependence (Nelson and Winter 1982), it is a key to our model to see how this mechanism affects the GPT adoption. Last not least, the role of the demand side effects is not

[3] Introduction of production costs into the model is left for further extensions.

clear. From a policy perspective this work focuses on the following question. Is it beneficial for the knowledge discovery process in general and the GPT adoption in particular that society starts favoring a certain product development as it was the case, e.g., for nuclear power plants in the 1950s (Cowan 1990) or renewable energy generation in the last two decades (Herrmann and Savin 2016)? In both cases, the policy maker was providing large subsidies to discover a product with certain characteristics, while actual choice among different technological trajectories were left to innovating firms. Clearly enough, none of the four factors shall be considered in isolation from the others, and the rest of the study devotes particular attention to the interplay between those forces on the GPT emergence.

The rest of the chapter is organized as follows. Section 2.2 describes the basic set up of our model and formulates four propositions on factors triggering the process of GPT adoption. We provide results of the numerical analysis of our baseline model in Section 2.3 additionally extending it by the introduction of an increasing knowledge base over time. In Section 2.4 we outline some stylized facts that our study reproduces, while Section 2.5 discusses implications of the results and concludes.

2.2 The model of general purpose technology emergence

2.2.1 Technology network

In this model we focus on the process of knowledge discovery. In particular, it is assumed, that to satisfy the consumer needs, the certain population of product types (P) is necessary to be introduced into the market (innovation as a problem-solving process (Dosi 1988b, p. 1125)). For each product type to be discovered and introduced onto the markets, some intermediates (I) need to be combined, which in reality are typically combinations of other intermediates. We simplify our modeling by considering only two layers (see left panel of Figure 2.1): the product types (final goods e.g. Internet, transportation) and the intermediates (technologies used to produce the intermediate input: transistor, combustion engine).[4]

From the beginning, the technologies are present in the model as yet not connected

[4] Henceforth, we use the terms ‘technology’ and ‘intermediate’ as synonyms.

nodes of the technology network (mid panel of Figure 2.1). For these technologies to find practical application, they need to become interconnected with other technologies forming fully connected component (*clique*)[5], which we schematically demonstrate on the very right panel of Figure 2.1.

The intuition of the assumption is that constituent elements of a product shall be all "adjusted" to one another so that the connected component of those elements exhibits a larger value than those taken separately. One of the famous examples of this nature is a printing press invented by Johannes Gutenberg. The major contribution was in the ability to combine the existing elements of various industries and specialties existed before, bringing them together to produce a commercially viable technology. Another famous examples include the internal combustion engine and the digital computer.[6] or a smartphone. Thus, the discovery of new products becomes an incremental process of figuring out the combinations of intermediaries.

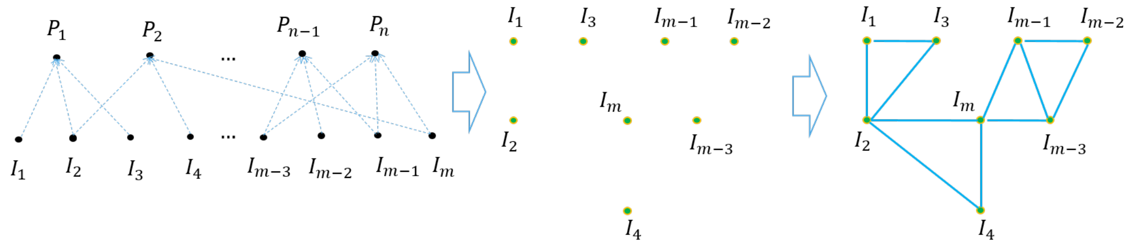


Figure 2.1: Layers of products and intermediates.[7]

We make another assumption that each product type has more than one way of

[5] In a similar way of reasoning, one could consider a fraction of technological links from the clique also forming a fully connected component to be themselves technologies of a higher complexity (combining more than one technological input) and necessary to be discovered for the respective good to become producible. For simplicity, however, we avoid such a discussion to keep our argument clear and simple.

[6] As stated by Kauffman (1995, p. 24): "The whole is greater than the sum of its parts". An illustration of that definition in reality is another quote from Holland (1995): "Take two technological innovations that have revolutionized twentieth-century society, the internal combustion engine and the digital computer. The internal combustion engine combines Volta's sparking device, Venturi's (perfume) sprayer, a water pump's pistons, a mill's gear wheels, and so on. The first digital computers combined Geiger's particle counter, the persistence (slow fade) of cathode ray tube images, the use of wires to direct electrical currents, and so on. In both cases most of the building blocks were already in use, in different contexts, in the nineteenth century. It was the specific combination, among the great number possible, that provided the innovation."

[7] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

production, i.e. there is more than one technology combination satisfying a certain need (compare technology network from Figures 2.1 and 2.2 consisting of the same product types and intermediates). The intuition is that there is no consumer need to be satisfied in a unique way. Those alternative technology combinations satisfying the same need can be anything from having very different inputs (e.g., paper towel vs textile one vs electric hand dryer) to fairly similar ones (different types of cheese, all fermented out of milk by yeast).

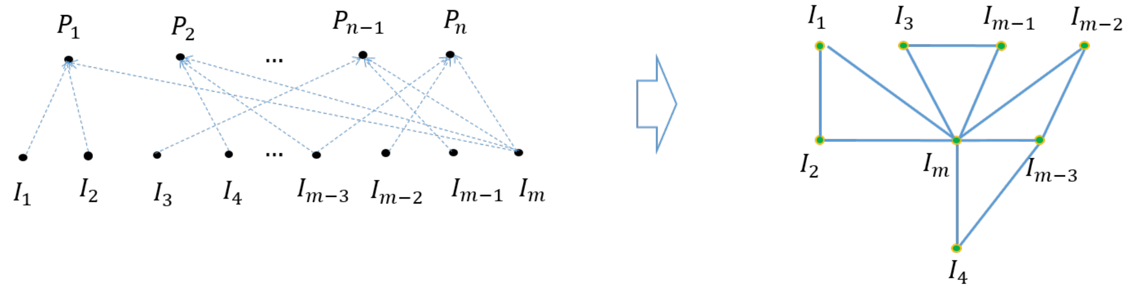


Figure 2.2: Alternative combinations of intermediates.[8]

Important to stress is that combination of two distinct technologies (like $I_{m-1} - I_{m-3}$ on Figure 2.1 or $I_m - I_{m-3}$ on Figure 2.2) may enter more than one product both, within one way of technological combination but also between them. This model feature reflects the fact that in real world we may utilize the complementarity arising from the combination of two technologies in more than one application.[9] Combining all alternative technology combinations together (constructing a multiplex network) one obtains a '*potential technology network*' - mapping of all possible combinations producing added value, (see left panel of Figure 2.3). The resulting network can be considered as a *technological paradigm* in accordance with Dosi (1982, p. 148)'s definition: "an 'outlook', a set of procedures, a definition of the 'relevant' problems and of the specific knowledge related to their solution",[10] while each single way of technological combination as a *technological trajectory* – "the direction of advance within a technological paradigm". Clearly, the position of each technology in such

[8] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[9] For example, combining tubes and lenses for telescopes, microscopes, photo equipment etc.

[10] In Dosi (1988a, p. 1127), words on ".'pattern' of solution of selected techno-economic problems based on highly selected *principles derived from the natural sciences*" are used.

a network is different. In accordance with the arguments presented in Section 2.1, we consider GPT to be the one with largest generality of purpose, thus, *potentially* entering all product types in at least one technological combination (right panel of Figure 2.3).[11] However, there is no guarantee that the GPT will eventually be included in any product type discovery. Our aim is to identify factors fostering fulfillment of the GPT's potential application in largest possible number of final goods.[12]

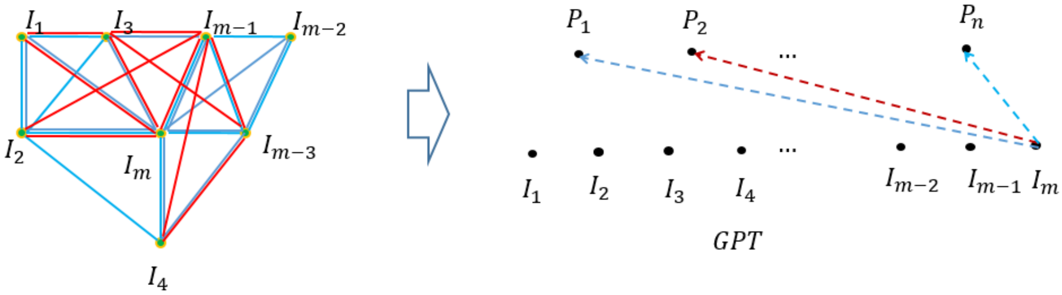


Figure 2.3: Potential technology network and GPT.[13]

2.2.2 Discovery process

The process of knowledge (and eventual product) discovery is the process of satisfying consumer needs. To keep the demand side simple, we consider each product type having a certain value (V), proxying an expected profit from its discovery. These values are the driver for profit-oriented agents (anyone able to conduct R&D: firms, entrepreneurs, scientists etc.) to conduct the discovery process upon the technology network. In the baseline model agents are considered to be able to see all alternative ways of production (thus, setting the whole potential technological network to be *visible* to all agents), an assumption that is necessary to test the mechanics of

[11] At the same time we rule out the option that GPT enters all product type within any single way of technological combination to make its inclusion (in all products) a less trivial task.

[12] Henceforth, we refer to GPT as the technology with largest pervasiveness *potential*. While examining to what extent this potential has been fulfilled, we interchangeably call it 'GPT' and 'potential GPT'.

[13] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

our model and which is to be relaxed later (see Section 2.3.3). Discovery of each technology combination has certain difficulty (d) – the resistance of the link to be discovered.[14] This difficulty is not known to agents so that agents can only compare alternative trajectories within one product type in terms of number of links yet to be discovered. The latter introduces *uncertainty* into the model since the best strategies are unknown, and agents can at most rank opportunities. The values, the difficulties and the technologies are assigned independently from each other. Thus, it may turn out that new knowledge necessary for a very valuable product type can be invented with a small effort (e.g., as penicillin discovered accidentally by Alexander Fleming) and the other way around. Also, the GPT is not necessarily attributed with more or less difficult technological links, differentiating our model from the existing studies attributing an *ex ante* advantage to the GPT, while the only virtue of a potential GPT we allow is its *a priori* larger scope of application.[15]

Over time, agents try to discover a certain technological combination from those being visible for each product types, where the order of the products to be considered is random and set anew each cycle. The effort applied is equally distributed among all yet undiscovered links so that once one of the constituent links becomes discovered, the effort is redistributed among the remaining ones creating a cascade effect of product discoveries in time (increasing number of innovations per period over time, see Figure 2.4).

The probability to discover a certain technological link x being a part of the product type y discovery is modeled stochastically as a uniform random number $Pr_x \in U[0, 1]$ and turns this link into a discovered one *if*:

$$Pr_x < \frac{V_y}{d_x \times L_x} \tag{2.1}$$

[14] Note that this does not necessarily introduce a discrete complexity ladder: goods consisting of 3 or 4 technologies would require 3 and 6 technological combinations, respectively, to be discovered. One, however, can smooth the product complexity by randomly assigning zero difficulty values to a certain fraction of edges. We conduct such an exercise as a robustness test.

[15] In perhaps the most related to us studies by Bresnahan (2012, p. 629) combinations of technologies (products) also have values and 'there are two potential ways to create new value': a 'compromise' way does not involve GPT and has lower value than an 'efficient' one including GPT. Thus, the model assumes a higher expected profits to production of a good with GPT pointing out that generality is expensive.

[16] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

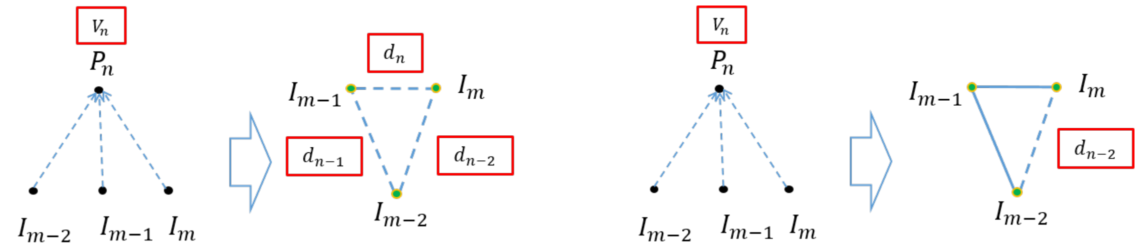


Figure 2.4: Fixed amount of R&D effort is redistributed among three undiscovered links (dashed lines on the lefthand side) or appllied to only one (dashed lines on the righthand side) because discovered links (solid lines on the righthand side) are known to agents and require no R&D.[16]

where L_x is the total number of yet undiscovered links in the clique in which the link x is located. Hence, the higher the product type value or the smaller the resistance of the respective link or the smaller the number of yet undiscovered links in the respective clique, the higher is the chance of that link to be discovered.

The described mechanisms introduce a strong *path dependence* in terms of past decisions and outcomes (which cliques to concentrate effort on and which links become discovered earlier) driving further results (which technology combinations become invented first). Given that the present study is a model of discovery, once a certain product type is discovered along one of its technological trajectories, the related pressure from the demand side disappears. We are only interested in first product type discoveries and those are analyzed in terms of GPT adoption. Though the history of innovation has many examples when new products were displacing the existing ones (smart phones against standard mobile phones, alternating current against direct one or Video Home System (VHS) against Betamax), this has normally had to do with functional superiority (where it becomes increasingly difficult to compare goods in satisfying exactly the same need) or cost advantage, which are not the focus of the present work. In contrast, we argue that if a technology becomes adopted in as many products as possible at the period of first invention, this does not only give it time and cost advantages but also allows it to become a new GPT.

We model agents in a very simplified way assuming no heterogeneity or interaction among them.[17] Once certain knowledge piece is discovered, it is upon the knowledge property, and not the agents, whether everyone or none of them gets access to this knowledge. Similarly, coordination is made not with respect to which agents shall

[17] Similar to production costs, we leave this aspect aside of the model to concentrate on the technology network effect first. In an extension, it will be certainly interesting to explore the issue of heterogeneity and interaction among those agents.

try to discover which technology link, but in terms of which technology clique one shall try to discover first (see Section 2.2.3). Thus, one can think of a 'representative' agent having the same incentives (V) and difficulties (d) in R&D process.[18] Also, no budget constraint for the agents is considered.

The notion of time is also present in our model and is kept simple. In particular, at period $t = 1$ agents start discovering new technological combinations (as aforementioned, none of them is present at the beginning of the simulation) and at each period can apply effort only to *one* way of producing a product. In this way, the model runs until all visible product types become producible (discovered in one of the production ways).

2.2.3 Factors affecting GPT adoption

Knowledge diffusion

One of the key questions to address in the case of knowledge discovery process is whether and to what extent does this knowledge diffuse to other products. A historical example could be a steam engine which was initially invented to pump water from coal mines, but with improvements spread to other applications finally powering locomotives and revolutionizing transportation. In the model we have already mentioned that some technology combinations can be utilized in more than one good and more than one way of production. A relevant question in such a case is whether the link between the two technologies I_{m-1} and I_m being discovered once (i.e. for one way of producing the respective product type, see Figure 2.5) opens this link for any other way of technological combination or product type. In the technology network context such a knowledge flow is contingent upon two conditions:

- functional similarity in combining the two technologies is sufficiently high to apply the same knowledge to other contexts: in the example of lenses and optics it means that this knowledge is directly applicable in cameras, telescopes, microscopes etc. This leads us to the discussion on technological standards and dominant design (for an overview, see Abernathy and Clark (1985) and

[18] Alternatively, one may think of a number of agents with a perfect information flow that act one at a time and all newly discovered knowledge becomes immediately available for everybody.

Anderson and Tushman (1990)), where being discovered once certain technology combination becomes *universal* and does not have to be rediscovered for other purposes (e.g., Global Positioning System (GPS) usage from military to civilian applications and from weather forecasting to time synchronization);

- the knowledge discovered flows freely within the population of agents, i.e. there are no firm- or institutional-based barriers preventing the flow of knowledge (so-called knowledge spillovers). This condition addresses the public good property (i.e., not appropriated by the owner) of knowledge coming back in the literature to at least Arrow (1962). This property is typically studied in the context of the network of agents (see Cowan and Jonard (2007)) and it's magnitude depends on the extent to which it is codified and the effectiveness of the mechanisms by which knowledge is protected, including the appropriability conditions (Dosi 1982).

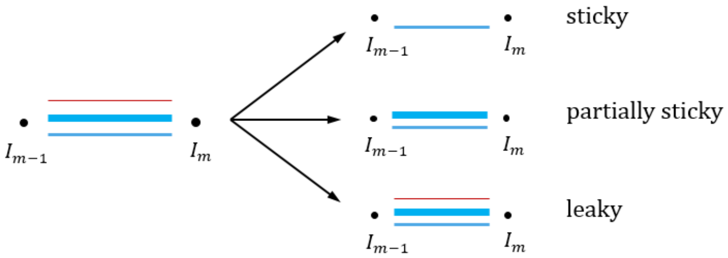


Figure 2.5: Knowledge diffusion regimes.[19]

We distinguish between three main regimes of knowledge diffusion (see also Figure 2.5):

1. *sticky knowledge*. In this regime there is either no functional similarity between products, or no knowledge spillovers preventing the possibility that knowledge discovery for one particular product (one of its production ways) can contribute to a discovery of any other product containing the same link;[20]

[19] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[20] In such a case, inventor literally has to 'reinvent the wheel' for every new product.

2. *partially sticky knowledge.* In this case, while the functional similarity between goods is still limited, the flow of knowledge is not. To distinguish that regime from the previous one, we make an assumption that limited functional similarity allows to apply the discovered knowledge to other product types, but only within the respective way of technological combination. This should reflect the intermediate status of the regime with imperfect knowledge diffusion;

3. *leaky knowledge.* This is the regime with perfect knowledge diffusion – once a certain technological link is discovered in any specific product, it becomes available in all product types across all ways of technology combinations.

It is worth pointing out that while we consider the aforementioned regimes of knowledge diffusion to be the result of innovation policy (affecting those through the technological standards and appropriability conditions), we treat those regimes as exogenous in our model, separately considering each of the three scenarios and analyzing implications for the emergence of GPT. In particular, we make the following proposition with respect to the effect that the knowledge diffusion has on GPT:

Proposition 1 *The larger the extent of knowledge diffusion, the more likely that a potential GPT becomes an input of many different product types at the stage of their discovery.*

Proposition 1 has the intuition that GPT, having in the present study the only distinct property of highest pervasiveness resulting in a large number of links connecting it to many other technologies in the network of intermediates, is also expected to have the largest number of links entering more than one product type in more than one way of production and, thus, must be the major beneficiary (among technologies) of the knowledge diffusion process.

Coordination of R&D efforts

Another mechanism, which plays a major role in technology emergence, is the decision heuristic of agents on how to innovate. Trying to solve a particular problem, an agent might concentrate on the easiest trajectory trying to use a lot of existing knowledge no matter for which initial purposes this knowledge was discovered. Alternatively, an agent might pursue an exploration strategy allocating its efforts equally among many alternatives. Let us take an example of a steam engine, which was initially competing with wind and water power. One can work on the improvements of wind, water, or steam power sources or concentrate its efforts on one.

Agents in the model do not know the difficulty of discovering a technological link and, thus, can take into account only the number of links yet to be discovered. However, the choice in favor of 'smaller' cliques (with least number of links yet to discover) may not always be optimal. First, given the strong uncertainty with respect to the difficulty of links, some cliques being larger in size may still be easier in terms of the amount of effort to be applied. Second, agents may prefer *knowledge breadth* over *knowledge depth* because of the interconnectedness between technological problems and the potential to utilize the gained knowledge in other applications. We introduce the factor of coordination in R&D effort through a logistic function determining the probability of the respective trajectory to be chosen by agents:[21]

$$Pr_i = \frac{e^{10\beta(L-L_i)}}{\sum_j^W e^{10\beta(L-L_j)}} \tag{2.2}$$

where a parameter $\beta \in [0,1]$ varies the scenarios from no (in favor of knowledge breadth) to the perfect coordination (knowledge depth), L stands for the maximum number of links to be discovered across all possible production combinations W and L_i is the number of yet undiscovered links in the trajectory i. This is illustrated in Table 2.1. Clearly, with $\beta = 0$ trajectories are chosen randomly without any account for already accumulated knowledge, while for $\beta = 1$ agents always will concentrate on the smallest clique. Intermediate values of β will squeeze probability distributions towards cliques with the least number of undiscovered links.

Table 2.1: An example of how probabilities are distributed for different β. L_j - number of edges yet to be discovered, W_i - number of ways a single product can be produced.

W	W_1	W_2	W_3	W_4	W_5
L_j	4 edges	2 edges	3 edges	4 edges	5 edges
$\beta = 0$	0.2	0.2	0.2	0.2	0.2
$\beta = 0.2$	0.019	0.850	0.110	0.019	0.002
$\beta = 1$	0	1	0	0	0

[21] Thus, at one period of time effort can be applied only to one technological trajectory in one of the product types. The fact that agents may not all coordinate in pursuing one technological trajectory is, thus, represented by no coordination.

We propose the following proposition with respect to the coordination of R&D efforts:

Proposition 2 *Preference for knowledge depth over its breadth fosters adoption of GPT under condition that knowledge spillovers are present between different technological combinations for different product types, and that those spillovers do not change over time.*

Proposition 2 consists of three parts: the first one conjectures that under no knowledge diffusion between different problems agents' coordination on any trajectory is purely random;[22] the second part postulates that in the presence of knowledge diffusion coordination may force agents to switch the trajectory in favor of the one with positive externalities[23] in the form of accumulated knowledge from a different product type. Since our GPT is potentially the most pervasive technology, those positive externalities are expected to be the largest for technological trajectories containing it, resulting in a higher adoption of the potential GPT in the first product type discoveries; finally, the third part ensures that those spillovers do not change over time: for example an agent in the model discovers a technological link which can be utilized in this period in one way of technological combination and one product type only, but many years later people find a different application for this technological combination in a different product type not considered before. In such a case, the time gains an importance in our model, while for invariant knowledge spillovers discovery of the link (or more generally, pursuing the technological trajectory with this link) may look not as attractive originally, this changes if the spillovers alternate over time.[24]

Potential technology network structure

Another core factor affecting the knowledge discovery process is the structure of technological network. Apart from the number of product types (N), intermediates

[22] Remember that earlier we assumed no relation between values, difficulty and technologies involved. Hence, in choice between two ways of production with the same amount of links agents will be indifferent, otherwise they pursue trajectory with smallest amount of edges given R&D coordination.

[23] In the words of Dosi (1982, p. 154) "complementarities among trajectories".

[24] In other words, we follow Carlaw and Lipsey (2006, p. 159) in that because of Knightian uncertainty agents do not have a "foresight about an unknowable future" and take decisions based on the externalities "as being constant at the current period level".

(M) and the ways of technological combination (W), this shall be affected by at least two more parameters: number of intermediate technologies in each technological clique forming a technological combination (clique size, CS) and the pervasiveness of the present intermediates within the product types. To keep the modeling simple, we assume in the baseline model that all product types in all technological combinations consist of the same number of intermediates,[25] while pervasiveness of other technologies is modeled via two opposite views. In particular, while GPT per assumption potentially enters all product types at least once and has the highest *potential* pervasiveness, other technologies (from 2 to M) may either all be very similar or very different in this respect. Based on the latter distinction we formulate the third proposition:

Proposition 3 *The larger the difference between the potential GPT and other technologies in terms of their technological pervasiveness, the more likely that the GPT becomes an input of many different product types at the stage of their first discovery.*

Proposition 3 is based on the intuition that the less potential synergy is concentrated between non-GPT technologies, the easier it must be for the GPT to fulfill its potential. Similar to knowledge diffusion, we consider the technology network structure as an exogenous factor. However, we do not argue that a policy maker may have an impact on technology network structure, as it represents the knowledge space itself; rather this network structure could be indirectly identified in order to adjust policy decisions.

Changes in expected profits

Finally, one may expect some effect on GPT adoption from the demand side. The expected profits for each product type proxy the priority from the side of society (both, consumers and policy makers) on which needs shall be satisfied first. Thus, any change in the rank of priorities can reflect either changes in preferences or institutions.[26] As an example of enforced preference change let us take the one

[25] In the robustness checks we relax this assumption highlighting that the main results remain valid.

[26] Institutional arrangement change incentives of entrepreneurs and investors to develop new products. For example, policy instruments introduced in the German energy sector made it profitable to concentrate on the renewable energy technologies (above all, wind and solar) (Herrmann and Savin 2016).

considered by Cowan (1990) on the nuclear power reactors. Because of the Cold War and fierce competition with the Soviet Union for the technological leadership, the U.S. government was heavily subsidizing the nuclear industry in the end of 1950s to foster building of the first commercial prototype and securing the global market. However, to enable such a swift discovery of the product type, a critical decision with respect to the preferred technological trajectory had to be taken (in this case between light water, heavy water and gas graphite). Given that "typically [...] when a technology is introduced its future payoffs are not well known" (Cowan 1990, p. 544), the choice has been made mainly based on knowledge accumulated by the U.S. Navy adopting the light water for submarine propulsion. As history illustrated, due to that exogenous shock introduced by the policy maker the market eventually became locked into the inferior technology.

To examine such an exogenous effect on the knowledge discovery process and the adoption of a potential GPT, but at the same time to keep the model simple, after a fixed number of periods (throughout the experiments we keep it equal to 100) for a certain fraction of product types we allow exchanging their expected profits, proxied by parameter *Value Dynamics* (VD) between 0 to 100%.[27] Thus, some less 'valuable' needs may instantly gain in priority and the other way around. All other characteristics of the model remain unchanged. Having introduced this mechanism in the model and keeping the example described by Cowan (1990) in mind, we formulate the following proposition:

Proposition 4 *Frequent changes in the rank of product type expected profits negatively affect the adoption of GPT and may lead to a technological lock-in in the long term.*

The intuition behind Proposition 4 is that due to instant change in the product type's expected profit its discovery becomes faster and essentially random with respect to the technological trajectory chosen, leaving not enough time to take an advantage of positive externalities through the knowledge diffusion. Thus, we conjecture that those changes in the rank of priorities diminish the effect of knowledge diffusion combined with coordination of R&D efforts and may lead to a technological lock-in.[28]

[27] This is done to prevent any volatility in the overall amount of effort the agents can apply to discover all product types in at least one production way.

[28] To address the possibility of a technological lock-in, we define as a *lock-in* the situation

2.3 Numerical analysis and model extensions

In what follows we describe how we set up the numerical experiment and which parameters we use as a default (Section 2.3.1). Afterwards, results of the simulation exercises (Sections 2.3.2-2.3.3) and robustness tests (Section 2.3.3) are presented.

2.3.1 Numerical experiment

At the beginning of each simulation, a large network of potential technological interconnections has to be generated. For this a subset of technologies (of the size CS) has to be sampled for each product and each of its ways of technological combination. In doing so, three conditions are ensured:

1. The sampling replicates one of the two sampling functions, which are chosen in line with Proposition 3. In particular, both sampling procedures start from ensuring that potential GPT enters first technological combinations for each product type. Afterwards, one either follows a highly skewed distribution function or sets the sampling probability of them to be constant. Analytically this is achieved by following one of the two probability distributions, respectively:

$$Mprobability_1 = \frac{1}{\sqrt{seq(1, M, M)}} \text{ or } Mprobability_2 = (1, seq(pm, pm, M-1)) \qquad (2.3)$$

   where $seq(a, b, l)$ generates a vector of equally distant elements between a and b of size l. As a result, the sampling function to the left in equation (2.3) creates highly skewed distribution, where the potential GPT still has the largest scope of application and is followed by a small subset of 'competitor' technologies also pretending to become included in many different product types and technological combinations. The sampling function to the right in (2.3), in contrast, generates a single 'champion' with other technologies having equal chance[29] to be included in any technological combination. Needless to

where the process of knowledge discovery is hampered (e.g., lower number of technological links is discovered), which eventually leads to no or delayed product discovery.

[29] This is proxied by the parameter $pm = \left(\sum\left(\frac{1}{\sqrt{seq(1,M,M)}}\right) - 1\right)/(M-1)$ chosen just to ensure that in both sampling functions GPT has the *same* potential pervasiveness (number of times being sampled for distinct technological combinations. For example, for $M = 100$ $pm \approx 0.178$.)

say that no technology can enter any technological combination more than once.

2. After all W technological combinations for all N final product types are constructed, they are rearranged randomly to ensure that GPT is equally present in all of them.[30]

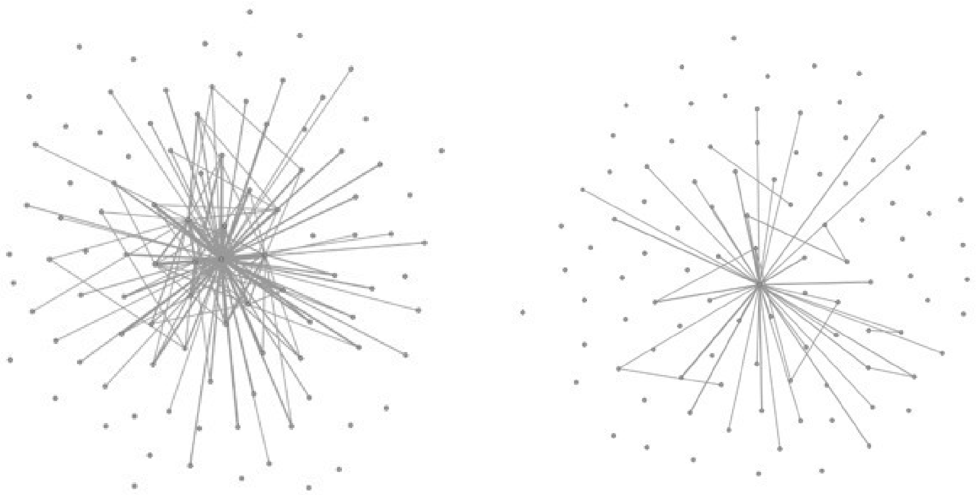
3. While creating the technological combinations, the code ensures no combination is repeated. The motivation behind that is to keep at least moderate technological differences between discovered goods in our model.

The exercise results in a complex weighted network, having both the bipartite (product-technology; presented in Figures 2.1-2.2) and multiplex (W alternative technological combinations consisting of the same number of nodes and links but having different link allocation; Figure 2.3) structure. As the default values we consider the number of product types $N = 60$, the number of intermediate inputs (technologies) $M = 100$, five ways of technological combination ($W = 5$) and four technologies to be recombined per product ($CS = 4$) so that the resulting network of possible technological links is a highly interconnected graph. To illustrate the difference between the two alternative sampling approaches described above, consider a network of technologies where the weight of an edge represents the amount of times this link is used in products. We examine the two network structures by filtering edges with the weight below $k = 5$. [31] This allows one to concentrate only on those edges which enter several product types. Clearly, in the case of equal pervasiveness the potential of other $M - 1$ technologies, a 'star-type' network structure is observed (see right plot in Figure 2.6). Almost all 'heavy weighted' links lead to a GPT. In the alternative network structure of core-periphy type there is a highly interconnected core of five-ten technologies including the potential GPT (left plot in Figure 2.6), which are also well connected to technologies outside the core (periphery). Henceforth, we denote the two alternative network structures as 'star network' and 'core-periphery network'.

[30] This is primarily done to avoid any strong assumption that GPT may benefit a lot from limited knowledge diffusion within just one way of technological combination.

[31] The exact value of k is chosen just for visualization convenience. For different parameters of the network, some different value of k may be chosen instead.

[32] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technolo-



a) *core-periphery* type b) *star* type

Figure 2.6: Potential technological networks after filtering links with a low weight.[32]

Then for the resulting networks we randomly distribute values (among final product types) and difficulties (among links). Afterwards, starting from period $t = 1$ agents apply R&D effort in a sequential order to discover final product types. To avoid any effect from specific product type order, each *cycle* the ordering of not yet discovered product types is rearranged randomly anew. For the basic model described, R&D agents continue inventing new technological links until for each product type at least one way of production is discovered.[33]

To start exploring the basic model with regard to Propositions 1-4, one first has to fix some further parameters we use. We assume expected profits of product types to be exponentially distributed with the parameter rate equaling 10, while the difficulties to discover each of the links are normally distributed with $\mathcal{N}(100, 25)$. These parameters, thus, are chosen to keep the numerical simulation sufficiently fast, avoiding discovery of many technological links within one cycle. Given the stochastic nature of the model and unless specified otherwise, in what follows results are reported for 50 restarts.

Describing the results, we primarily look on the (actual) pervasiveness of the potential GPT (percentage of first product type discoveries where GPT becomes an

gies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[33] As discussed before, any subsequent technological combinations, which can be discovered as a byproduct of the R&D process directed on discovery of different product types, are not taken into consideration.

input). Furthermore, to account for the fact that for different network parameters (like M, W or CS) the potential of GPT to enter all products *relative* to the potential of other technologies varies, we introduce an additional indicator, called $GPT_{Adoption}$, measuring to what extent GPT has fulfilled its potential in comparison to an average other technology in the technological space doing the same:

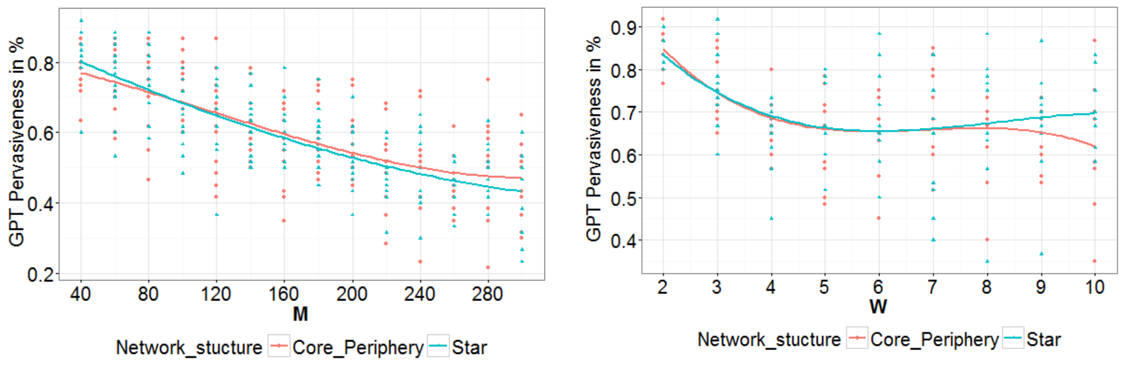
$$GPT_{Adoption} = \frac{Actual\ Pervasiveness_{GPT}}{\frac{1}{M-1}\sum_{m=2}^{M} Actual\ Pervasiveness_{m}} \Big/ \frac{Potential\ Pervasiveness_{GPT}}{\frac{1}{M-1}\sum_{m=2}^{M} Potential\ Pervasiveness_{m}} \tag{2.4}$$

Thus, $GPT_{Adoption}$ indicates not just how much more pervasive GPT has become in comparison to an average other technology (after the discovery process is finished and one calculates the 'actual pervasiveness'), but compares this ratio with the one using 'potential pervasiveness', i.e. in how many different technological combinations a given technology had a potential to be included.[34] Additionally, we report information on the discovered network size (in terms of number of links discovered) or amount of time spent by agents, which complement the picture on the intuition behind the results we obtain.

2.3.2 Results of the basic model

To understand the effect of network structure on GPT pervasiveness, one should look at how the variation in network parameters affects it under *ceteris paribus* principle for *core periphery* and *star* network structures (demonstrated on Figures 2.7 and 2.9). Increasing the number of technologies M naturally reduces the density of the technological network, thus, lowering the externality effects that favor GPT (left chart in Figure 2.7). A similar result with a level-off effect is obtained if we increase the number of alternative ways of production (right chart in Figure 2.7). Here the explanation is also simple. The more alternative ways of production we have the more competition a GPT has with other technologies and the harder it is for it to become pervasive. A level-off effect appears because we keep a ratio of

[34] For example: GPT had the potential pervasiveness of 60 and other technologies on average only 10, while actual pervasiveness is 30 and 3, respectively. The resulting value of $GPT_{Adoption} \approx 1.67$ implies that in comparison to its 'competitors' GPT has fulfilled its potential 67% better. Note here that the indicator value of 1 means that technologies have fulfilled their potential equally well.



(a) The effect of number of technologies.

(b) The effect of number of ways of production.

Figure 2.7: The effect of variation in the number of products (N) and the number of ways of production (W) on the GPT pervasiveness.[35]

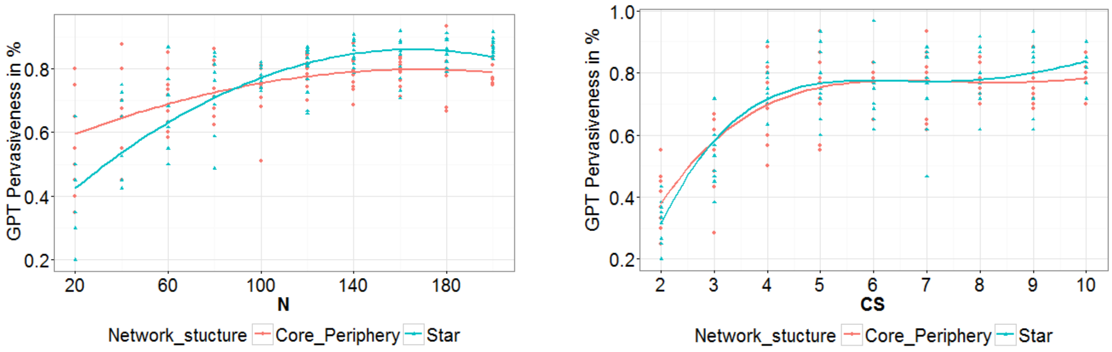
Note: This result is produced under no dynamics in product values. The network parameters used: $N = 60, M = 100, CS = 4, W = 5$.

products to technologies constant and at some time GPT starts to enter not one but several ways of producing the same product type in a potential network increasing the variance in the outcomes. An opposite trend is observed if one increases either the number of products (N, left chart in Figure 2.9) or the number of technologies each product can be made of (CS, right chart in Figure 2.9): as the network density rises leading to larger externality effects, GPT becomes adopted in larger number of final goods.

Hence, two conclusions can be made. First, one can observe a little difference between two alternative network structures, namely *core-periphery* and *star* types of network, thus, rejecting Proposition 3. Second, the more dense is the network in terms of the amount of weighted links, the more likely is the GPT adoption. By 'density' here we mean the amount of links with the weight larger than 1. It is clear that a typical definition of a network density employed from graph theory will not fit to our problem. This definition says that density is a ratio of existing links to all potential links. In our set up we are more interested in which links lead to GPT and which do not. Thus, we construct an index that sums the differences in those occurrences in favor and against GPT adoption, weights it according to the

[35] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[36] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



(a) The effect of number of products.

(b) The effect of clique size.

Figure 2.8: The effect of variation in the number of technologies (M), and cliques size (CS) on the GPT pervasiveness.[36]

Note: This result is produced under no dynamics in product values. The network parameters used: $N = 60, M = 100, CS = 4, W = 5$.

likelihood to encounter in the technological network and normalizes it to the total number of unique links in that network (Ψ):

$$Multiplicity\ Index = \frac{\sum_{\psi=1}^{\Psi} \omega_\psi \left[\max(\omega_\psi^{GPT} - 1, 0) - \max(\omega_\psi^{NoGPT} - 1, 0)\right]}{\Psi}, \quad (2.5)$$

where ω_ψ is the number of occurrences of a unique link ψ (the same pair of technologies) in our network of potential technological edges, ω_ψ^{GPT} is the number of times this link leads to cliques containing GPT and ω_ψ^{NoGPT} is the number of times the same link leads to cliques without GPT. In this way, we attempt to capture the effect of knowledge externalities between competing technological trajectories in our model. The larger the resulting *Multiplicity Index* the larger the actual GPT pervasiveness is expected to be.

Figure 2.9 illustrates how actual GPT pervasiveness depends on the index. Again little difference can be observed regarding two contrast network structures. The dependence is not linear and once the index exceeds a value of 1 GPT pervasiveness levels off around 80%. It is important to note that the index reacts to variation in all key network parameters discussed earlier and can serve as a good *ex ante* estimate of GPT adoption under leaky knowledge and coordination.

[37] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

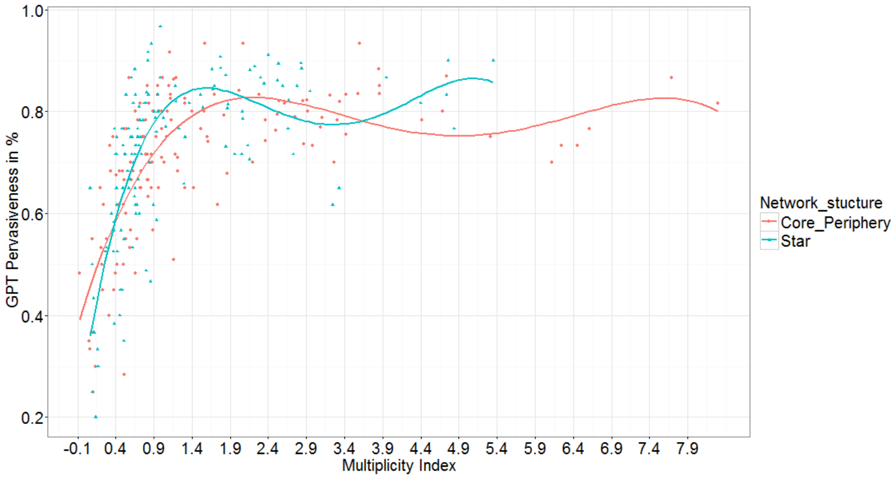


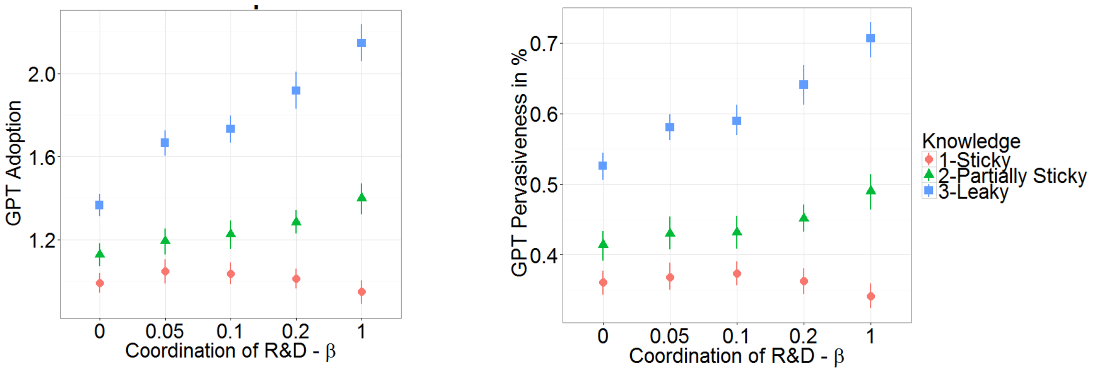
Figure 2.9: Multiplicity index reflecting resulting GPT pervasiveness under leaky knowledge regime.[37]

Note: We fit two polynomial lines for a better illustration purposes.

Figures 2.10 and 2.11 demonstrate the effect of the extent of knowledge diffusion and coordination of R&D efforts on GPT pervasiveness and adoption for different network structures. Since we do not observe a big difference among both network structures, let us concentrate on Figure 2.10. Start from the case of no coordination ($\beta = 0$): the more *leaky* is the flow of knowledge among technological combinations, the more pervasive is GPT and the better its potential is fulfilled. New knowledge embodied in discovered technological edges and applicable in different technological combinations becomes available for agents working on different technological problems and enforces earlier discovery of products containing larger proportion of links with such a multiple application. GPT is the main beneficiary of that 'knowledge propagation' process due to the network structure where by definition it potentially has the largest amount of technological links used in more than one product type. Thus, with leaky knowledge and no coordination GPT becomes a part of a much larger number of new products, while in comparison to an average competing technology GPT fulfills its potential 1,3 times better (Figure 2.10 at $\beta = 0$). This result fully supports Proposition 1.

[38] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[39] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

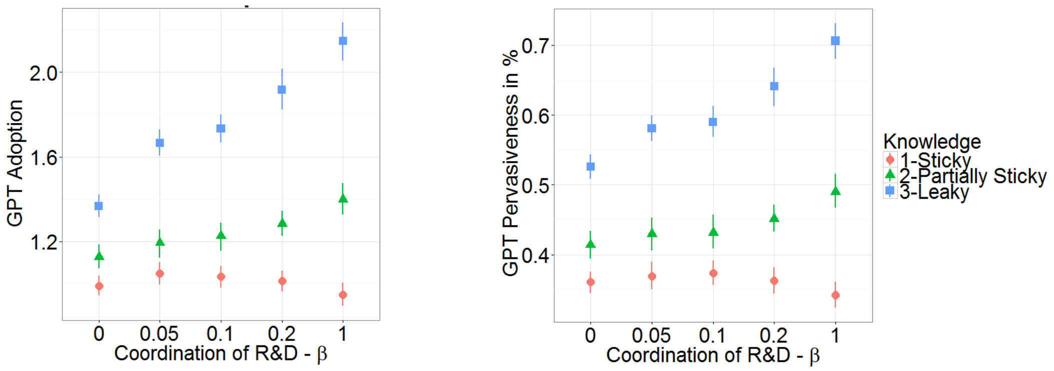


(a) GPT Adoption.

(b) GPT Pervasiveness.

Figure 2.10: The effect of knowledge diffusion and coordination on GPT adoption for core-periphery type network. Dots represent the mean values while vertical lines represent the range of values.[38]

Note: This result is produced under no dynamics in product values. The network parameters used: $N = 60, M = 100, CS = 4, W = 5$.



(a) GPT Adoption.

(b) GPT Pervasiveness.

Figure 2.11: The effect of knowledge diffusion and coordination on GPT adoption for star type network. Dots represent the mean values while vertical lines represent the range of values.[39]

Note: This result is produced under no dynamics in product values. The network parameters used: $N = 60, M = 100, CS = 4, W = 5$.

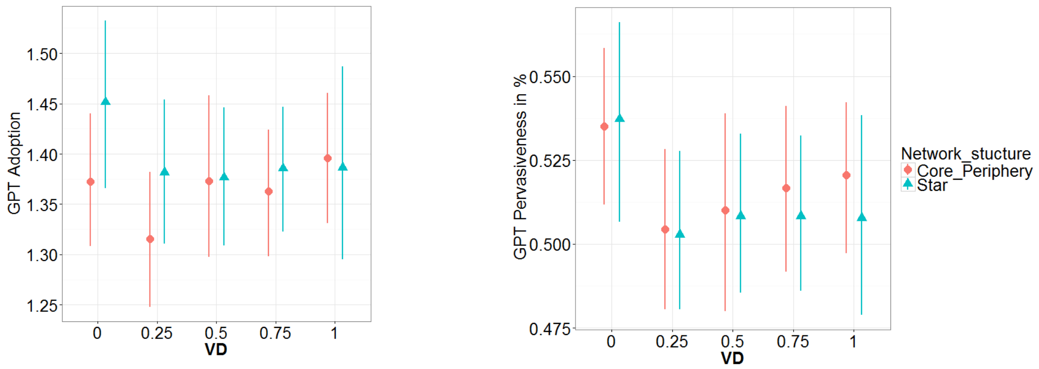
Furthermore, given that the knowledge diffusion propagates discovered solutions (technological links) to many other applications, it is worth testing whether coordination of R&D can strengthen GPT adoption under leaky knowledge even further and whether this is contingent on the presence of knowledge diffusion (Proposition 2). This factor is tested by varying the β parameter between 0 (preference for knowledge breadth) and 1 (preference for knowledge depth) for the three different knowledge diffusion regimes. Clearly, under leaky knowledge (Figure 2.10 blue square dots), an increase in coordination contributes to a larger GPT pervasiveness and adoption. The more 'sticky' is the knowledge, the smaller this contribution is until it vanishes completely confirming our Proposition 2.[40]

Finally, we explore the effect of variation in expected profits by setting the VD parameter equal to values between 0 (no variation) and 1 (all N product types change randomly their rank in expected profits every 100 periods). Results of the exercise are presented in Figure 2.12. The absence of a clear effect on GPT adoption has an explanation. In our model the demand side is interested in discovery of products (to be precise, first discovery for each product type satisfying a certain need), but puts no difference on which inputs shall be used to do so, leaving this choice to agents doing R&D. The agents, in their turn, pursue trajectories with lowest expected difficulty. As a result, this variation in expected profits has almost no impact on the agent's discovery choices. Hence, one has to reject Proposition 4 given that the network of potential technological interrelations is fixed and knowledge spillovers are constant over time.

What the variation in expected profits does affect, however, is the period of time within which at least one technological combination for each product is discovered (see right chart in Figure 2.13). Given that a high pressure from the demand side rotates between different product types over time, some more difficult edges become discovered much faster reducing the overall amount of time spent. A similar effect on the time of discovery have the knowledge diffusion and coordination of R&D efforts. Right charts on Figures 2.14 and 2.15 demonstrate for both network types that coordination of R&D reduces time, and this effect is enhanced if knowledge diffusion increases. Yet, the nature of those two effects is different. In the case of diffusion, present knowledge diffusion stands for the possibility of utilizing knowledge discovered elsewhere for a specific technological problem at hand. The coordination force leads to a focus on technological trajectories where knowledge is already accumulated and results in faster invention.

Another informative result is the size of the discovered graph reflecting the amount of knowledge accumulated in model's simulated economies. Knowledge diffusion logically increases the knowledge base discovered. Left charts on Figures 2.14 and 2.15 show that under all coordination regimes ($\beta = 0...1$) the diffusion has a positive

[40] Note here that varying the extent of knowledge spillovers, we keep those constant in time implying that if a technology combination becomes discovered and the knowledge spillovers regime allows the link to be applied elsewhere, this externality was taking place immediately (without any time lag). In an extension of our model (Section 2.3.3) we illustrate how such a time delay (in terms of knowledge externalities to be utilized) can be taken into account.



(a) GPT Adoption

(b) GPT Pervasiveness

Figure 2.12: The effect of value dynamics on GPT adoption. Dots represent the mean values while vertical lines represent the range of value.s[43]

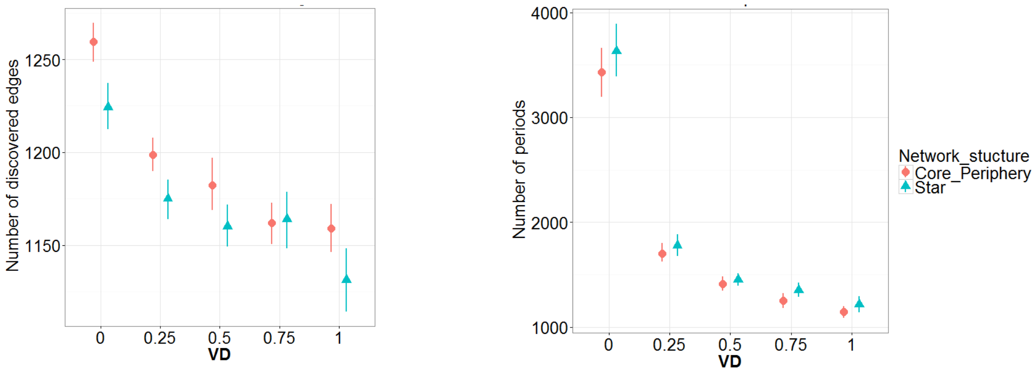
Note: The result is obtained under leaky knowledge and full coordination ($\beta = 1$). A similar result with quantitatively smaller values for GPT adoption is observed for no coordination ($\beta = 0$).

impact on knowledge base for both network structures. Even though agents were not aiming to discover all possible applications of a unique technological combination, this is done automatically.[41] The same chart demonstrates that coordinating R&D efforts (focus on knowledge *depth*) reduces the discovered base because agents always follow the (seemingly) 'least resistant' clique not trying to discover edges in alternative ways of production of the same product. Finally, left chart in Figure 2.13 demonstrates the negative effect of variation in expected profits on the amount of accumulated knowledge, which is due to the high pressure from the demand side, leading to fast product discovery, preventing agents to work more on different technological trajectories. This result is important to understand our findings for the technological network growing over time.[42]

Thus, one could conclude that in order to invent all products in a fastest way and promote adoption of potential GPT, one shall promote knowledge diffusion, stimulate agents to concentrate their innovative efforts on the technological trajectories with largest amount of accumulated knowledge and in parallel stimulate rotation in the demand side pressure towards discovering distinct product types. Yet, as we show in Section 2.3.3, such a conclusion would be too delusive in the long term perspective.

[41] If in contrast, we would have counted only all unique edges (between unique pairs of technologies), the presence of knowledge spillovers would result in the smallest network discovered.

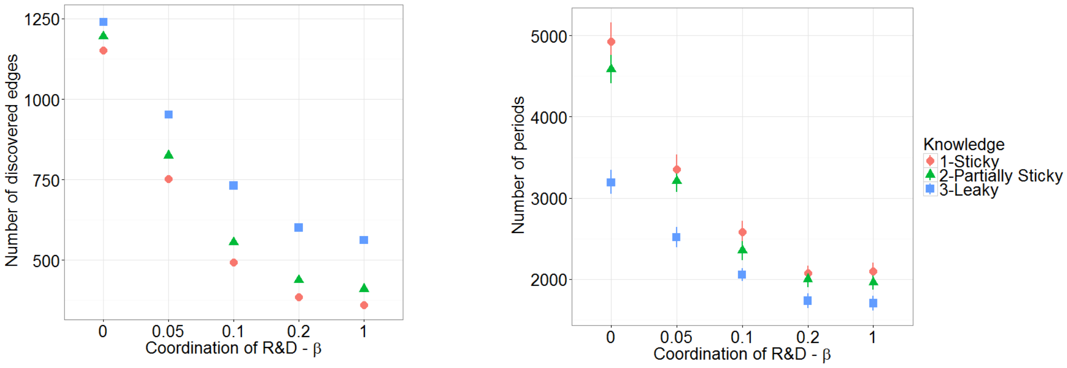
[42] Note that when coordination of R&D efforts is switched on, variation in profits has no clear effect on the discovered knowledge base since under coordination agents quickly start disregarding alternative trajectories.



(a) Number of discovered edges.

(b) Number of periods.

Figure 2.13: The effect of value dynamics on discovered graph and time of discovery. Dots represent the mean values while vertical lines represent the range of values.[44]



(a) Number of discovered edges.

(b) Number of periods.

Figure 2.14: The effect of knowledge diffusion and coordination on time of discovery and discovered graph for a core-periphery network type. Dots represent the mean values while vertical lines represent the range of values.[45]

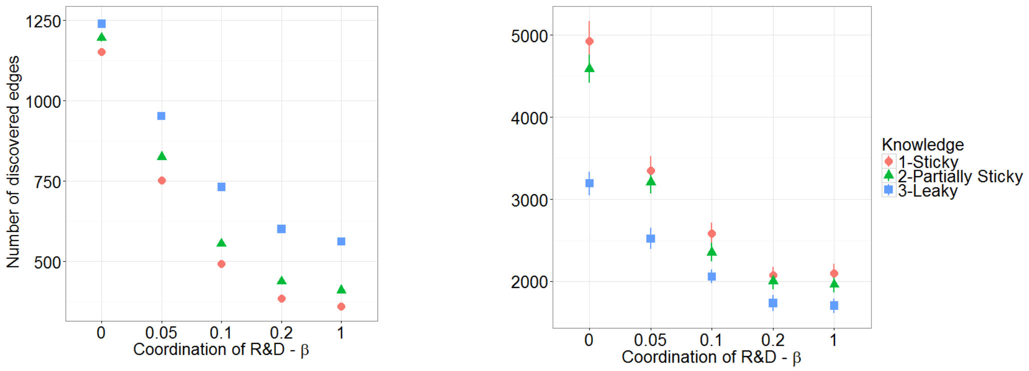
c

[43] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[44] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[45] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[46] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



(a) Number of discovered edges.

(b) Number of periods.

Figure 2.15: The effect of knowledge diffusion and coordination on time of discovery and discovered graph for a star network type. Dots represent the mean values while vertical lines represent the range of values.[46]

2.3.3 Growing technological network and modeling the arrival of new ideas

Up till now in our model all ways to produce a singe product were known or visible to agents *ex ante* and the discovery process stopped when at least one technological combination was found for each product. However, innovation process is dynamic and during the course of technological progress we come up with new ideas of new products and new ways of technological combinations. The use of railroads would be completely different from what we see today without a combination with a steam engine which allows the development of locomotives - something that has never been thought of before. Therefore, in the following we relax the assumption of fixed number of technologies (as in Section 2.2.2) and allow a visible network to grow (both in terms of number of visible technological combinations for a given product type, but also in terms of new product types/needs arising) calling this scenario ’*growing technological network*’. This scenario is logically close to the description by Arthur (2015, p. 140) of an economy as a complex evolving system, where ”structural change is [...] a chain of consequences where the arrangements that form the skeletal structure of the economy continually call forth new arrangements”.

We implement this extension into the model by adding a third layer to our multiplex technological network (Figure 2.16). The model then constitutes *discovered* network (consisting of combinations already discovered by agents), *visible* network (links that agents become aware of, i.e. *realize* those links as we will call henceforth; so far we were considering it to be the entire potential network and fixed over time) and a third *potential* network (all possible technological combinations, including visible ones but also those that agents are not yet aware of).[47] Thus, invisible network contains hidden ideas on new possible technological recombinations of existing but also new product types.

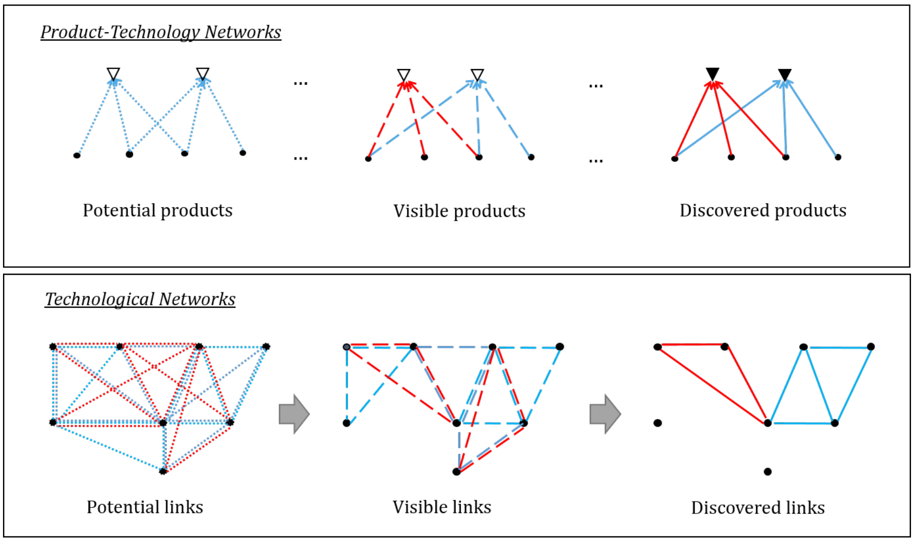


Figure 2.16: Three states of technological links.[48]

While the edge transition from visible to discovered state has been addressed in detail in our baseline model, here we discuss the transition from the invisible state to the visible one. In other words, we model the arrival of new ideas to our agents. This process is contingent on the knowledge being already accumulated by them. Thus, the growth of a visible network is highly dependent on a size and a structure of a discovered one. In particular, agents tend to learn about new possible product types or new ways of production of known product types depending on the extent they are using constituent technological combinations. One important difference of the mechanism making links visible to the one transforming them into discovered ones is that edges become visible in cliques, while links become discovered individually through practical tests more like an applied knowledge. The second difference comes from our assumption that the process of recognizing new technological combinations requires no R&D effort from the agents.[49] In particular, agents can recognize a new

[47] Obviously, the latter network is most general one, while the former two represent its fractions (discovered network - part of the visible one, while visible part of invisible potential network).

[48] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[49] The process can be better compared with 'Eureka' moments preceding the application of R&D

technological combination v at each period of time if the probability $Pr_v \in U[0,1]$:

$$Pr_v < \alpha \exp\left(v(s_1 - 1)\right) + \beta \exp\left(v(s_2 - 1)\right) + \gamma \exp\left(v(s_3 - 1)\right), \tag{2.6}$$

where s_1, s_2 and s_3 are shares of discovered, visible and invisible links[50] in the technological clique v, while α, β, γ, v are parameters specifying the function's shape, so that it increases exponentially and the more links in v are visible and discovered by agents.[51]

Figure 2.17 illustrates the equation (2.6) for different shares of s_1, s_2, s_3. Suppose a product consists of five technologies, which makes a clique size of ten edges. If all links of this clique are visible in other products than its probability to be recognized by agents as visible is approximately 10%. If 70% of links are already discovered and remaining 30% are visible in other products than this probability raises to almost 23%. All exact values can be found in Appendix A, Table A.1. Thus, if agents are aware of the fact that a pair of technologies has an innovation potential (it is visible for agents), or they have already discovered that link, it is more likely they will once recognize that there is a new product type that can be created. Note here that even if all edges in a clique are completely unknown to agents this probability is different from zero. We can't deny the fact that there is always a chance of the arrival of new radical idea from different technological paradigm. This chance increases if all respective links are already visible in different products, and by the time 100% of those links are discovered the clique v becomes visible with certainty. In the words of Atkinson and Stiglitz (1969) or Nelson and Winter (1982) agents search locally for new knowledge trying extensions of existing one close to what they already possess and use in some space of technological characteristics. The model now runs until agents discover all product types that they see.

The exercise below takes into account the results of our baseline scenario. We fix knowledge as 'leaky' for the rest of the analysis given that without knowledge

effort.

[50] Once a new link has become either visible or discovered, one automatically updates the probability of yet invisible technological combinations containing this link to become visible.

[51] In particular, we set $v = 5$, $\alpha = 0.01$, $\beta = 0.1$, $\gamma = 1$.

[52] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

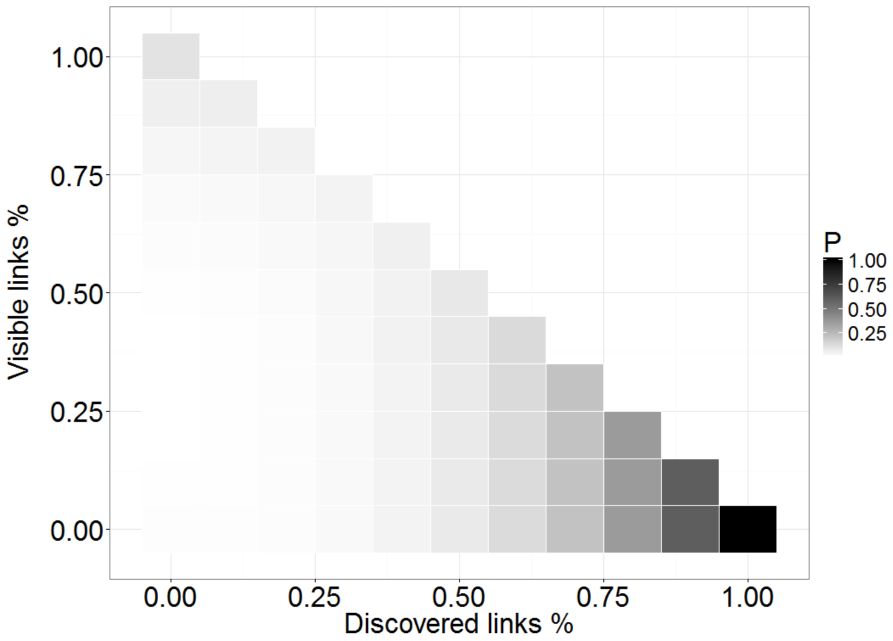


Figure 2.17: An Example of a probability of a product to become visible.[52]

Note: The figure illustrates the probability of a product consisting of five technologies and ten edges to become visible. Values on the x-axis denote the percentage of discovered links in the clique, while values on the y axis show the percentage of visible links. The remainder is by definition - invisible. Color intensity indicates the probability of this clique having certain fractions of visible and discovered links to become visible to agents. All exact values can be found in Appendix A, Table A.1.

diffusion the role of other factors vanishes and discovery process turns random. We also consider only core-periphery network structure as a more realistic one where a potential GPT is followed by competitors.

Results for growing knowledge base

In this scenario we extend our baseline model adding a second generation of products that is not visible to agents from the beginning. New generation has the same ratio of products to technologies, namely $N = 60$ and $M = 100$ and mainly consists of new technologies that were not present in the first generation (using technologies from 96-195 and the potential GPT itself). Hence, there are only six common technologies between the first and second product generations (see Figures 2.18 and 2.19 respectively). Those two generations are meant to represent two distinct technological paradigms with former of complexity $CS = 4$ and the latter of $CS = 5$ reflecting the fact that consumer products become more complex over time. The value distribution of product types in the second paradigm is taken twice larger than in the former one, primarily to compensate for the complexity and boost the simulation speed.

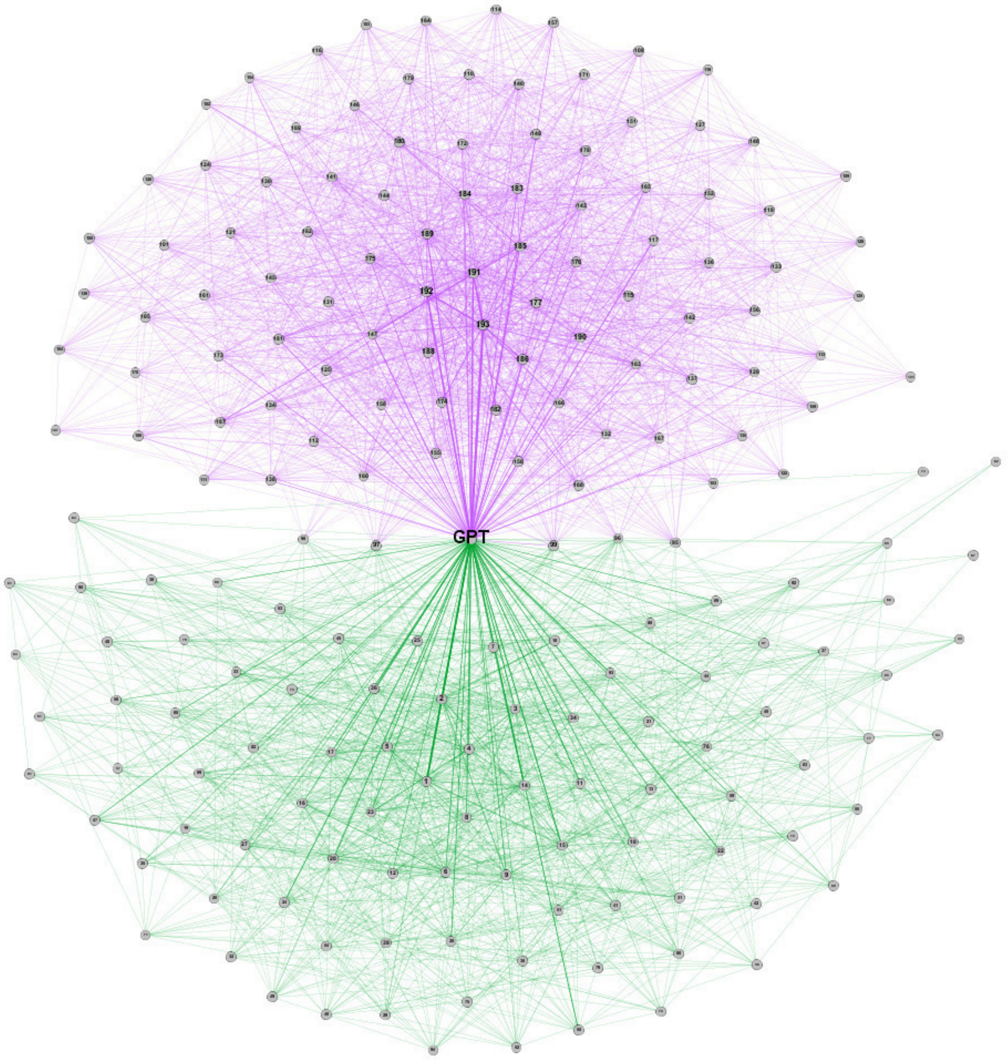
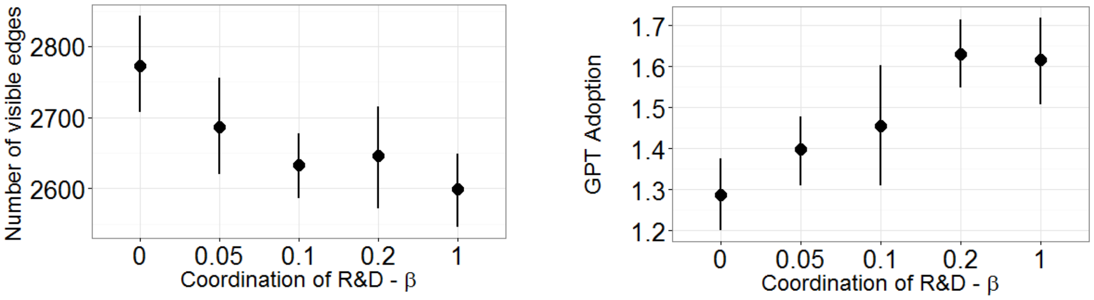


Figure 2.18: Technological network of intermediaries for the case with two product generations and a single GPT. Green network represents initially 'invisible' technological combinations or ideas that economic agents do not have at the start of each simulation.[53]

The Figure 2.19b demonstrates that coordination of efforts instead of a strictly

[53] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[54] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



(a) Number of visible edges.

(b) GPT Adoption.

Figure 2.19: The effect of coordination on GPT adoption for the case with two product generations and a single GPT. Dots represent the mean values while vertical lines represent the range of values.[54]

positive (recall left-hand side of the Figure 2.10) exhibits a 'level-off' effect on GPT adoption. A key to understand the nature of this finding is on the upper right chart of the figure. Being more focused on knowledge depth strategy (and, as a consequence, discovering less technological links), one reduces the size of the visible network in the second product generation, thus, limiting the externality effect that one can exploit. In other words, in the dynamic perspective high coordination hampers agents in discovering technological combinations with more pervasive technologies (including potential GPT). The positive effect of coordination in the first product generation is compensated by the negative effect in the second generation because an agent cannot discover something it has no idea about (yet).

Results for two GPTs with different product generations

The negative effect of coordination becomes more pronounced if we consider the second product generation to have its own potential GPT (for an illustration see Figure 2.20). Results of the experiments are demonstrated in the Figure 2.21 and 2.22 focusing on the GPT adoption in the second product generation since for the first GPT the results repeat the pattern described in Section 2.3.2. Figure 2.21 demonstrates that by increasing coordination the size of the visible network falls. As a result, we observe a pattern similar to an inverted U-shape form illustrating the adoption of the second GPT in coordination. Thus, while moderate coordination is better than no coordination at all, this trend changes its direction once coordination approaches its maximum level, demonstrating, that neither no nor full coordination is optimal. This trade-off between exploiting externality effects and keeping the size of the visible technological network large enough (a sort of proxy for 'new ideas' in our model) reminds the classical ambidexterity trade-off known in the literature on organization theory (see, e.g., the seminal paper by March (1991)).[55]

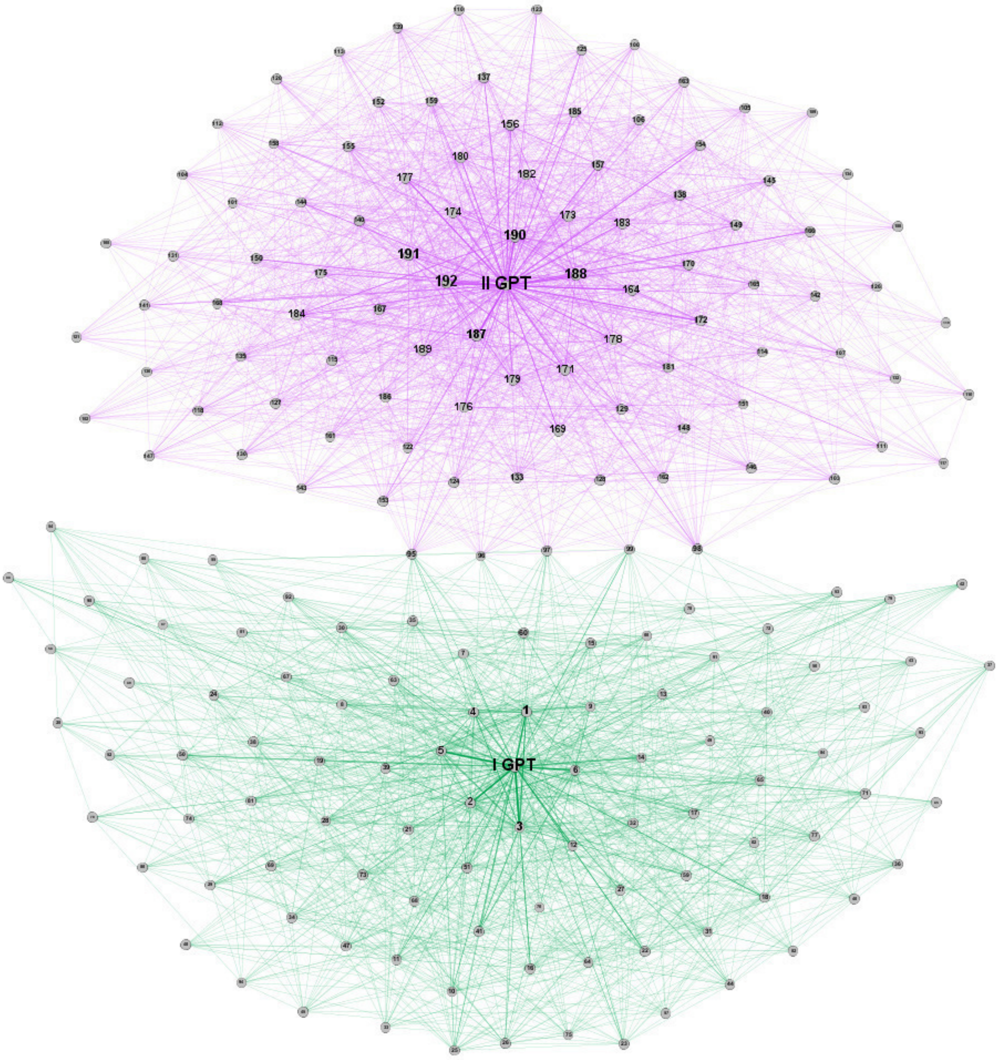


Figure 2.20: Technological network of intermediaries for the case with two product generations and two distinct GPTs. Green network represents initially 'invisible' technological combinations (ideas) and a new GPT that economic agents do not have at the start of each simulation.[56]

[55] According to (March 1991, p. 72), "choices must be made between gaining new information about alternatives and thus improving future returns (which suggests allocating part of the investment to searching among uncertain alternatives), and using the information currently available to improve present returns (which suggests concentrating the investment on the apparently best alternative)".

[56] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

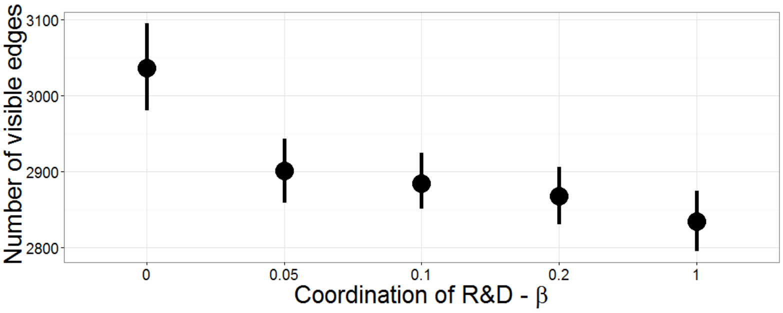


Figure 2.21: The effect of coordination on GPT adoption for the case with two product generations and two distinct GPTs on the number of visible edges. Dots represent the mean values while vertical lines represent the range of values.[57]

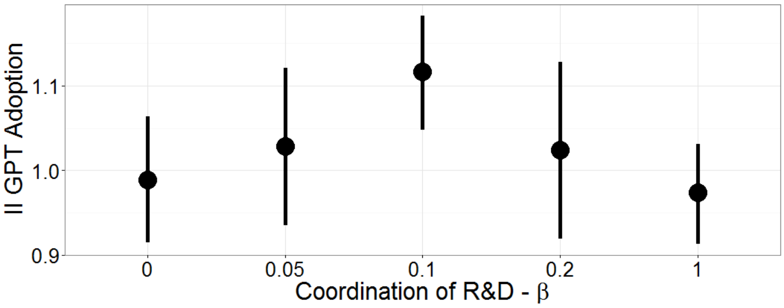


Figure 2.22: The effect of coordination on GPT adoption for the case with two product generations and two distinct GPTs on the adoption process.[58]

As we know from Section 2.3.2 the variation in expected profits has no clear effect on GPT adoption in the short term while in the long term it reduces the size of the discovered network and the period of time spent on the discovery process (Figure 2.13), it is easy to foresee that the effect of value dynamics on GPT adoption in the long term is strictly negative.[59]

[57] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[58] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[59] For brevity reasons we do not include those results here, but they are available on request.

Robustness tests

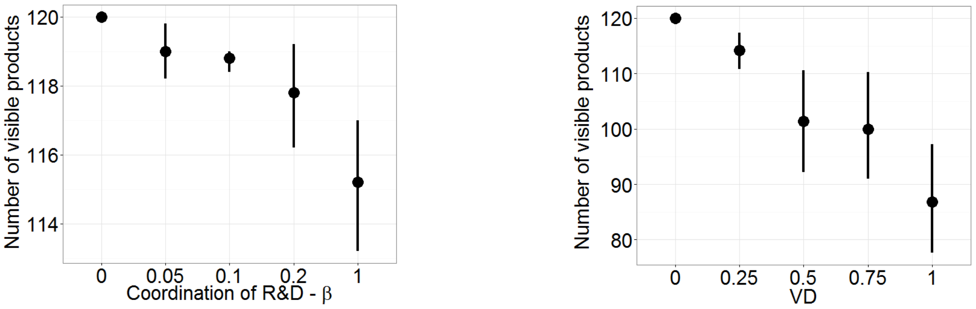
It is important to point out that robustness test were conducted in all the numerical experiments described above for different clique sizes of different product types within one technological paradigm. Though on average the GPT adoption has reduced over the experiments (because GPT was randomly allocated between combinations of different size), the major results hold. We also conducted the experiments with different distributions of difficulty and expected profits. Among others, we considered the difficulty distribution being exponential reflecting the situation where only few innovations are hard to discover. Additionally, we have considered the case where certain percentage of technological links are given 'for free' implying that their difficulty equals zero. Those modifications affect the speed of discovery process but do not change our findings with respect to Propositions 1-4. Furthermore, we considered alternative parameters for equation (2.6) and also modified the shape from exponential to logarithmic one. Our main findings do not change as long as our main assumption that arrival of new ideas depending on the visible and accumulated knowledge holds.

2.4 Stylized facts

Apart from theoretical results on GPT emergence, we would like to point here some of the stylized facts of innovation process that our model replicates and illustrate some steps in empirical verification of our predictions. In Proposition 4 we have already mentioned the *lock-in* effect. This effect is replicated by our model in the scenario with growing technological network, where low amount of knowledge accumulated (either due to coordination of R&D efforts or variation in expected profits) leads to many product types remained neither realized nor discovered (Figure 2.23).

The model also demonstrates clustering of innovations in time (see Silverberg and Lehnert (1993) for a literature review), which in the model is represented by the discovery of products. To ensure that we replicate the procedure by Silverberg and

[60] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



(a) Number of visible products.

(b) Value Dynamics.

Figure 2.23: *Lock-in* effect in the case of two product generations and two distinct GPTs.[60]

Verspagen (2003) in generating an innovation time series[61] (periods when a new product type has been discovered, see Figure 2.24 for an example) and sequentially fit the Poisson and negative binomial models with linear, quadratic and cubic time trends as explanatory variables. The linear and quadratic time coefficients are significant at the 5% level, while the negative binomial model is consistently preferred over the Poisson one for all the three model specifications.[62] The simple explanation of the temporal clustering of innovations by our model is that those innovations share a common knowledge (technological edges), and agents coordinating their R&D activity exploit the knowledge externalities by discovering several product types within a short period of time.[63] This confirms ideas dating back to the concept of 'technological convergence' described by Rosenberg (1976) and shows the power of knowledge diffusion mechanism.

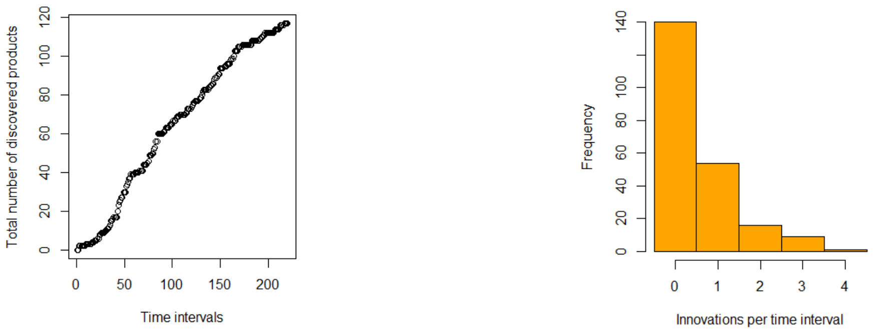
In addition to the aforementioned facts, we compare structural similarity of the networks we generate with those we observe empirically. In particular, networks of technologies and product relatedness are of interest. For many reasons (mainly because of *invisible* and *visible* networks representing an ex-ante state of knowledge we can only hypothesize), we concentrate on the final ex-post *discovered* networks

[61] Note that by definition of a time period in our model, it is unlikely two innovations to happen at the same period. Therefore, without loss of generality we consider each twenty periods as one time interval.

[62] This finding holds for the majority of parameter values we use. The notable exception is variation in expected profits. If those change often, the process of discovery becomes close to linear in time.

[63] This is particularly true if the technological link had a relatively large difficulty and, thus, likely remaining one of the last barriers to introduce a new product.

[64] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



(a) Accumulation of the discovered products over time.

(b) Value Dynamics.

Figure 2.24: Number of innovations per time interval. The result obtained for the scenario with two distinct GPTs (N=120, M=195, CS=4, $\beta = 1$ and leaky knowledge). The left figure illustrates bursts in the number of innovations (y-axis) in time (x-axis), while the right one is the histogram of innovations, i.e. frequency of periods with 0,1, 2 or more innovations.[64]

drawing parallels with the works of Hidalgo et al. (2007), Hidalgo and Hausmann (2009) and Boschma et al. (2014) based on trade and patent data. Hidalgo et al. (2007) and Hidalgo and Hausmann (2009) consider a product as a combination of some hidden technological capabilities that economic agents possess. In our model these capabilities are represented by technological combinations (a link between two technologies being discovered). Boschma et al. (2014) investigates networks of patent International Patent Classification (IPC)[65] classes and their relatedness providing structural characteristics of those networks constructed by employing similar techniques as Hidalgo et al. (2007). Here we focus on technological networks, where technologies are seen to be related if they share a patent. Thus, we see our model as a mechanism by which these empirical networks of products and technologies are formed.

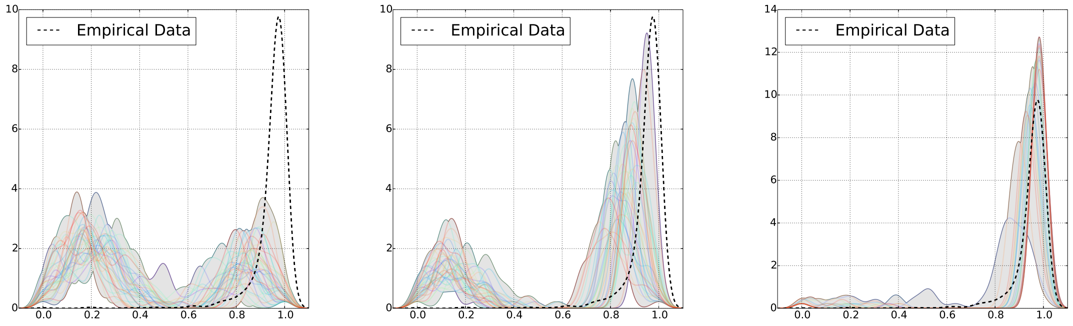
There is no consensus in literature about graph comparison due to the nature of the subject of study. This problem is tackled differently across scientific fields (Mernberger 2011). In particular, three main strategies are identified: exact graph matching, inexact graph matching, and feature-based approaches. The latter is preferred because it allows observing empirical networks on an aggregate level, where each IPC class or product is already a collection of knowledge pieces or smaller products. Hence, we expect our simulated graphs, matching some general structural characteristics of empirical graphs, to be an indication that forces behind the formation

[65] International Patent Classification

of those graphs are similar. We choose four features of graphs to compare: *density*, *degree assortativity*, and *degree distribution*. Density tells us about the interconnectedness (interrelatedness) of technologies in discovered products. Degree assortativity reveals whether more pervasive technologies tend to be connected with less pervasive ones. Average clustering illustrates to what extent do technologies cluster. Finally, degree distribution reveals possible 'hubs' - few technologies dominating the others in terms of their interconnectedness. Note that comparison of those network features does not require to have the same number of nodes in the simulated and empirical networks. We compare our product networks to the 'product space' taken from the atlas of economic complexity (Hidalgo et al. 2007, Hidalgo and Hausmann 2009). In particular, the data comes from the website of the observatory of economic complexity (Simoes 2016). Our product networks have similar high density and are disassortative (Table 2.2). Figure 2.25 demonstrates how degree distribution changes for simulated graphs with different knowledge diffusion regimes. Only *'leaky'* knowledge ensures products to be technologically highly interconnected as in empirical networks.

Table 2.2: Comparison of simulated (product) and empirical ('product space') networks

Network parameters	Empirical Network	Simulated Network Mean(Standard Deviation)
N of nodes	773	60(0)
N of edges	282402	1720.4(23.8)
Density	0.967	0.972(0.013)
Degree assortativity	-0.041	-0.042(0.004)



(a) *Sticky* (b) *Partially leaky* (c) *Leaky*

Figure 2.25: Kernel-density estimations of degree distributions for product networks compared to the 'product space' (dashed line) under different knowledge diffusion regimes.[66]

To validate the produced technological networks we compare their typical ex-post (discovered) structure (for core-periphery structure, N=60, M=100, CS=4) with empirical networks of patent classes (kindly provided by P.-A. Balland for United States Patent and Trademark Office (USPTO) data for the years 1976 - 2010). A patent class can be considered as a piece of knowledge needed for production of goods. Characteristics comparison is presented in the Table 2.3. Both networks have similar density, implying almost the same ratio of existing links to all potential links. They are also similarly disassortative meaning that technologies with high degree centrality tend to be combined with technologies with low degree centrality. This result demonstrates again that technologies become pervasive only when they are combined with many infrequently used ones. We also report a typical for empirical networks heavy-tailed degree distribution (Figure 2.26). Figure 2.26 b) also illustrates that empirical technological network has a core-periphery structure: there are important 'gateway' technologies that are connected to the core and peripheral ones.[67]

Table 2.3: Comparison of simulated (technological) and empirical (IPC) networks

Network parameters	Empirical Network	Simulated Network Mean(Standard Deviation)
N of nodes	438	100(0)
N of edges	12295	292(12)
Density	0.068	0.059(0.009)
Degree assortativity	-0.152	-0.150(0.026)
Average clustering	0.479	0.501(0.038)

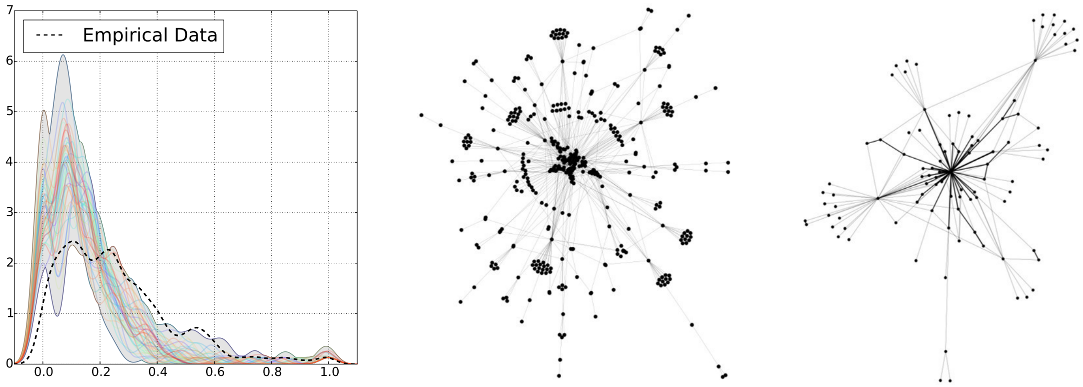
2.5 Conclusions on modeling emerging GPTs

General Purpose Technologies proved to be crucial for the process of technological development providing a structure for other technologies and supporting economic growth. Earlier GPT models emphasized their influence on 'productivity paradox',

[66] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.

[67] If in contrast one looks at the star type network, such gateway technologies are absent.

[68] Reprinted with the permission from Korzinov, V., Savin, I. (2018) General Purpose Technologies as an emergent property. Technological Forecasting and Social Change, Vol. 129, 88-104. © 2017 Elsevier Inc. All rights reserved.



a)

b)

Figure 2.26: a) - Kernel-density estimations of degree distributions for technological networks compared to empirical network (dashed line) of IPC classes. b) - Visualization of a network of IPC classes among patents (left) and simulated (right) technological networks.[68]

accounting for a 'residual' in aggregate production functions, focused on GPTs' evolution under a stream of innovations as well as explained the 'dual inducement' mechanism between GPT and its application sectors. Despite this extensive body of literature, the emergence of these technologies deserved only a little attention so far. Our study sheds light on this issue by concentrating on the pervasive nature of GPTs. Introducing its emergence as a continuous process of technology adoption, we look for mechanisms fostering technological convergence employing methods of network science representing knowledge discovery as a growing technological graph.

Our results demonstrate that *knowledge diffusion* is absolutely necessary for GPT emergence, since being discovered once the knowledge spills over to many other applications, benefiting most those technologies, having the potential to be used in many distinct products and industries.

The structure of our knowledge should have a *sufficient density* for a GPT to become pervasive, where by structure we mean interconnectedness of our ideas and by density – interchangeability of our knowledge among various applications. With the novel metric (*the Multiplicity Index*) we demonstrate how to measure that density. Given the presence of knowledge spillovers and sufficient density of the network structure, *concentrating on technological trajectories*, where more knowledge is accumulated also favors GPT in the short term. However, once the technology network is modeled as a growing knowledge base where agents become aware of novel possibilities to combine technologies through inventing simpler products, a trade-off between coordinating on existing trajectories and pursuing novel technological combinations emerges. This transforms the pure positive effect of coordination into an inverted U-shape form, echoing the classical ambidexterity trade-off between exploration and exploitation. Similar to firms in the organizational theory (see, e.g., Sidhu et al. (2007)), countries shall apply differentiated technological policy depending on whether the economy is in a more or less dynamic environment. Thus, in the 'path-following' catching-up process (Lee and Lim 2001) countries aiming to discover certain product types in the knowledge base where most of technological trajectories are known from experience of advanced economies will find exploitative strategy (high coordination on trajectories with most accumulated knowledge) most attractive. In contrast, if the economy is currently at the technological frontier, seeking to identify the next GPT, it shall put more focus on exploration of new opportunities and provide incentives for sufficient knowledge breadth. For the same reason, policy maker shall avoid supporting any specific product need, before the economic agents accumulate enough information on alternative ways of producing goods to satisfy that need and payoffs to adoption of the respective technologies. Otherwise, the choice of the technological trajectory turns random and due to the increasing returns to adoption described by Arthur (1989) the economy risks to be locked-in to inferior technologies.

Our model reproduces well-known stylized facts accompanying innovation processes such as S-shaped curve of technology adoption, temporal clustering of innovations in time and lock in effects. Furthermore, our model replicates many structural features of the empirical product graphs (Hidalgo et al. 2007) and those graphs constructed based on networks of relatedness between technological IPC classes (Boschma et al. 2014).

One shall also point out that the current analysis is limited in a number of ways. First, no production costs are taken into account. This together with explicit budget restriction on the side of agents shall provide a more complete picture of the technological competition, and help to explore 'growth bottlenecks' reported by Bresnahan and Yin (2010). We preferred to abstract ourselves from those issues here for the sake of clarity. Furthermore, so far we have neglected heterogeneity among agents in terms of their accumulated knowledge and possible cooperation/competition between them. All these aspects provide a natural direction to further develop the present model opening a fruitful trajectory of further extensions in the direction of GPT competition/succession.

Chapter 3

A patent search strategy for service robotics[1]

3.1 Modern general purpose technologies

Examples of famous general purpose technologies are three-masted sailing ship, steam engine (Rosenberg and Trajtenberg 2004, Crafts 2004), electricity (Moser and Nicholas 2004), ICT (Vuijlsteke et al. 2007) and currently bio or nanotechnologies (Shea et al. 2011, Lipsey et al. 2005). Another potential GPT of our time is robotics. In line with theoretical background (Bresnahan 2012) robotics is seen as a potential driver of the fourth industrial revolution, being considered in a cluster of technologies, together with artificial intelligence and big data. These technologies can alter modern production chains and organizational routines as well as global leadership. Robotics has all necessary characteristics of a general purpose technology. Due to its potential broad application it has a pervasive character entering many downstream products. Especially, a branch of robotics - service robotics (hereafter Service Robotics (SR)) - has a general application potential. It shows a significant technological dynamism. In recent decade a huge development is seen in robots and their applications (IFR 2016). International Federation of Robotics (IFR) estimates that the worldwide number of domestic household robots is rising up to 31 million between 2016 and 2019 (IFR 2016). Finally, robotics exhibits innovation comple-

[1] Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. doi 10.1007/s11192-017-2268-3.

mentarity while advances in robot development induce significant improvements in downstream sectors and the same holds in reverse. Advances in manufacturing of machines and new materials allows construction of better and safer robots.

On top of that robotics can be considered a structural technology as seen through the lenses of eleven characteristics highlighted by Lipsey (Carlaw and Lipsey 2011). While some of these characteristics are useful for modeling purposes others can serve as a criteria for detecting a structural technology. We elaborate on these criteria in the following numbering them in brackets. Robotics has been developed through the endogenous research and development process (1) and its efficiency increases gradually overtime (Ott 2012) (2). The use of robots spreads slowly in the economy. It will take time before many of the modern inventions such as driverless cars or service robots, will be fully commercialized and their markets will mature. Only after these technologies enter many sectors all the advantages that they bring will be reaped (3). In addition to robotics several potential non-identical GPTs exist nowadays (4). Advances in nanotechnologies and MicroElectroMechanical Systems (MEMs) may revolutionize many production sectors. Bio-technologies and chemistry developments in pharmaceutical industry promise a revolution in various medical applications and in a disease diagnosis and treatment. Moreover, robotics itself can be seen as consisting of classes of technologies. Figure 3.1 demonstrates technologies constituting robotics based on the patent data and technology classifications developed by Schmoch (2008).[2] In the words of T. Bresnahan (Bresnahan 2012) GPT itself is a cluster of technologies, which satisfies the fifth characteristic of Carlaw and Lipsey (5). On top of that modern robots heavily stand on the advancements and developments of previous GPTs, such as electricity and ICT (7). Invention, innovation and diffusion of robots in the economy involves many uncertainties (8). Firms, dealing with robotics, can not really maximize their returns over a life time of a technology, due to many uncertainties involved in its development path (9). For example, who would have known a decade ago that we will have real examples of driverless cars on our roads by 2013, when several US states passed laws permitting autonomous vehicles on their roads.

Thus we can see that robotics, especially considering its future developments in service sectors, has a great potential as an emerging disrupting technology. The

[2] Patents are taken from the PATSTAT database of the European patent office. The retrieval query was IPC class "B25J" or a substring "robot" in title or abstract of a patent. For more information see subsection 3.3

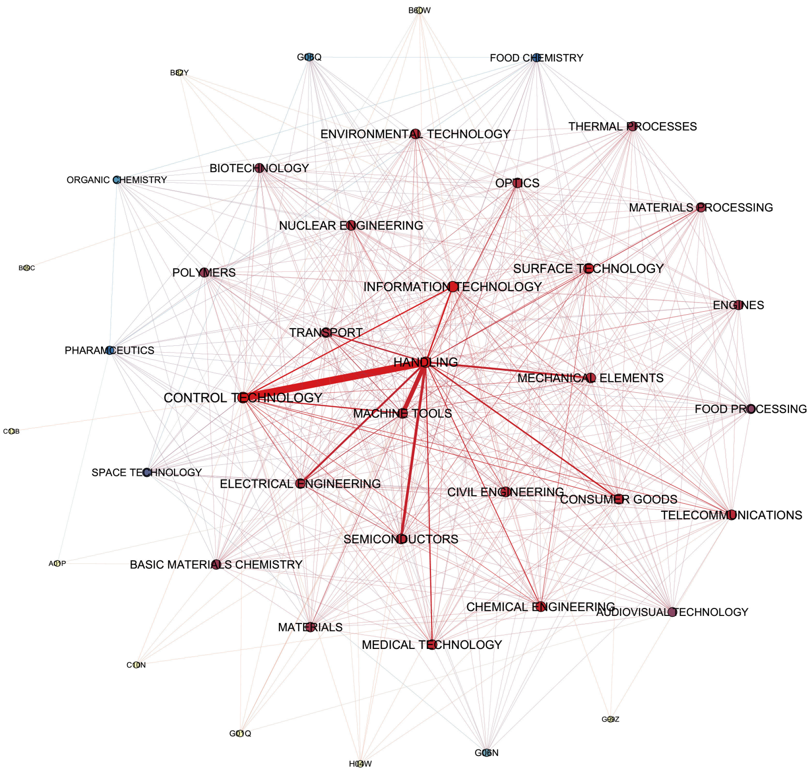


Figure 3.1: Robotics represented as a cluster of technologies based on the IPC classes of patents from PATSTAT database (version April 2016). Each node is a technology where technology-to-IPC concordance is taken from Schmoch (2008). Each edge connecting two nodes is a patent. Edge width reflects the number of patents.

following chapter demonstrates a methodology that helps to identify service robotics patents within modern relational databases as well as provides some basic descriptive statistic on patents identified with this methodology.

Innovation policies that address promising emerging technologies serve to reach macroeconomic objectives such as promoting sustainable growth and prosperity. They are legitimated due to the various uncertainties associated with new technological fields that result from coordination problems in complex innovation chains with scale economies, multilateral dependencies, and externalities. In order to develop effective policy measures, one has to carefully recognize emergence patterns and assess possible downstream effects. This is a demanding task since these patterns vary across technologies, time, scale, and regional and institutional environments. It is important that the policy advises rely on credible data sources that accurately depict early research and innovation results at the very beginning of value creation. However, as long as a new technology has not yet been specified within official statistical schemes, the identification of delineating boundaries in respective data bases is a nontrivial problem.

Service robotics is a current example of an emerging technology. The International Federation of Robotics (IFR) has been working on a service robot definition and classification scheme since 1995. A preliminary definition states that a service robot is a robot that performs useful tasks for humans or equipment excluding industrial automation applications. Industrial automation applications include, but are not limited to, manufacturing, inspection, packaging, and assembly (compare `www.ifr.org` and ISO 8373:2012). Service robots can be further subdivided into those for non-commercial personal use like domestic servant robots or automated wheelchairs, and those for professional commercial services, for which they are usually run by trained operators like fire-fighting or surgery systems in hospitals. Hence, SR contribute to both traditional and new types of services.

Beyond its potential productivity effects SR is believed to induce visible changes in employment structures (Autor et al. 2003, Frey and Osborne 2013, Graetz and Michaels 2015). SR has a potential to change organization of processes in firms and everyday life of people by the diffusion of at least semi-autonomous physical systems out of industrial fabrication and into service economies. Using the advances of modern digital economy robotics can move from a professional use to a more private use. In order to understand SR one needs to identify its scope and detect it within various databases.

As a result of the arising multiplicity, the technology field so far is not clearly confined and thus neither part of any existing official industry, patent or trademark classification system nor of any concordances not to mention national account systems. Having said that, distinguishing SR from industrial robotics (hereafter Industrial Robotics (IR)) is hardly possible. This so far has impeded a comprehensive assessment of the economic impacts of SR diffusion, especially with respect to the magnitude, timing and geographical localization.

This work makes SR tractable by developing a search strategy to identify it within the patent databases. Moreover, we model the approach not to be limited to patents but to be applicable for scientific publications as well. In addition, the general methodology is not even confined to the field of robotics, but could be applied to any similar identification problem. Differentiating from classical lexical and citational approaches used by other scholars, our approach introduces a machine learning algorithm that is utilized as a classifier. Being trained on some sample data this classifier acts as an 'expert'. The machine is able to decide whether a patent belongs to the category of service robotics or not – with a certain degree of precision. Since there are several approaches in the scientific literature which deal with analogous problems of technology detection and classification, we hereby set out to (1) limit expert bias regarding vested interests on lexical query methods (with respect to term inclusion and exclusion), (2) avoid problems with citational approaches such as the lack of portability, and (3) facilitate evolutionary changes.

The following sections are organized as follows: First, we give an overview of previous technology identification approaches referring to examples of similar emerging fields that lacked classification schemes in its infant phase. Second, we present our step-by-step methodology for identifying developments in an emerging field characterized only by its early applications. It successively describes the use of patents as data source, the retrieval of a structured core dataset, and the use of an automated machine learning algorithm, namely a support vector machine (hereafter Support Vector Machine (SVM)). Finally, we present results of our pioneering approach and conclude with future scope for improvement.

3.2 Detection of emerging technologies

There is no widely agreed-upon definition of emerging technologies (Halaweh 2013). The initial lack of common knowledge, standards, and specifications entails uncertainties along various dimensions (Stahl 2011). Future costs and benefits, relevant actors, adoption behaviour, and potential socio-economic implications such as creative destruction are highly unclear (Srinivasan 2008). Therefore, scientific studies have been using bibliometrics to monitor trends for a variety of domains and assess the nature of emerging technologies already within scientific research and early development.

No matter what the paramount aim, all analyses greatly rely on well-founded data acquisition, which first and foremost identifies the technology under consideration. With ongoing technological advancements as well as computational power more and more elaborated strategies have accrued. Most often, technology detection within patent or publication databases is predicated on either (1) lexical, (2) citationist, or mixed search strategies.[3] For example, early conceptions of apt queries for nan-

[3] With respect to scientific publications another common strategy is to identify core journals.

otechnology proved to be difficult, as the first specific IPC-subclass B82B[4], which basically refers to nano-structures and their fabrication, was not introduced before the year 2000 and did not incorporate applications from former years (Noyons et al. 2003). In its infancy, it contained only estimated 10 percent of all relevant documents. Hence, the first scientific identification approach for nanoscience and technology relied instead on a lexical query developed in 2001 by the Fraunhofer Institute for Systems and Innovation Research (Fraunhofer Institute for Systems and Innovation Research (ISI)) in Germany and the Centre for Science and Technology Studies (Centre for Science and Technology Studies in Leiden (CWTS)) at Leiden University in the Netherlands.

A lexical query is a search for specified terms, which in the most simple case might consist of only one word (like 'nano*' for nanotechnologies) or a basic combination (like 'service robot*'). This primal string is applied to titles, abstracts, keywords or even the whole text body of examined documents. Some of these documents might prove to be relevant in the eyes of experts and, thus, offer additional terms starting an iterative process.[5] Considering emerging fields, the number of terms within a search string that is developed in such a lexical manner naturally grows rapidly. More and more scholars and practitioners become attracted by the field [6] adding alternatives and broadening interpretations in the course of time. For example, in order to keep track of the dynamically spreading nano-fields Porter et al. (2008) comprised a modular Boolean keyword search strategy with multiple-step inclusion and exclusion processes, which was subsequently enhanced and evolutionary revised (Arora et al. 2013). In addition, both authors of scientific publications as well as applicants of patents are interested in some rephrasing. The former, because they might benefit from a serendipity effect if their label establishes itself in the scientific

All articles within those journals are then considered relevant. For patents though, this search strategy is obviously not feasible, which is why we do not deepen it further.

[4] Only in 2011 a second sub-class, B82Y, focusing on specific uses or applications of nanostructures was introduced for IPC and the Cooperative Patent Classification (Cooperative Patent Classification (CPC)). Previously, related nano patent documents could only be identified if they were classified via the European Classification System (European Classification System (ECLA)) with the specific sub-class Y01N.

[5] Such a search strategy is called evolutionary, if subsequent researchers may build upon existing query structures by progressively incorporating terms that better specify the technology and widen its scope (Mogoutov and Kahane 2007).

[6] For the instance of nanotechnology, to which we refer throughout, (Arora et al. 2014) measure the growth in nano-prefixed terms in scholarly publications and find that the percentage of articles using a nano-prefixed term has increased from less than 10% in the early 1990s to almost 80% by 2010.

community, and the latter because of encryption and legalese issues. Applicants may want to re-label critical terms, both to hide relevant documents and technical information from actual rivals and to build patent thickets of overlapping Intellectual Property Rights (IPR) which precludes potential competitors from commercializing new technology altogether.

A lexical query can be enriched adding documents and inherent terms by citational approaches, for instance, by including new publications, that are cited by at least two authors belonging to the initial database (Garfield 1967, Bassecoulard et al. 2007)[7] or, regarding patents, by including applications, that refer prior art that has been a part of the previously established core. In the example of nanotechnology Mogoutov and Kahane (2007) enriched an initial nanostring by a number of subfields, automatically identified and defined through the journal inter-citation network density displayed in the initial core dataset of nano-documents. Relevant keywords linked to each subfield were then tested for their specificity and relevance before being sequentially incorporated to build a final query.

The example of nanotechnology illustrates well how much effort the development of an evolutionary query yields. Lately, private interests – rather than governmental or scientific research – have driven even more elaborated technology identification procedures: companies that seek to monitor competitors or investigate latest research trends have started to rely on more cost-efficient processes in order to lower resulting expenditures. As a side effect, some encompassing literature on specialized text mining techniques has emerged, which goes beyond lexical and citation based procedures. To name just a few, Li et al. (2009) attempt to find significant rare keywords, considering heterogeneous terms used by assignees, attorneys, and inventors. Yoon and Park (2004) argue that citation analysis has some crucial drawbacks and propose a network-based analysis as alternative method, that groups patents according to their keyword distances. Lee (2008) uses co-word analyses regarding term association strength and provides indicators and visualization methods to measure the latest research trends. Lee et al. (2009) transform patent documents into the structured data to identify keyword vectors, which they boil down to principal components for a low-dimensional mapping. These facilitate the identification of areas with low patent density, which are interpreted as vacancies and, thus, chances

[7] This approach naturally harbors the risk of including generic articles of any scientific field that somehow happen to be cited in a technologically unrelated context. Bassecoulard et al. (2007), therefore, incorporate a statistical relevance limit relying on the specificity of citations.

for further technical exploitation. Erdi et al. (2013) use methods of citation and social network analysis, cluster generation, and trend analysis. Tseng et al. (2007) attempt to develop a holistic process for creating final patent maps for topic analyses and other tasks such as patent classification, organization, knowledge sharing and prior art searches. They describe a series of techniques, including text segmentation, summary extraction, feature selection, term association, cluster generation, topic identification and information mapping. For the field of robotics, Ruffaldi et al. (2010) is a good instance: They visualize trends in the domains of rehabilitation and surgical robotics identified via text mining. Following Mogoutov and Kahane (2007), the relative performance of different identification approaches may be compared via (a) the respective degree of intervention of experts, (b) their portability, (c) their transparency regarding core features and respective impacts on final results, (d) their replicability, (e) their adaptability, meaning its ability to produce valid results while the technology in question keeps evolving, (f) their updating capacity, and (g) the extent and relevance of the data obtained. Certainly, no single best approach exists, since any method has its advantages and drawbacks according to these criteria. We will conclude on the relative performance of our approach at the end of this chapter.

Instead of purely lexical, purely citationist or mixed query, that are similar to the current text mining literature, we propose a machine learning algorithm. For this purpose, we first identify a small core patent dataset consisting of 228 patent applications and then let automated algorithms identify emerging technology boarders.

3.3 Machine learning for technology detection

Patents as Data Source

As soon as a technology is sufficiently well specified, generically distinguishable, and ideally properly classified there are various techniques to map ongoing advancements. However, if such a delineation is not yet established and no broadly accepted consensus has been reached so far, economists most often rely on lexical, citation based, or mixed search strategies for prior identification purposes that help to trace related emerging fundamental and application knowledge in academic articles and patent documents.[8] As regards the technology under consideration, it is important to acknowledge that according to the IFR, the intended use, and as a consequence, the factual field of application determines the delimitation of SR from IR. Thus, patents are the data source of choice for an automated SR identification, since patentability requires an indication of the intended commercial implementation. Despite all difficulties that arise in the use of patents and their interpretation, they are widely accepted as indicator for innovative activity Griliches 1990, Hall et al. 2005. Especially citation structures facilitate tracing knowledge flows (see, for instance, Jaffe et al. 1993, Thompson 2006, Fischer et al. 2009, Bresnahan 2010) and thus make technology development patterns visible. Hence, we started with a patent search strategy with a vision to extrapolate it to other lexical sources.

Building a structured core dataset that is suitable for the later application in machine learning, requires the identification of a sufficiently large number of documents, that are validated as part of the technology and capture most of its hitherto variety of developments. This validation is granted by independent technological experts, who can either provide those documents themselves or may be given a predefined assortment to adjudicate on. The latter decreases a potential expert bias with respect to multifaceted preferences but might give rise to a negative influence of the researcher himself, who has to develop a search method for this primal assortment. In this work experts are provided with a predefined core dataset.

Retrieval of a core service robotics patent dataset

All unstructured patent text data as well as related document meta data were extracted from the 'EPO Worldwide Patent Statistical Database' (PATSTAT), version April 2013.[9] First, we extracted all patents that were either sorted in IPC class B25J[10] or contained a substring like 'robot*' in their respective title or abstract.[11] Hence, we established a set of documents describing robotic devices. Second, in

[8] Consequently, the adequate data sources for this identification process are the same that comprise the targets of subsequent analyses which might give cause for some criticism.

[9] This database encompasses raw data about 60 million patent applications and 30 million granted patents, utility models, Patent Cooperation Treaty (PCT) applications, etc. filed at more than 100 patent authorities worldwide.

[10] MANIPULATORS; CHAMBERS PROVIDED WITH MANIPULATION DEVICES. See http://www.wipo.int/classifications/ipc/en/

[11] According to the USPTO, most of the manipulators classified in B25J are industrial robots. See http://www.uspto.gov/web/patents/classification/cpc/html/defB25J.html.

order to identify a subset of potential SR patent documents that comprise most of the hitherto existing developments we created 11 sub-queries, based mainly upon IFR application fields for service robots. These queries consisted both of IPC subclasses (mostly on 4-digit-level) and stemmed lexical terms, combined modularly in a Boolean structure.[12]

The second step provided us with 11 non-disjunct subsamples of potential SR patents. While other approaches regarding similar tasks of technology identification from there on further evaluate candidate terms by testing, assessing and adjusting terms and class codes to address weaknesses and follow emerging research trails manually (Porter et al. 2008), we did not alter the primal modular Boolean search. Instead, as indicated above, we left it to technological experts to verify the underlying categorization. Two independent academic expert groups with 15 scientists, affiliated with the

- High Performance Humanoid Technologies (H2T) from the Institute for Anthropomatics and Robotics at KIT, Germany, and the
- Delft Center for Systems and Control / Robotics Institute at TU Delft, Netherlands,

took on the task to decide which of the patents belonged to SR and which belonged complementarily to IR. The above experts were specialized in humanoid robotics, computer science, and mechanical engineering. Their experience in the field of robotics varied between 1 and 15 years. We provided them with 228 full body versions of potential SR patents from all over the world, extracted with the primal subsample queries. All patents listed in PATSTAT disclose at least English titles and abstracts. Thus, the judging scientists could always refer to these text parts as well as to all engineering drawings, independent from the language of the remaining text body.

For the application of automated machine learning approaches we then transformed the unstructured patent document text into structured data. This included several steps, namely (1) combining titles and abstracts in one body and splitting the resulting strings into single terms in normal lower cases, (2) removing stop words, (3) stemming, i.e. reducing inflected words to their stem, (4) constructing n-grams of

[12] The queries are available upon request.

term combinations (up to 3 words in one), and (5) deriving normalized word and n-gram frequencies for each document.[13]

With these normalized frequencies a matrix was constructed with columns, being variables, and rows, being their observations. This matrix, shown in table 3.1, together with the binary vector indicating which observations had been identified as SR patents, served as a training input for the machine learning algorithm.

Table 3.1: Structure of patent word and n-gram frequency matrix with binary decisions as input for machine learning. The lighter gray shaded area indicates an example of a subsample, on which the machine is trained. The darker gray area is then a respective example for a subset of data which is used for testing the fitness of the classification process. The non-shaded area at the bottom refers to new data, on which the SVM is able to decide based on the previous training[14].

	Attribute vectors $\mathbf{x}$									binary
patent	word$_{w1}$	word$_{w2}$	...	bigram$_{b1}$	bigram$_{b2}$	...	trigram$_{t1}$	trigram$_{t2}$	...	decision y
1	freq$_{\cdot 1\|w1}$	freq$_{\cdot 1\|w2}$	...							1
2	freq$_{\cdot 2\|w1}$	...								-1
...	...									...
205	freq$_{\cdot 205\|w1}$	...			freq$_{\cdot 205\|b2}$	...				-1
206	freq$_{\cdot 206\|w1}$	...			...					1
...	...									...
228	freq$_{\cdot 228\|w1}$	...						freq$_{\cdot 228\|t2}$	...	-1
xxx	freq$_{\cdot xxx\|w1}$	...						...		?
xxx	freq$_{\cdot xxx\|w1}$	...								?
...	...									...

Statistical classification, using machine learning algorithms, has long been implemented for the purpose of solving various problems and tasks, such as computer vision, drug discovery or handwrite and speech recognition. Numerous different methods were developed and new ones still appear. However, there has been no one, at least to our knowledge, using statistical classifiers on the basis of a primal lexical query for the purpose of detecting an emerging technology. We considered a number of alternatives (Kotsiantis 2007) to the aforementioned SVM, such as k-nearest neighbor, neural networks, and genetic algorithms before starting with our particular algorithm. According to the so called *no-free-lunch* theorem (Wolpert and

[13] We also tried to incorporate another step (6), which added IPC dummy variables to indicate class belongings. These additional attributes where later abandoned by the following feature selection process, which suggests that these IPC class belongings are not significant for the categorization at hand.

[14] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

Macready 1997), there is no general superior machine learning method and every problem has to be tackled individually depending on its properties. We assessed the aforementioned algorithms according to run-time performance, sensitivity to irrelevant or redundant features, and ability to overcome local maximums. In a nutshell, SVM proved to be the most suitable algorithm and this decision was in line with computer science experts' opinions from robotics groups at the Karlsruhe Institute of Technology.

The k-Nearest neighbor classifier was not chosen due to a poor run-time performance, its sensitivity to irrelevant or redundant features and weaknesses compared to the SVM, regarding difficult classification tasks (Cunningham and Delany 2007). Although, the first disadvantage of the algorithm is not that important for our problem, the next two are highly relevant. We do not know in advance, which keywords or features will be significant within the identification process of SR patents and it is also hard to assess the difficulty of our task up front. Despite the popularity of deep learning algorithms, the second abandoned algorithm was a neural networks classifier (Rojas 1996), which is difficult to retrain and hard to extend. This is important for our problem since we would like to diversify and expand our sample size as well as the expert pool if this proves to be advisable. On top of that, this algorithm may get stuck on local maximums and requires quite large datasets to be trained. Finally, we rejected genetic algorithms which give no assurance of an optimum solution in terms of a best fit function (Rojas 1996). Nevertheless, this algorithm is probably the best substitution for a SVM and an implementation of it could offer some future improvement of our methodology.

Support Vector Classification

The method of support vectors was introduced in the middle of the 1960s (Guyon et al. 1993, Cortes and Vapnik 1995). The original approach together with its various extensions is now one of the most acknowledged and recommended tools among modern machine learning algorithms. In the following we briefly describe its core concept and discuss some advantages that are found relevant for the problem at hand. The core idea of the method is to create a unique discrimination profile (represented by a linear function) between samples from different classes.

The result is a line – or more generally a hyperplane – which is constructed in such a way, that the distance between two parallel hyperplanes touching nearest samples becomes as large as possible. This way the method is trying to minimize false classification decisions. The "touching" data points are termed *support vectors*. In fact, the resulting separation plane is shaped only by these constraining (= supporting) points. Below we provide the mathematical notation of a support vector machine, following Hsu et al. (2010).

Formally defined, we have a training set (x_i, y_i) of $i = 1, \ldots, l$ sample points, where every $\mathbf{x}_i \in \mathbb{R}^n$ is an attribute vector (consisting of our normalized word and n-gram frequencies) and $y_i \in \{-1, 1\}^l$ is a decision for that specific data point, which, thus, defines its class. Each point represents a patent. The SVM then yields the solution to the following optimization problem (Boser et al. 1992, Guyon et al. 1993):

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ s.t. \quad & y_i\left(\mathbf{w}^T\Phi(\mathbf{x}_i) + b\right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned} \tag{3.1}$$

where w is the normal vector between the separating hyperplane and the parallel planes spanned by the support vectors. The mapping Φ is related to so called Kernel functions, so that $K(\mathbf{x}_i, \mathbf{x}_j) \equiv \Phi(\mathbf{x}_i)^T\Phi(\mathbf{x}_j)$. For problems in which the data under consideration are not linearly separable, Φ maps the training attributes into a higher dimensional space, where a hyperplane may be found. Table 3.2 summarizes common Kernel functions and their respective parameters γ, r, and d (Burges 1998, Ali and Smith-Miles 2006, Pedregosa et al. 2011, Manning et al. 2008)[15].

Table 3.2: Kernel functions used for the SVM[16]

Kernel function	Formula
Polynomial	$(\gamma\langle x, x'\rangle + r)^d$
Radial basis function (rbf)	$\exp(-\gamma\lvert x - x'\rvert^2)$
Sigmoid	$\tanh(\langle x, x'\rangle + r)$

[15] Since there is no possibility to determine in advance which Kernel function should be used, the choice of the depicted functions was mostly motivated by their popularity in classifiers and availability within the software package used.

[16] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy

The above version of the classification procedure also incorporates the so called *Soft-Margin* method (Cortes and Vapnik 1995), that allows for mislabeled training sample points. The approach introduces ξ_i as non-negative slack variables which measure the extent of incorrectly classified items in the training set. $\sum_{i=1}^{l} \xi_i$ is, thus, a penalty term, and C a penalty parameter. The chosen method provides certain advantages. It is generally accurate, does not get stuck on a local maximums and is tolerant to irrelevant or redundant attributes (Kotsiantis 2007). The latter is probably the most important for the purpose of patent detection, since it is impossible to know in advance which keywords or keyword combinations will be relevant for identification.

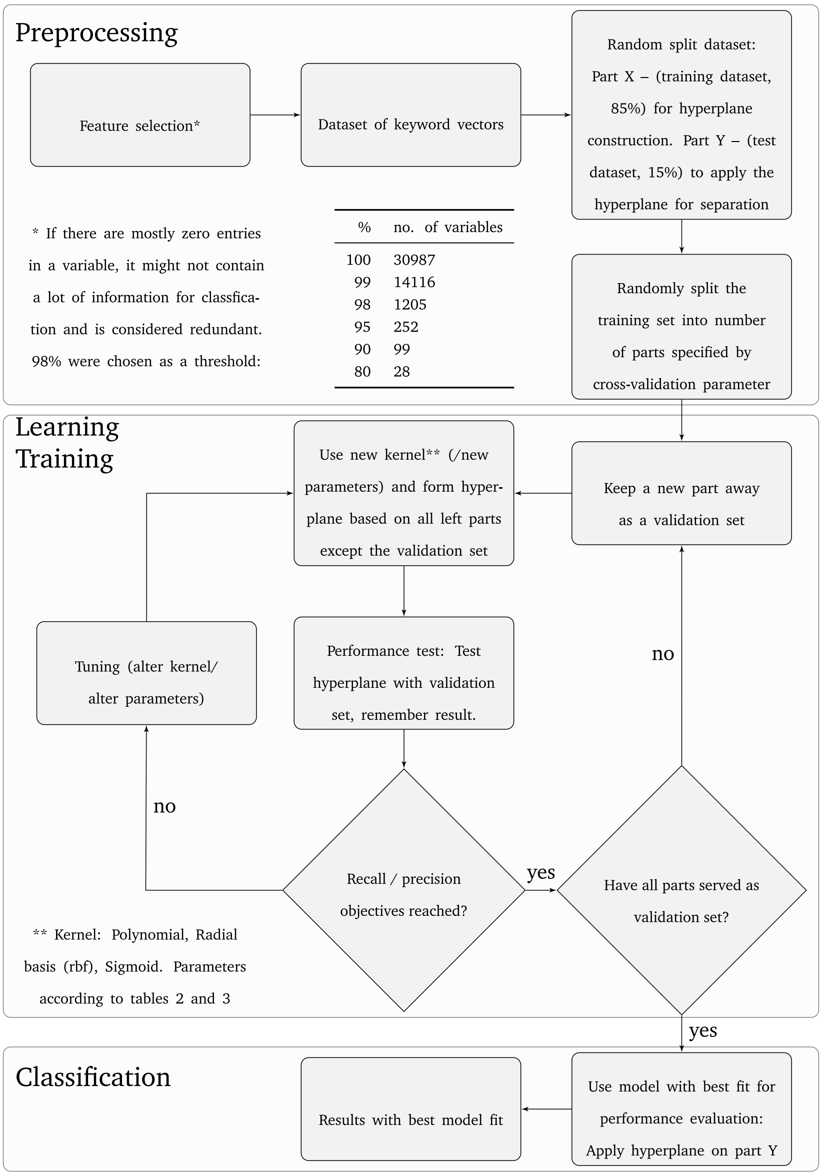
Training Algorithm, Classification, and Evaluation

Figure 3.2 depicts the flow chart of our algorithm. First, we preprocessed the data in order to eliminate irrelevant features and to obtain a final dataset of feature vectors. When we turn to the result section, the necessity of this preprocessing becomes clearer. In a second step we started the SVM training process, comprised of three iterative steps: training of the model, model evaluation and optimization. We realized all these steps for our SVM, using the python programming language and its library *scikit*-learn for machine learning (Pedregosa et al. 2011).[17] Finally, the classifier with the best model fit was applied on PATSTAT data to identify new service robotics patents.

Firstly, the algorithm randomly splits the training dataset X into training and test parts. Second, it fits the model based on the training dataset leaving out the test data. During the training process the data are again split into k parts. The algorithm then trains the model on k-1 parts and validates on the k-th part. The training is performed several times so that every part serves as a validation dataset. The number of training repetitions is reflected by a cross-validation parameter and can be specified and is subject to variation during the fitting of the model.

based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

[17] We do not discuss the exact implementation of the support vector machine algorithm using the python scikit-learn library. All necessary materials can be found in open access libraries following the reference provided (Pérez and Granger 2007).


Preprocessing
Feature selection*
Dataset of keyword vectors
Random split dataset: Part X – (training dataset, 85%) for hyperplane construction. Part Y – (test dataset, 15%) to apply the hyperplane for separation
Randomly split the training set into number of parts specified by cross-validation parameter
Learning Training
Keep a new part away as a validation set
Use new kernel** (/new parameters) and form hyperplane based on all left parts except the validation set
Performance test: Test hyperplane with validation set, remember result.
Tuning (alter kernel/ alter parameters)
Recall / precision objectives reached?
no
yes
Have all parts served as validation set?
no
yes
Classification
Use model with best fit for performance evaluation: Apply hyperplane on part Y
Results with best model fit

* If there are mostly zero entries in a variable, it might not contain a lot of information for classfication and is considered redundant. 98% were chosen as a threshold:

%	no. of variables
100	30987
99	14116
98	1205
95	252
90	99
80	28

** Kernel: Polynomial, Radial basis (rbf), Sigmoid. Parameters according to tables 2 and 3

Figure 3.2: Flow chart of the machine learning algorithm with preprocessing, support vector training, and final classification.[18]

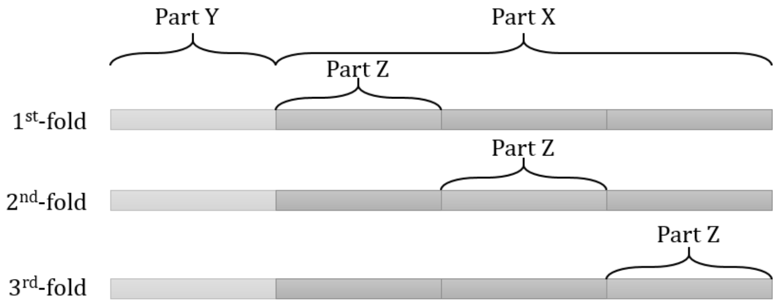


Figure 3.3: k-fold cross-validation process. Part Y serves as an independent test set, while part Z serves as a cross validation set.[19]

Figure 3.3 illustrates the k-fold cross-validation process. The evaluation of our model is based on the criteria of precision and recall. The former measures the ability of a classifier not to label objects as positive that should have been labeled negative. Formally, precision is the total number of true positives (tp) divided by the sum of all positives including false positive errors (fp).

$$precision = \frac{tp}{tp + fp} \tag{3.2}$$

The latter (recall) measures the ability of a classifier to find all positives or the number of true positives divided by a sum of true positives and false negative errors (fn).

$$recall = \frac{tp}{tp + fn} \tag{3.3}$$

On the one hand, a model with a good recall but bad precision will find all positive samples – but will have some of them being actually negative. On the other hand, a model with bad recall but high precision will not have false positive objects, however it will miss some of the true positives. In order to balance these two measures we used a f1-score that can be seen as their weighted average:

[18] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

[19] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

$$f1 = 2 \cdot \frac{precision * recall}{precision + recall} \tag{3.4}$$

To optimize our classifier we calibrated it to have the highest possible f1-score. Tuning of the model was done by varying the cross-validation parameter, the kernel functions, and their respective parameters.

3.4 Resulting learned model of patent classification

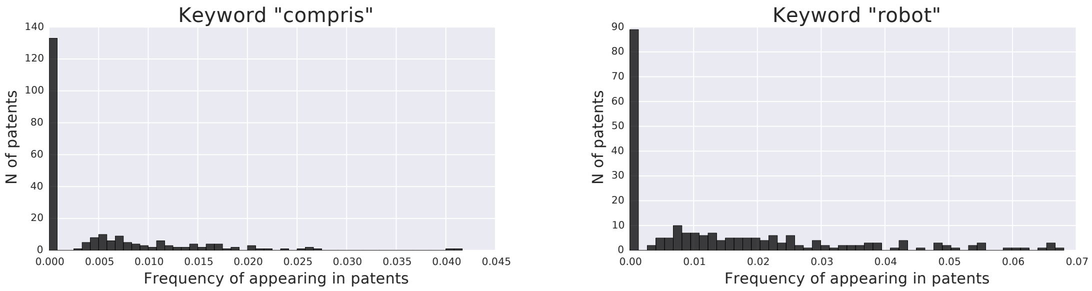
The sample used in the machine learning process consisted of 228 patents with valid expert decisions. It contained 98 SR patents and 130 IR patents, according to our expert group's validation. As a result of the transformation of unstructured patent text into structured data, we observed 30,987 different features (or variables) within these patents, which included keywords, bigrams, and trigrams.[20].

The resulting matrix (228 x 30,987) had to be pre-processed before serving as an input for the SVM, due to its sparsity. This means, that only a small number of keywords and n-grams are shared by a majority of the patents. At first glance this information could appear confusing. The explanation lies in the variety of SR applications: Descriptions of significantly different service robots with very unlike applications contain a huge number of dissimilar keywords and keyword combinations. Most of these are uniquely used in their specific contexts and, thus, appear with a very low frequency. Figure 3.5 illustrates this fact by showing typical relative appearances of normalized frequencies of two randomly chosen variables.

Thus, some variables contained too little information and introduced noise instead. Consequently, these insignificant features had to be excluded from the data, since they do not contain information relevant for classification purposes. For example, if a keyword (or n-gram) appeared in only one patent, this variable would not have helped in solving the problem of classification. Our feature selection process served

[20] We even included IPC classes in an early stage of development, but did not find any of these classifications to become part of the support vectors. They turned out to irrelevant for the discrimination procedure and were thus removed during the feature selection process

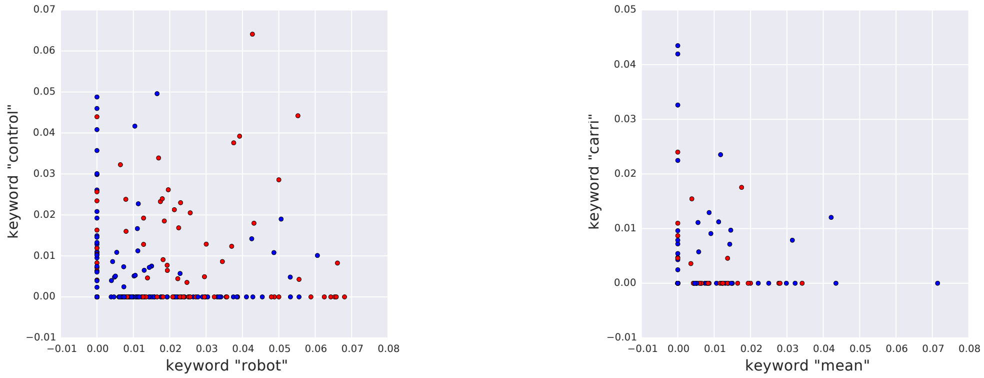
[21] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017



(a) Key word 'compris'

(b) Key word 'robot'

Figure 3.4: Two histograms of exemplary chosen keyword.s[21]

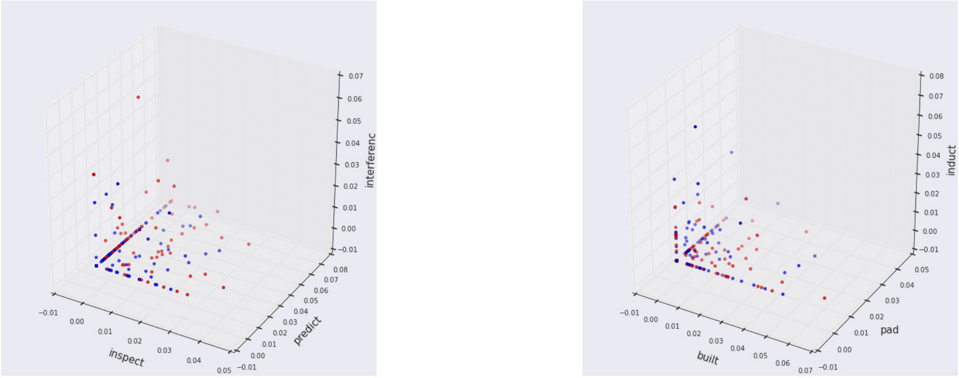


(a) Key words 'robot' and 'control'

(b) Key words 'mean' and 'carri'

Figure 3.5: Normalized frequencies of randomly chosen attribute pairs – here keywords. Colored dots indicate the expert classification as SR (red) and IR (blue).

to exclude such a redundant feature. We implemented a threshold that at least 2% of the entries of a variable in each class (SR vs. IR) should have non-zero entries. The table in the flow chart (figure 3.2) shows the dependency between the number of variables and different thresholds. With this selection process the resulting matrix was reduced to 1206 variables for our 228 observations/patents. We provide these variables/terms in the tables B.5 to B.11 in the appendix. Finally all variable frequencies were scaled to the interval [0, 1], so that a second normalization process set the maximum frequency in the sample to 1. Figures 3.5 and 3.6 show normalized frequencies of attribute pairs and groups of three, respectively. Colored dots indicate the expert classification as SR (red) and IR (blue).



(a) .

(b) .

Figure 3.6: Normalized frequencies of randomly chosen attribute groups of three – here keywords. Colored dots indicate the expert classification as SR (red) and IR (blue).

SVM specific outcomes

In order to eliminate negative influence of the unbalanced dataset we introduced weights in our SVM proportionate to SR and IR classes. Following the cross-validation procedure the support vector machine was fit on to a 85% of the original dataset. The remaining 15% were kept for testing purposes. The split was random and its ratio is an arbitrary choice.

Table 3.3: Model tuning parameters and respective values[22]

Parameter	Varied values	Chosen values
cross-validation (cv)	$3, 4$	3
complexity (C)	$10, \ldots, 1000$	10
γ of rbf kernel	$10^{-6}, \ldots, 10^{-2}$	0.005
γ of polynomial kernel	$10^{-6}, \ldots, 10^{-2}$	not chosen
d of polynomial kernel	$1, 2, 3$	not chosen
r of polynomial kernel	$1, 2, 3$	not chosen
γ of sigmoid kernel	$10^{-6}, \ldots, 10^{-2}$	not chosen
r of sigmoid kernel	$1, 2, 3$	not chosen

The cross-validation parameter was set to 3 and 4, determining the amount of random splits of training dataset into a training and evaluation sets. Another param-

[22] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

eter, that was varied while searching for a better model, is so called C parameter. The following citation nicely explains the main properties of this penalty parameter: *"In the support-vector network algorithm one can control the trade-off between complexity of decision rule and frequency of error by changing the parameter C"* (Cortes and Vapnik 1995, p. 286).

Table 3.4: Classification report[23]

	precision	*recall*	*f1-score*	*No. of patents in test set*
SR	75%	94%	83%	16
IR	93%	74%	82%	19
Avg. / total	85%	83%	83%	35

Finally, the three different kernel functions from Table 3.2 were considered. In particular, the first was a polynomial function and its γ, degree, and r coefficient. The second was a radial basis function (rbf) and its γ constant. The third was a sigmoid function and its γ and r constant. Table 3.3 presents all kernel parameters and their values, that were considered to find the best performing classifier – as well as all eventually chosen values.

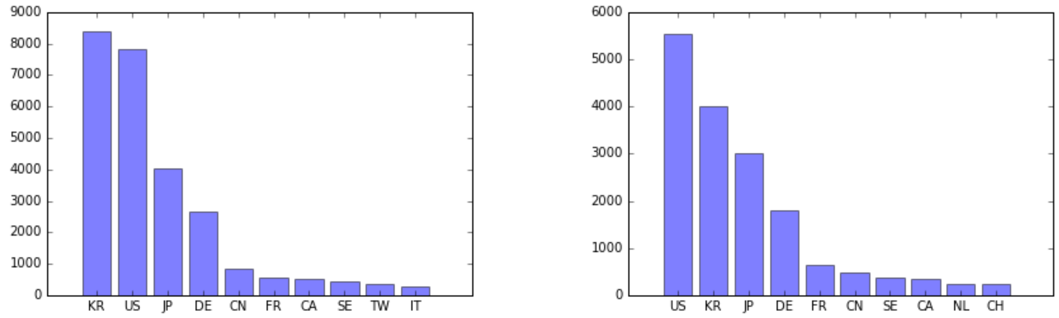
The best f1-score of the model was obtained after a grid-search, performing simulations with all possible combinations of the above mentioned parameters. The final model showed an 85% precision and 83% recall. It contained a radial basis function kernel with γ equal to 0.005 and C equal to 10. The training set was randomly split into 3 equal parts for cross validation. The resulting discrimination plane between the two classes of patents was constructed using 192 support vectors, meaning that only these sample observations were significant for classification. Table 3.4 presents a classification report after classifying the test set of our sample.

Service Robotics Patents: Overview

Below we provide some descriptive statistics of patents identified based on our methodology without elaborating on them since this is out of scope of this research. We identified 21286 priority patents in the period 1980 - 2010 in the world data.

[23] Reprinted with permission from Kreuchauff F., Korzinov V. (2017) A patent search strategy based on machine learning for the emergent field of service robotics. Scientometrics, Vol. 111, Issue 2, 743-772. Copyright © Akadémiai Kiadó, Budapest, Hungary 2017

Figure 3.7 shows that majority of the inventors are coming from the Republic of Korea and United States of America followed by Japan and Germany. Similar picture is obtained when considering applicants (Figure 3.7) or in other words companies, that patented a particular invention.



(a) Inventors

(b) Applicants

Figure 3.7: The number of inventors and applicants in service robotics in top 10 countries. KR=South Korea, US=United States of America, JP=Japan, DE=Germany, CN=China, FR=France, CA=Canada, SE=Sweden, TW=Taiwan, IT=Italy.

Figure 3.8 demonstrates a geographical distribution of German applicants in Service robotics depicting their postcodes on the map. We can see, that knowledge production activity is concentrated in the south-west part of the country, which is known to be economically developed. In the Chapter 4, section 4.2 we assess to what extend these clustering activity differs from an overall pattern of knowledge production activity in Germany and show that there is a significant deviation from a random pattern.

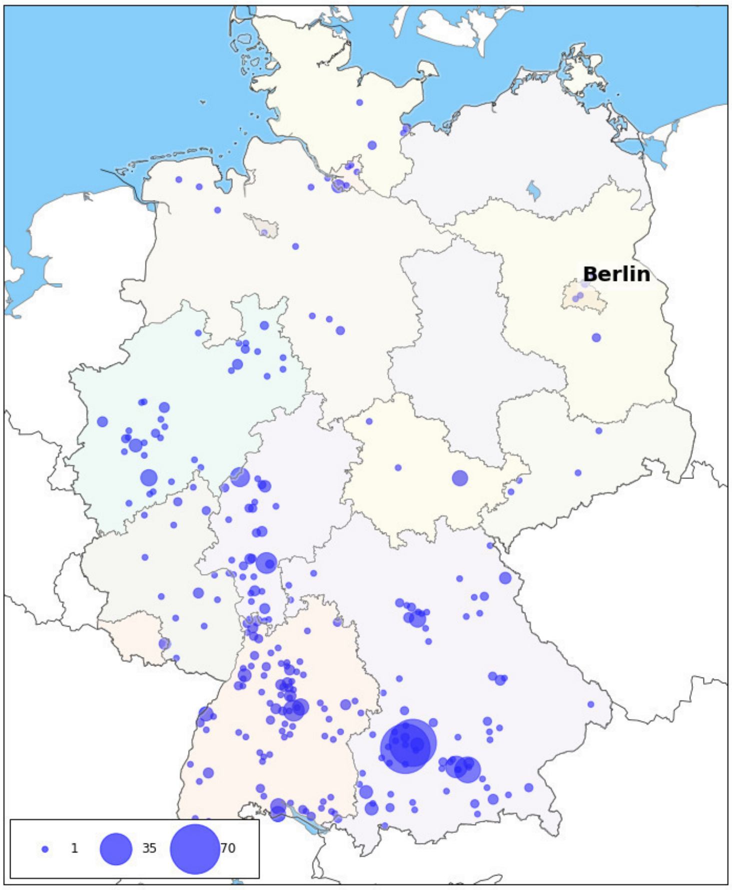


Figure 3.8: Location of firms patenting in service robotics in Germany. The size of the circle corresponds to the number of firms.

3.5 Conclusion on patent search strategy

In this Chapter we proposed a novel methodology for detecting early developments of an emerging technology in patent data. Our method uses a support vector machine algorithm on the example of robotics patents. The resulting model was able to find 83% of service robotics patents and classify them correctly with a probability of 85%.

There are several advantages of our method regarding technology classification tasks, which we discuss along the criteria of Mogoutov and Kahane (2007). Firstly, experts do not choose, which terms should be added to or excluded from the primal search, hence, the typical lexical bias towards preferred subfields is limited. Speaking of lexical versus citationist approaches, our method, also avoids a major drawback of citational methods, which circle around a core dataset and rely on future works explicitly referring to this prior art. Since citations in patents are generally rare[24], for young emerging technologies in particular the citation lag decreases the expected number of citations for any given document to a negligible amount. Secondly, the procedure offers strong portability, so that it can easily be applied to scientific publications. Moreover, our step-by-step classification method can basically be applied to any emerging technology – not only those, that arise as an initially small subset consisting of niche applications like SR emerging out of robotics. Nanotechnology would have been hard to detach from some well-defined mother technology. The same is true for Industry 4.0, which is a superordinate concept describing digitally cross-linked production systems and, thus, enveloping various heterogeneous subtechnologies, that are hardly classifiable. One of our future tasks will thus comprise the application of our method on historical nanotechnological patent sets as well as on Industry 4.0 technologies in order to demonstrate the general applicability and robustness of our method. Thirdly, our algorithm approach shows high adaptability. Due to its learning nature it is able to produce valid outcomes although the technology under consideration is constantly evolving. Fourth and of capital importance, the proposed method performs well in terms of recall and precision scores, proving sufficient extent and relevance of the obtained data.

There is some scope for an even more precise technology identification. First, there is still room to increase the performance of the SVM method, namely regarding the kernel functions. Although there have not been any successful attempts to introduce automatic kernel selection algorithms yet (Ali and Smith-Miles 2006), it is probably possible to find a better kernel function for our problem at hand. Second, the support vector machine can be seen as a first-tier machine classifier that we just started with. Other methods like genetic algorithms, neural networks or boosting as well as their combinations could be applied in additional steps. Finally, applying Principal Component Analysis (PCA) to our matrix of variables could provide some insights about a similar behavior of different key words in patents. This, however, comes at a costs of interpretability of results. Nevertheless, words could be grouped and analyzed together, to see whether these groups of variables are significant in

[24] Within PATSTAT, for instance, more than 90% of the listed patent applications are followed by less than three forward citations, 74% do not show any at all.

identifying an emerging technology, which is a subject for further research.

Chapter 4

Patterns in innovation clustering and job creation

4.1 Introduction

'God does not play dice ...'
A. Einstein

The methodological research of this thesis, started in Chapter 3 with the development of a technique for the detection of an emerging technology, continues in chapter Chapter 4. In the following a broader perspective is taken on the concept of pattern detection in economics. In particular, the problem of technology detection is a part of a broader set of problems in identifying micro trends and patterns, that emerge on a macro level. Macro- and microeconomics, unfortunately, largely remain separated from each other, with one being concerned with aggregate economy behavior and the other being focused on an single markets and people behavioral patterns. One of the primary reasons for that is a complexity of the smallest unit of study, which, in social systems, is often a person who's behavior we still can't fully describe and predict. Modern technological advancements provide us with more and more micro data as well as tools to proxy and simulate the real world social systems. The discussion in the following chapter continues methodological research of this thesis in the direction of pattern detection in micro data. What is meant by pattern detection is a statistically significant deviation from a benchmark generated with a process driven by chance.

In economics Ellison and Glaeser (1997) apply those techniques to construct an index of industrial concentration. The so called "dartboard approach" compares locations of actual industry establishments to average location of hypothetical establishments allocated by a random process within a given administrative region. This idea was further developed by Duranton and Overman (2005) were space is treated homogeneously and the frequency of bilateral distances between all establishments is benchmarked against a random frequency. Various similar works apply this methodology to a different datasets as well as improve it (Duranton and Overman 2008, Albert et al. 2012, Nakajima et al. 2012, Barlet et al. 2013, Koh and Riedel 2014). To the best of our knowledge, the only two studies by Buzard and Carlino (2009) and Carlino et al. (2012) apply a similar approach to analyze innovations and emerging technologies which is a primary concern of this thesis.

Section 4.2 addresses this research gap by taking a geographical perspective on technological change and identifying clusters of innovations, that can be significantly differentiated from the ones expected by a random allocation. It is known that general economic activity tends to be geographically concentrated, and innovation-related activities are even more spatially clustered (Audretsch and Feldman 1996). This work demonstrates how some observed clusters of industrial innovations significantly deviate from an overall spatial distribution in the case of Germany, opening a new perspective on innovation clustering and contributing to the literature on spatial organization of innovative activities. A high clustering of service robotics innovators is also depicted, which indicates the presence of agglomeration forces in this field.

The second section of this chapter concerns with another application of the benchmarking idea in the economics context. Namely, it focuses on the topic of a sustained superior performance of a firm. The work of Henderson et al. (2012) draws first attention to this interesting phenomena. In order to illustrate the problem arising in this topic, let us take a broader perspective on pattern perception.

People tend to be misled by chance, while looking for a meaningful patterns. For instance, there is a bizzare but an illustrative example. The letters in the name of William Shakespeare can be rearranged in a sentence 'Here was I, like a Psalm'.[1] The 46th word from the top of the Psalm 46 in the King James Bible is 'shake'

[1] Psalm 46, Wikipedia (2018, July 1st.) Retrieved from https://en.wikipedia.org/wiki/Psalm_46

and 46th word from the bottom is 'spear' (Psalm 46, Bible (1999)). Sir William Shakespeare was 46th years old, when the first version of this Bible was completed. This is a completely random fact. Actually, given the vast amount of information and enough patience for a search, one can find a numerous examples of such strange 'facts'. As noted by Tversky and Kahneman (1971), it is easy, thus, to be misled by this randomness perceiving patterns were they do not exist. Another examples of randomness misconception are the 'hot hand fallacy' and the 'gambler fallacy'. Many fans, coaches and even sports commentators will be positive about the statement, that, if a basketball player hits five shots in a row, he is more likely to hit another one. This is wrong. This phenomena is known as a 'hot hand fallacy' and is studied since the work of Gilovich et al. (1985). It is shown that "... people not only perceive random sequences as positively correlated, they also perceive negatively correlated sequences as random." (Gilovich et al. 1985, p. 311) This mismatch between beliefs and facts is partly explained by the law of small numbers coined in the paper of Tversky and Kahneman (1971). If we now replace a player throwing a ball with a tossing coin or turning roulette, many people tend to say that the opposite is likely, which is widely known as a 'gambler fallacy' (Roney and Trick 2009). Both phenomena are well studied and deal with people's perceptions of a chance.

What if one now asks a question of whether a firm, that has been growing in terms of employment higher than others in the economy for a number of years, will continue to do so? The famous examples are companies like Google or Amazon. This question is very relevant, given that unemployment and job creation are at the core focus of the political agenda and high growth firms are at the center of a European policy debate (EU 2013). The literature provides a scant evidence regarding this phenomena, showing that it is difficult to identify determinants of sustained job creation. It is observed that firms do fail often (Parker et al. 2010) and superior growth is typically a temporary phenomenon in the life of a firm (Hölzl 2014). The persistence of growth is smaller for small firms and larger for large firms (Acs and Mueller 2008). Even more profounding is the evidence that persistent job creators exist, but do not differ from other (non persistent) high growth companies (Capasso et al. 2014, Bianchini et al. 2016).

The research in Section 4.3 answers a question of whether the number of persistent high growth firms can be explained by a simple random process. A method based on Markov property is developed allowing to abstract from distributional assumptions. We find a mixed evidence of presence and absence of factors determining sustained superior growth performance. In some countries firm dynamics can't be explained with a model driven by chance, indicating that firms might possess superior operating capabilities and/or technological traits as well as better managerial and organizational strategies. In contrast, the data from other countries (Italy and Spain, for example) indicate that we cannot rule out chance as an explanatory mechanism for the firm's growth trends, pointing out that it could be a merely temporary phenomenon, implying that firms create new jobs, but very likely these jobs will be lost. Finally, we also find a contradictory evidence where depending on the underlying model, growth measure and confidence level, the results may vary.

The chapter continues as follows, Section 4.2 first takes a dive into the economics literature on clustering and discusses classical theoretical constructs. It then describes the data and methodology applied to service robotics patents and German R&D data, while appendix tables report on all findings, results subsection focuses on a listing of the most interesting ones. The Section concludes with discussion about contributions to the geographical economics literature. The Section 4.3 develops the idea of random benchmarking elaborating on the literature about persistent growth and its implications. It then describes the methodology developed to study persistent growth using data of four European economies, followed by presentation of the resulting mixed evidence. Conclusion subsection summarizes main methodological contributions of this chapter.

4.2 Spatial distribution of innovative activities: The example of Germany[2]

4.2.1 Spatial distribution of innovative activities

Economists have theoretically and empirically demonstrated a positive relationship between investments in research and development (R&D), resulting innovations and economic growth. Models of endogenous growth lead to the conclusion that R&D is one of the main drivers of national welfare (Romer 1990b, Grossman and Helpman 1991, Aghion and Howitt 1992). These theories are supported by multiple empirical studies that also confirmed the importance of R&D for technological progress and productivity (see, for example, Akcay 2011 for a survey of this literature).

Given the broad literature regarding the spatial distribution of innovation, R&D, and industrial activity, this work aims to fill a gap regarding locational patterns of R&D input by empirically exploring micro-geographic data for Germany. In order to measure spatial concentration early studies, as for example Krugman (1991) and Audretsch and Feldman (1996), use a locational Gini coefficient. However, as argued by Ellison and Glaeser (1997), one problem with this coefficient is that it may spuriously indicate the localization of an industry resulting from the lumpiness of plant employment[3]. Ellison and Glaeser (1997) improve this approach by offering an alternative index, that controls for the organization of an industry by adopting a so-called dartboard approach (Ellison and Glaeser (EG) approach).

The approach compares the degree of spatial concentration of employment in a given sector with the degree of concentration that would arise if all plants in that sector were located randomly across locations. In other words, it answers the question of whether a location behavior of plants can be distinguished from a random distribution of plants in a given country. However, the approach has mainly been criticized as it relies on a discrete definition of space and is, thus, affected by the underlying spatial zoning system, i.e. shape, size and relative position of spatial units.[4]

[2] Parts of this work benefited from collaboration with Dr. Andrea Hammer and Dr. Florian Kreuchauff

[3] The expression lumpiness of plant employment relates to different patterns of plant size distributions each leading to the same amount of total employment.

[4] For further elaborations on the so-called Modifiable Areal Unit Problem (MAUP) see Briant et al. (2010).

The critique, together with enhanced availability of micro-geographic data sets, has lead Duranton and Overman (2005) to develop an approach (Duranton and Overman (DO) approach) that is based on continuous space by utilizing address data of establishments. In order to assess statistical significance of the deviation from randomness, the density distribution of bilateral distances is compared to counterfactuals constructed by randomly distributed establishments with the help of simulations.

Although, both the EG and the DO approach have been widely adopted in the literature in order to measure industrial concentration[5], a few studies use them to determine agglomeration patterns of innovation-related activities. Moreover, the scarce evidence on innovation-related activities based on the DO approach mostly refers to patent data and technology classes (Murata et al. 2014, Kerr and Kominers 2015). Only two studies by Buzard and Carlino (2009) and Carlino et al. (2012) relate to the DO approach in order to analyze locational patterns of R&D establishments. However, these studies only cover geographic partial areas of the United States and do not differentiate between industries.

This work conducts the analysis using the data provided by the "Stifterverband für die Deutsche Wissenschaft" (Donors' Association for the Promotion of Sciences and Humanities in Germany) that constitutes the most comprehensive database for private R&D in Germany. In total, the analyses is based on 19,804 company R&D establishments in Germany that employ 476,575 researchers in all economic sectors – agriculture, production industries and service industries.

It is revealed that with reference to the overall spatial distribution of R&D, 40.8% of 3-digit industries exhibit significantly different patterns of spatial R&D organization. In general, deviations occur more often in the production industry than in the service sector. Moreover, production industries exhibit a higher propensity to cluster in geographical space. However, taking distances into account, clustering of R&D activities in production industries mostly occurs at relatively high distances of around 100km. Deviations from spatial randomness in service industries tend to exhibit dispersion, i.e. for service industries, we do find statistically significant larger distances between R&D establishments than we would expect from taking the overall spatial distribution of R&D as a reference.

[5] See, for example, Duranton and Overman (2005, 2008) for the UK, Albert et al. (2012) for Spain, Nakajima et al. (2012) for Japan, Barlet et al. (2013) for France and Koh and Riedel (2014) for Germany.

This section is organized as follows. *Section 4.2.2* introduces the database, descriptive statistic and the basic estimation methodology. Results on spatial patterns of industry-specific R&D are presented in *Section 4.2.3*. Finally, conclusions are derived in *Section 4.2.5* together with the policy implications.

4.2.2 Data and basic estimation methodology

In the following we discuss the data used in our analysis together with the basic estimation methodology (DO approach). The introduction of the database includes both the description of the database and the first descriptive statistics on R&D on the level of industry divisions for Germany. Subsequently, the basic estimation methodology is presented to depict industry-specific location patterns of company R&D establishments. It implies estimating industry-specific estimations of kernel density functions and counterfactuals based on measures of great-circle distances. The methodology is illustrated by exemplary location patterns of R&D on the level of 3-digit industries.

R&D-survey and descriptive statistics of R&D in Germany

In order to identify location and size of R&D establishments in Germany we use data from the biennial survey conducted by the "Stifterverband für die Deutsche Wissenschaft" (Donors' Association for the Promotion of Sciences and Humanities in Germany) which constitutes the most comprehensive database for private R&D in Germany. By means of a standardized written survey the Stifterverband collects data reflecting different aspects of company R&D activity – e.g. internal and external expenditures, personnel, location and size of establishments – on behalf of the German Federal Ministry of Education and Research. The survey is designed as full census, so that it raises the claim to cover the whole population of companies conducting R&D in Germany. Reporting unit on company level is usually the smallest independent accounting unit. All companies in Germany that are assumed to conduct R&D are included in the survey. They are identified by preceding surveys and auxiliary variables – including industry, company size and information on public R&D funding. However, as pointed out by the Stifterverband, the detection of all companies in Germany that conduct R&D remains a challenge as no complete database exists. Thus, although the survey is designed as a full census, the coverage

Table 4.1: Size distribution of companies and R&D companies in Germany

	Employees subject to social insurance				
	0 to <10	**10 to <50**	**50 to <250**	**250 and more**	**Total**
No. of companies [Germany, 2013]	3,290,579	268,263	57,712	13,112	3,629,666
Share [%]	90.7	7.4	1.6	0.4	100.0
No. of companies [Sample]	3,139	7,431	5,510	2,790	18,870
Share [%]	16.6	39.4	29.2	14.8	100.0

might be incomplete, especially with respect to small and medium-sized companies (Stifterverband für die Deutsche Wissenschaft 2015).

Table 4.1 compares the overall company structure in Germany in 2013 to the R&D company structure extracted from the database provided by the Stifterverband. The size distribution of R&D companies in the database is skewed towards bigger companies. This leads us to assume that – compared to the overall company size distribution – bigger companies are more likely to conduct R&D activities. This conclusion is in accordance with evaluations for Germany based on the KfW panel[6] over the years 2005 to 2012 where shares of companies conducting R&D increase from 24.0% for companies with 0 to less than 10 employees over 41.0% for companies with 10 to less than 50 employees up to 60.0% for companies with 50 to less than 250 employees (Baumann and Kritikos 2016).

In order to identify spatial patterns of private R&D activity, the adequate unit of analysis is not the company but the company's R&D establishments. Because the survey collects information on the postcodes of a company's R&D establishments and of the fraction of total R&D workforce employed in these establishments, we are able to identify not only the location of R&D establishments, but also their size in terms of the number of researchers employed. Thus, for every R&D establishment we know its postcode, its 2- and 3-digit industrial classification (Statistical classification of economic activities in the European Community (NACE) Rev. 2), and its size. We assume that a private R&D activity is a long term investment and, therefore, subsume five consecutive surveys of the years 2005, 2007, 2009, 2011 and 2013. This allows us to gather data on 19,804 R&D establishments that occupy in total 476,575 researchers in Germany.[7] As each establishment is assigned a unique identifier, we

[6] The KfW SME panel ("KfW Mittelstandspanel") is a representative survey of micro, small and medium-sized companies in Germany that have an annual turnover of up to 500 Million Euro.

[7] Note that by merging five consecutive surveys we implicitly assume that the spatial distribution

Table 4.2: Size distribution of R&D establishments

	Number of researchers				
	0 to <10	**10 to <50**	**50 to <250**	**250 and more**	**Total**
No. of R&D establishments	14,398	4,042	1,080	284	19,804
Share [%]	72.7	20.4	5.5	1.4	100.0

exclude multiple entries by taking the most current information available in the database. In total, we identify 81 industries at the 2-digit and 235 industries at the 3-digit level of industrial classification with at least one R&D establishment. Out of the 235 industries, that we identify at the 3-digit level, 140 have more than ten R&D establishments. *Table 4.2* shows the size distribution of R&D establishments. The majority of R&D establishments (72.7%) employ less than 10 researchers, while the fraction of big R&D establishments with 250 and more researchers employed is only 1.4%.

Table C.1 and *Table C.2* in the Appendix C depict the descriptive statistics at the 2-digit industry level, i.e. statistical divisions, for agriculture, production industries and the service industries in terms of the number of establishments, R&D establishment-company ratio, average number of researchers per R&D establishment and number of 3-digit industries contained. Analyzing the number of establishments and the number of researchers employed shows a dominance of production industries, especially of manufacturing (divisions 10 to 33), concerning not only the number of R&D establishments but also and even more the number of researchers employed. While 67.0% of all R&D establishments and 81.1% of the R&D workforce are in manufacturing (70.6% and 82.0% in production industries), 25.6% of R&D establishments and 17.6% of R&D workforce are in the service sector. However, the biggest divisions with more than 1,000 R&D establishments[8] are not only in manufacturing, but also in the service sector. In the production industries, the majority of industry divisions shows an R&D establishment-company ratio greater

of R&D establishments in space is solid and they are not easily moved in geographical space. This is a quite restrictive assumption on spatial dynamics of R&D. However, merging of data is necessary in order to collect information on as many R&D establishments as possible as the DO approach requires at least ten establishments per industry in order to derive significant results on location patterns.

[8] *25 Manufacture of fabricated metal products, except machinery and equipment*, *26 Manufacture of computer, electronic and optical products*, *28 Manufacture of machinery and equipment n.e.c.*, *62 Computer programming, consultancy and related activities*, *71 Architectural and engineering activities; technical testing and analysis*, *72 Scientific research and development*

than one, indicating that in most divisions the number of R&D establishments exceeds the number of companies conducting R&D. In contrast, in the service sector divisions the ratio often is exactly one indicating that in many service industries R&D companies only dispose on one R&D establishment. However, as the total R&D establishment-company ratio is 1.05, establishing several R&D establishments seems to be quite rare for most companies that conduct R&D. Looking at the average number of researchers per R&D establishment shows substantial differences among divisions ranging from 2.5 (*56 Food and beverage service activities*) to 212.6 (*29 Manufacture of motor vehicles, trailers and semi-trailers*) researchers per R&D establishment. The average size of R&D establishments in terms of R&D workforce is 24.1.

The locations of establishments are geocoded by using centroids of postcodes. In Germany, postcodes are very useful for locating establishments because they cover relatively fine grained areas. In comparison to 402 Nomenclature of Territorial Units for Statistics (NUTS) 3 regions, we have identified 8,212 postcode areas. In 4,865 of them at least one R&D establishment is located. *Figure 4.1* demonstrates the distribution of postcodes in geographical space differentiating between production and service industries. On average, each postcode belongs to 4.1 establishments with a minimum value of one R&D establishment for 30.9% of the postcodes and a maximum value of 106 R&D establishments for one postcode in Berlin. More than 90% of the postcodes are home to less than ten establishments.

Basic estimation methodology

Estimating kernel density functions

To assess the spatial concentration of R&D establishments in an industry, we first calculate great circle distances[9] between all R&D establishments in that industry which generates $\frac{n(n-1)}{2}$ unique bilateral distances. The great circle distances only serve as a proxy for true geographical distances, thus distributions are kernel-smoothed in order to estimate the industry specific distribution of bilateral distances between R&D establishments. The estimator of the density of R&D establishments in a given industry m at any distance d is:

[9] $d = acos(sin\phi 1 * sin\phi 2 + cos\phi 1 * cos\phi 2 * cos\Delta\lambda) * R$, with d=distance, ϕ=latitude, λ=longitude, R=radius

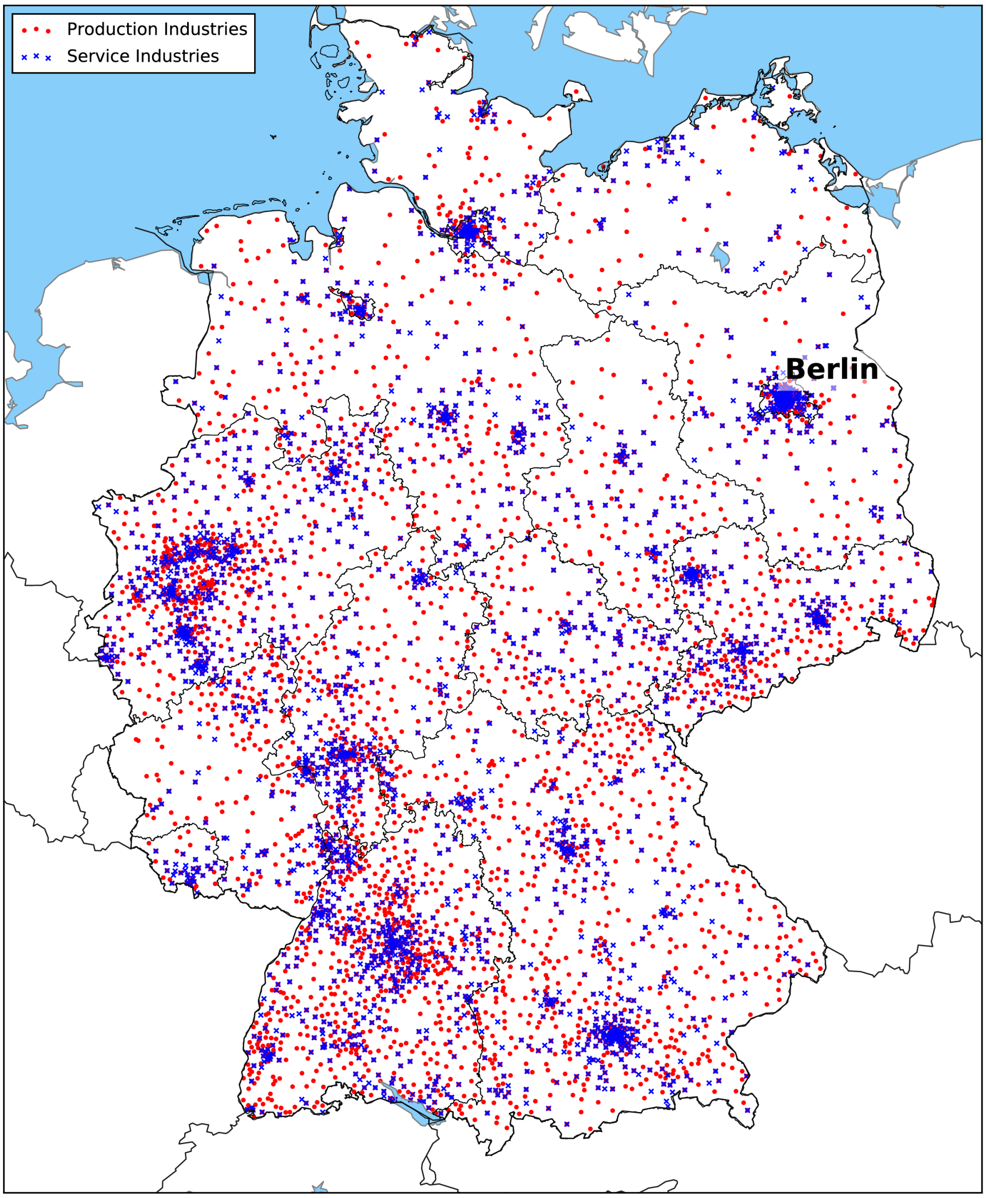


Figure 4.1: Location of production and service industry R&D establishments in Germany.

$$\hat{K}_m(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{ij}}{h}\right), \tag{4.1}$$

where h is a bandwidth parameter defined according to Silverman (1986), and f a standard Gaussian kernel function. $d_{i,j}$ depicts the bilateral distance between R&D establishments i and j. n is number of R&D establishments in a given industry m.

Counterfactuals for each industry are constructed in order to assess weather the estimated kernel density functions significantly deviate from the overall location pattern of R&D. We first determine all sites in Germany where R&D facilities could possibly be located as a reference. Note that the general spatial distribution of R&D activity in Germany – which we take as a reference – was formed historically being influenced by a variety of factors. *Figure 4.1* indicates that this basic distribution is not random and exhibits clustered as well as dispersed areas. However, addressing the question of why this general location pattern of R&D occurs goes beyond the scope of this work. Instead, by taking the spatial distribution of R&D as a reference, we explore whether there are deviations from this general picture, implicitly controlling for other factors that have influenced the observable spatial pattern of R&D. Counterfactuals are then constructed by randomly drawing n R&D establishments from the population of all R&D establishments in Germany, and determining kernel density functions for their bilateral distance distributions. To be able to draw statistically significant conclusions, we run 1000 simulations for each industry (Duranton and Overman 2005, 2008).

After calculating $\hat{K}_m(d)$ and constructing the counterfactuals, both need to be compared. To make comparison easier across industries and to account for the redundancy of information on long distances, we choose a threshold of 456km which corresponds to the median distance of all R&D establishments in Germany. This redundancy occurs as the area under each kernel density function needs to sum to unity. Thus, information on long distances is redundant if we know what happens at relatively short distances. In order to be able to make statements about deviations from randomness over the entire range considered in our analyses, we calculate and draw global confidence bands, so that only 5% of the randomly generated kernel density functions cross the upper $\bar{K}_m(d)$ and lower $\underline{\mathrm{K}}_m(d)$ global confidence bands for all $d \in [0, 456]$.[10] If $\hat{K}_m(d) > \bar{K}_m(d)$ for at least one distance $d \in [0, 456]$, R&D in

[10] In our work we implement *global* confidence bands meaning that we always focus on the whole

that industry is said to exhibit localization. Accordingly, if we find $\underline{\mathrm{K}}_m(d) > \hat{K}_m(d)$ for at least one distance $d \in [0, 456]$ and we do not detect localization, R&D in that industry exhibits dispersion. Perhaps another way of explaining it is that R&D localization (dispersion) in an industry is thus observed when there are more (less) R&D establishments at shorter distances than would be expected if firms would have chosen location sites at random. *Figure 4.2* illustrates examples of one localized (a), one random (b) and one dispersed (c) industry at the 3-digit level together with their respective maps of R&D establishments. In (a) we observe $\hat{K}_m(d) > \bar{K}_m(d)$ for all distances from 0km to 232km and thus localization of R&D activity. In (c) we detect no localization but $\underline{\mathrm{K}}_m(d) > \hat{K}_m(d)$ for all distances from 0km to 99km. This leads to the conclusion that the industry exhibits dispersion. Industry (b) exhibits a random location pattern. The shape of the confidence bands reflects the distribution of R&D in Germany for an average industry with the same amount of establishments as in industry m.

Following the reasoning set out above, we define

$$\Gamma_m(d) \equiv max(\hat{K}_m(d) - \bar{K}_m(d), 0) \tag{4.2}$$

as an index of localization and

$$\Psi_m(d) \equiv \begin{cases} max(\underline{\mathrm{K}}_m(d) - \hat{K}_m(d), 0), \text{if } \sum_{d=0}^{d=456} \Gamma_m(d) = 0 \\ 0, otherwise \end{cases} \tag{4.3}$$

as an index of dispersion. To reject the hypothesis of randomness of R&D for industry m at distance d because of localization (dispersion) $\Gamma_m(d) > 0$ ($\Psi_m(d) < 0$) is sufficient. In order to indicate to which degree an industry is dispersed or localized we define the following cross-distance indices $\Gamma_m \equiv \sum_{d=0}^{d=456} \Gamma_m(d)$ and $\Psi_m \equiv \sum_{d=0}^{d=456} \Psi_m(d)$ as indices of localization and dispersion across all distances $d \in [0, 456]$.

As noted already by Duranton and Overman (2005) the methodology described is sensitive to the number of R&D establishments in an industry. Industries with relatively few R&D establishments will show a very broad confidence band since there are many possible ways to randomly draw this small number out of the whole population of possible R&D establishment sites. We have thus chosen to analyze

range of distances. The literature also sometimes refers to *local* confidence bands defined for each distance independently. This constitutes a less strict definition of deviation from randomness.

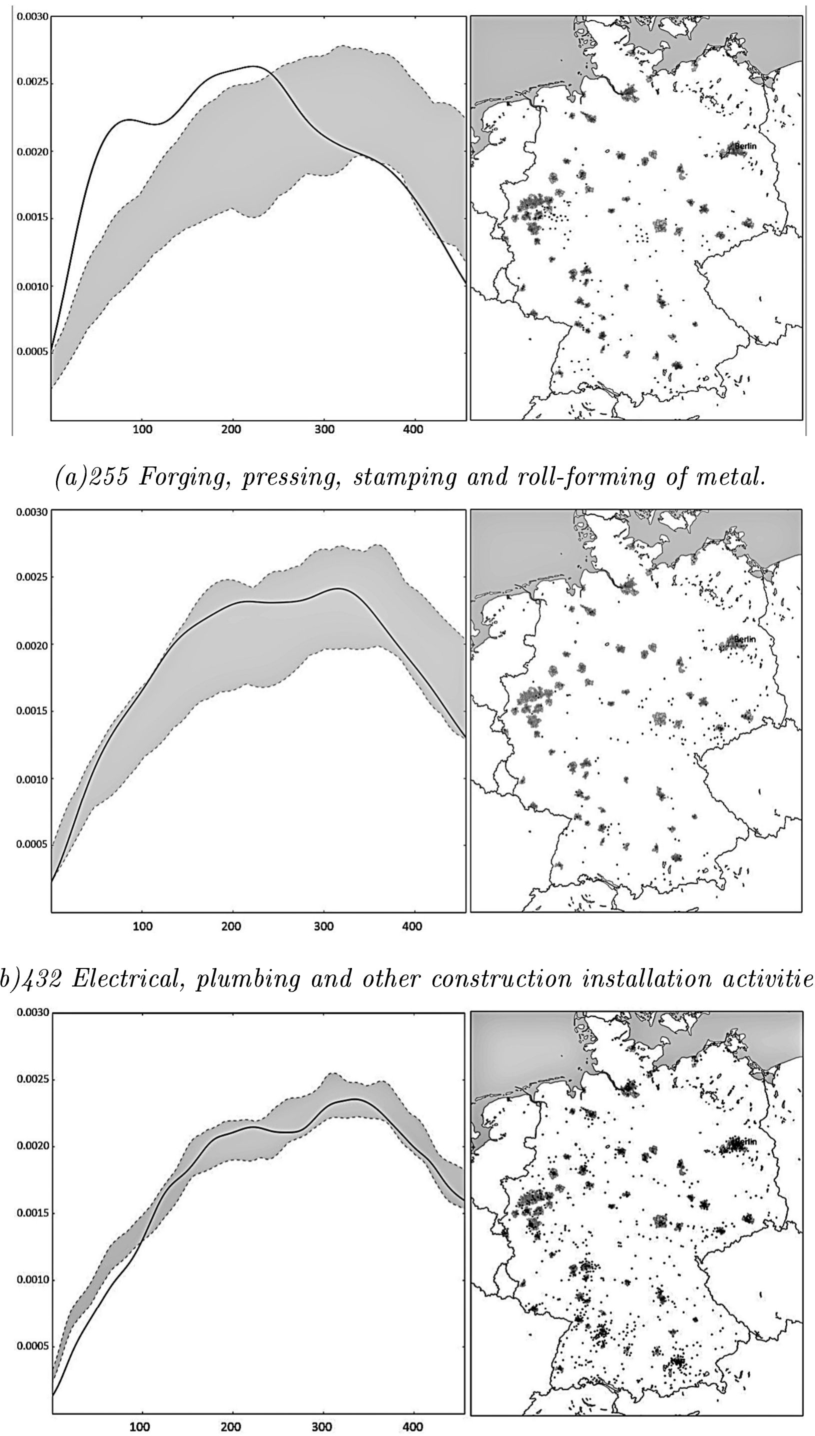
locational patterns of industries with more than ten establishments only since below that number we are not able to draw statistically significant conclusions. This leads us in total to simulating and analyzing 140 industries on the level of 3-digit industries of which three are in agriculture, 100 are in the production industry and 37 in the service sector.

4.2.3 Spatial patterns of industry-specific R&D in Germany

Two approaches are employed for our analysis. Firstly, we apply the basic estimation methodology introduced in *Section 4.2.2* in order to determine if industrial location patterns of corporate R&D are random, localized or dispersed in relation to the overall distribution of R&D. Calculating cross-distance indices of localized and dispersed industrial R&D activities leads to the identification of industries exhibiting relatively strong deviations from randomness. The results derived are put in relation to findings on locational patterns for economic activities in Germany. Secondly, by modifying the basic estimation methodology, we shift the level of analysis from R&D establishments to the individual researcher. This researcher-weighted approach leads us to derive further insights regarding the spatial organization of R&D.

Location patterns of corporate R&D establishments

We first explore the sectoral scope of location patterns in order to detect if 3-digit industries belonging to the same industry division – and thus conducting R&D for the same group of products or services – exploit identical spatial organization patterns of R&D. *Table C.3* and *Table C.4* in the Appendix C depict the shares of localized, dispersed and randomly distributed 3-digit industries within each industry division. In general, we can say that 3-digit industries in the same division do not follow identical patterns of R&D location. This leads us to conclude that, even within divisions, R&D activities follow their own specific spatial patterns. This holds especially for the production industry where we find diverse location patterns. For example, the six 3-digit sub-industries of industry division *20 Manufacture of chemicals and chemical products*, are to one third localized, dispersed and randomly distributed across geographical space.



(a)255 Forging, pressing, stamping and roll-forming of metal.

(b)432 Electrical, plumbing and other construction installation activities.

(c)620 Computer programming, consultancy and related activities

Note: X-axis on a graph indicates distances in km and Y-axis probability density.

Figure 4.2: Examples of the industrial location patterns of R&D establishments.

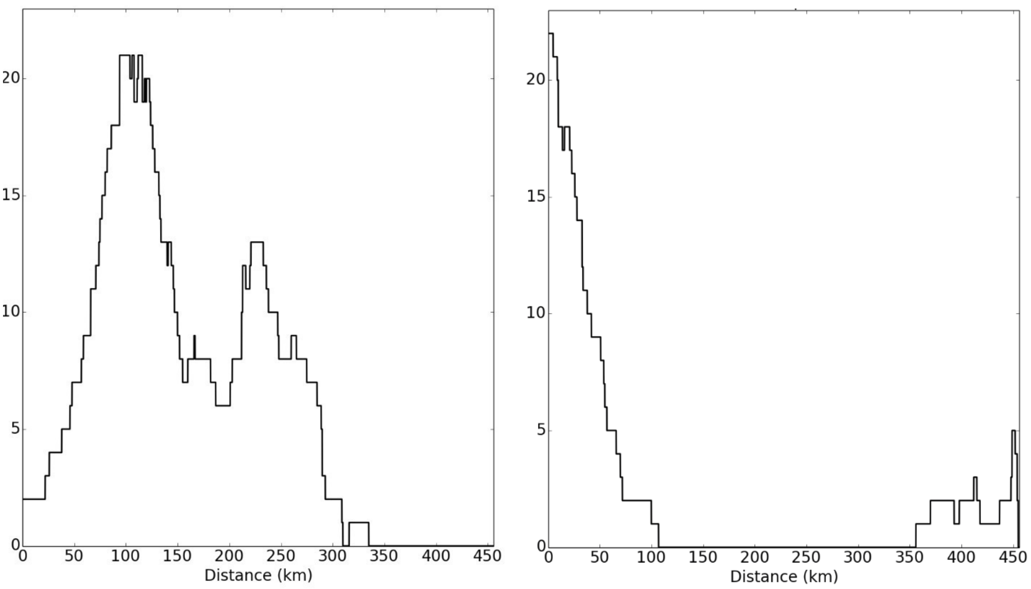
These diverse patterns lead us to relate to the 3-digit aggregation level in our analysis. We also report our results highlighting production and service industries because the nature of R&D activity in these sectors differs significantly regarding organization and content. In service industries, R&D is not always organized as formally as in the production industry; for example, it is unusual for firms in most service industries to have an own R&D department. Major developments are more likely to be conducted by temporary project development groups. Regarding content, social sciences and design activities play a more prominent role in service R&D than in production-oriented R&D.[11]

Comparing the kernel density estimates for R&D in every industry on the 3-digit level with the industry specific counterfactuals shows that R&D activities of 40.8% of industries deviate significantly from random spatial location patterns of total R&D being localized or dispersed. Deviation from randomness occurs more often in production than in service industries: While 50.0% of all industries in production deviate from random spatial distribution, the share of diverging industries in services is only 18.9%. In agriculture, spatial distribution of R&D activity is random for all industries implying that the location of innovation creation centers in the agricultural sector is influenced by factors affecting the overall spatial distribution of R&D in Germany.

Looking at the direction of deviations from spatial random distribution, we find 22.9% of all industries localized and 17.9% dispersed. Differentiating between production and service industries leads to further insights. With 30.0% of localized industries within the production sector, the share of localized industries is considerably higher than in services where we only find 5.4% of industries exhibiting localized R&D patterns. Regarding dispersion, we detect dispersed R&D activities in 20% of production and in 13.5% of service industries, leading to a conclusion that knowledge creation in production industries tends to be more localized than in services.

Taking a more detailed look at the spatial patterns of localized and dispersed industries we identify at which distances these patterns are observable. *Figure 4.3* shows the number of localized and dispersed industries at each distance for all 3-digit industries. Note that if both localization and dispersion occur in the same industry, localization drives out dispersion. Consequently, an industry is only defined as being dispersed, if for all distances $d \in [0, 456]$ no localization occurs. Thus,

[11] In their definition of R&D activity the Stifterverband follows the comprehensive concept put forward in the Frascati Manual (Stifterverband für die Deutsche Wissenschaft 2015, OECD 2002). Although this concept relates to a relatively broad definition of R&D aimed at covering both organizational and content-related differences between R&D in production and service industries, it may lead to under-coverage of R&D activity in the service sector. For a comprehensive review see Miles (2007). Bryson et al. (2004) list R&D sources in service industries.



(a)Localization *(b)Dispersion*

Figure 4.3: Distance patterns of industries exhibiting localization and dispersion in R&D.

dispersion occurring in localized industries is not represented by the distance-based frequency distributions set out in *Figure 4.3*. While only 6.3% of localized R&D activities are localized at a distance interval from 0km to around 20km, we observe a constant increase of localized industries up to a distance of approximately 95km where 65.6% of all localized industries show significant localization. The frequency distribution of dispersed industries shows a sharp decrease of dispersion with growing distance. Dispersion occurs at a range of distances from 0km (88% of samle) till 110km. These spatial patterns of localization and dispersion are observable for both R&D in production and service industries.

As we are the first to apply the DO approach in order to analyze spatial variations of industry-specific R&D activities with reference to the overall spatial pattern of R&D, we are not able to classify our results with respect to other studies. However, comparing our findings to locational patterns of economic activity leads to further interesting insights. For our comparison we mainly refer to Koh and Riedel (2014) who applied the DO approach on all plants in manufacturing and services in Germany with at least one employee subject to social insurance.[12] Taking the overall establishment distribution in Germany as a reference, they find that 78.0% of industries are localized and that the share of localized industries is substantially

[12] The results of Koh and Riedel (2014) are based on industry classification NACE Rev 1.1 (WZ 2003) at the four digit-level. Nevertheless, rough comparisons to our data are still possible.

higher in services (98.0%) than in manufacturing (71.0%). Accordingly, they find low shares of dispersed industries. Relating their observations to distances they – in accordance with other studies on the spatial distribution of economic activities (e.g. Duranton and Overman 2005, Barlet et al. 2013) – find localization at small distances and a quite equal distribution of dispersion over all distances. These results differ considerably from our aforementioned findings of industry-specific R&D activities. In general, the different findings lead us to conclude that industry-specific deviations from the general spatial distribution are rarer in R&D than in economic activity, and if deviations from the overall spatial patterns occur, than dispersion is of more relevance for industry-specific R&D patterns than for industry-specific economic activity. These general differences become even more pronounced when we look at services.[13]

Analyzing the geographical patterns of the most localized and dispersed industry-specific R&D activities identified by cross-distance indices Γ_m and Ψ_m leads to further interesting insights. *Table 4.3* and Table *4.4* each depict the ten most localized and dispersed 3-digit industries in production. With *243* [14], *255* [15], *259* [16] and *257* [17] four of the most localized industries in terms of R&D activity are part of the metal processing industry. The highest index of localization is measured for *243* [18] where spatial concentration of R&D establishments can be found in the Ruhr area. [19] For *255* [20], *259* [21] and *257* [22] we not only observe spatial concentration of R&D establishments in the Ruhr area but also in other parts of North Rhine Westphalia, Baden-Württemberg, Thuringia and Saxony. R&D in the industries *293* [23] and *222* [24] exhibit relatively high localization indices. However, taking into account the dis-

[13] Note that we do not compare the distribution of economic activity and R&D activity in general. I.e. our statements do not refer to the spatial concentration of the one relating to the other but on the within variation of activities with reference to the respective overall spatial distribution. Thus, our findings do not contradict the statement that R&D in general is more concentrated in geographical space than economic activity.

[14] Manufacture of other products of first processing of steel

[15] Forging, pressing, stamping and roll-forming of metal

[16] Manufacture of other fabricated metal products

[17] Manufacture of cutlery, tools and general hardware

[18] Manufacture of other products of first processing of steel

[19] Maps of localized industries where reference is made to specific regions or cities in Germany are depicted in *Appendix C.*

[20] Forging, pressing, stamping and roll-forming of metal

[21] Manufacture of other fabricated metal products

[22] Manufacture of cutlery, tools and general hardware

[23] Manufacture of parts and accessories for motor vehicles

[24] Manufacture of plastic products

Table 4.3: Most localized R&D activities in production industries

3-digit industry		**No. of R&D establishments**	Γ_m
243	Manufacture of other products of first processing of steel	36	0.1044
255	Forging, pressing, stamping and roll-forming of metal	143	0.0463
293	Manufacture of parts and accessories for motor vehicles	365	0.0395
222	Manufacture of plastic products	667	0.0226
231	Manufacture of glass and glass products	126	0.0163
284	Manufacture of metal forming machinery and machine tools	458	0.0148
259	Manufacture of other fabricated metal products	327	0.0127
139	Manufacture of other textiles	181	0.0122
143	Manufacture of knitted and crocheted apparel	20	0.0114
257	Manufacture of cutlery, tools and general hardware	343	0.0107

Note: An overview on all cross-distance indices of localization and dispersion is provided in *Appendix C.3*.

tance intervals of localization reveals that they are quite broad ranging from about 60km to 290km. These findings indicate a significant localization for R&D in these industries, however this clustering – in terms of distance – yet occurs on a relatively large geographical scale. R&D in industry *231* [25] is observable in Thuringia and Saxony. Like in *293* [26] and *222 Manufacture of plastic products* the distance interval of significant localization is broad and on a relatively large geographical scale starting at 86km and ending at 280km. For *284* [27] concentration of R&D establishments is observable in Baden-Württemberg. Finally, R&D activities in the textile related industries *139 Manufacture of other textiles* and *143* [28] in particular exhibit spatial concentration in the North of Bavaria and Saxony but also in some regions in Baden-Württemberg and North-Rhine Westphalia.

Indices of dispersion Ψ_m are on a lower level than indices of localization Γ_m indicating that deviations from randomness are weaker for dispersed than for localized R&D activities. In production, industries connected to the medical sector (*325 Manufacture of medical and dental instruments and supplies*, *212 Manufacture of pharmaceutical preparations)* and to the production of electrical equipment (*271 Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus*, *279 Manufacture of other electrical equipment*) as well as industries *251 Manufacture of structural metal products*, *236 Manufacture of articles of concrete, cement and plaster* and *205 Manufacture of other chemical products* are among the most dispersed. Compared to the overall spatial pattern of R&D in Germany, we see

[25] Manufacture of glass and glass products

[26] Manufacture of parts and accessories for motor vehicles

[27] Manufacture of metal forming machinery and machine tools

[28] Manufacture of knitted and crocheted apparel

less-than-usual concentrations of these industries in areas that are quite populated with R&D establishments (e.g. Ruhr Area and around Stuttgart). Although we observe significant dispersion for R&D in both industries *266 Manufacture of irradiation, electromedical and electrotherapeutic equipment* and *303 Manufacture of air and spacecraft and related machinery* distance intervals that exhibit dispersion start at relatively high distances, i.e. 370km and 356km. R&D activities in *108 Manufacture of other food products* show dispersion because they are located in more rural areas in North Rhine Westphalia and Saxony where general R&D activity is relatively low.

Table 4.4: Most dispersed R&D activities in production industries

3-digit industry		**No. of R&D establishments**	Ψ_m
325	Manufacture of medical and dental instruments and supplies	379	0.0028
212	Manufacture of pharmaceutical preparations	231	0.0023
266	Manufacture of irradiation, electromedical and electrotherapeutic equipment	124	0.0022
271	Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus	379	0.0017
251	Manufacture of structural metal products	247	0.0015
279	Manufacture of other electrical equipment	275	0.0014
303	Manufacture of air and spacecraft and related machinery	80	0.0011
108	Manufacture of other food products	122	0.0010
236	Manufacture of articles of concrete, cement and plaster	135	0.0007
205	Manufacture of other chemical products	281	0.0007

Note: An overview on all cross-distance indices of localization and dispersion is provided in *Appendix C.3.*

As mentioned above, the share of non-random spatial R&D distribution in service industries compared to production industries is relatively low. *Table 4.5* shows indices of localization Γ_m and dispersion Ψ_m for R&D in all service industries that deviate from randomness. The two service industries *711 Architectural and engineering activities and related technical consultancy* and *467 Other specialized wholesale* are the only service industries in which R&D activities are localized. However, distance intervals exhibiting localization start at 106 km and 166 km. This indicates that R&D activities in these industries are clustered at relatively long distances. Additionally, comparing the index values shows stronger localization of R&D in the ten most localized production industries than in the localized service industries. In total, we find five service industries with dispersion of R&D establishments. Interestingly, the four most dispersed industries *620 Computer programming, consultancy and related activities*, *721 Research and experimental development on natural sciences and engineering*, *712 Technical testing and analysis* and *631 Data processing,*

Table 4.5: Service industries exhibiting localized and dispersed R&D activities

3-digit industry		**No. of R&D establishments**	Γ_m
711	Architectural and engineering activities and related technical consultancy	1,175	0.0018
467	Other specialized wholesale	88	0.0002
3-digit industry		**No. of R&D establishments**	Ψ_m
620	Computer programming, consultancy and related activities	1,617	0.0061
721	Research and experimental development on natural sciences and engineering	994	0.0057
712	Technical testing and analysis	261	0.0022
631	Data processing, hosting and related activities	93	0.0021
702	Management consultancy activities	152	0.0008

Note: An overview on all cross-distance indices of localization and dispersion is provided in *Appendix C.3.*

hosting and related activities are all service industries that are identified as being research-intensive[29] and thus devote above average financial resources on R&D. In terms of index values, these dispersed service industries display index values higher or quite close to the index values of the ten most dispersed production industries. In *702 Management consultancy activities* the index of dispersion shows a relatively low value.

The comparison of results on geographical patterns of the most localized and most dispersed industrial R&D activities to patterns found in economic activities reveals new insights into spatial R&D organization. For example, economic activities in production industries, traditional manufacturing industries that evolved with the industrial revolution in the 19th century (e.g. industries connected to metal processing and textile) are among the most localized industries showing persistent localization patterns in traditional regions (Koh and Riedel 2014). Our analyses of localized industries reflect this observation regarding R&D activities in these traditional manufacturing industries. This leads us to conclude that relative spatial organization of R&D is partly congruent with relative spatial organization of economic activity in these traditional manufacturing industries. However, turning our attention to location patterns in services, our results on R&D distribution do not reflect the strong localization patterns regarding the administration of financial markets and the entertainment sector found for economic activity as we find R&D activities

[29] Gehrke et al. (2010, 2013) define research-intensive industries and services on a 3-digit level for Germany based on different data sources. The main criterion for identification is a threshold of 3% of R&D expenditures on sales. A complete list of research-intensive industries is provided in *Appendix C.2*, *Tables C.5* and *C.6*.

in these industries randomly distributed.

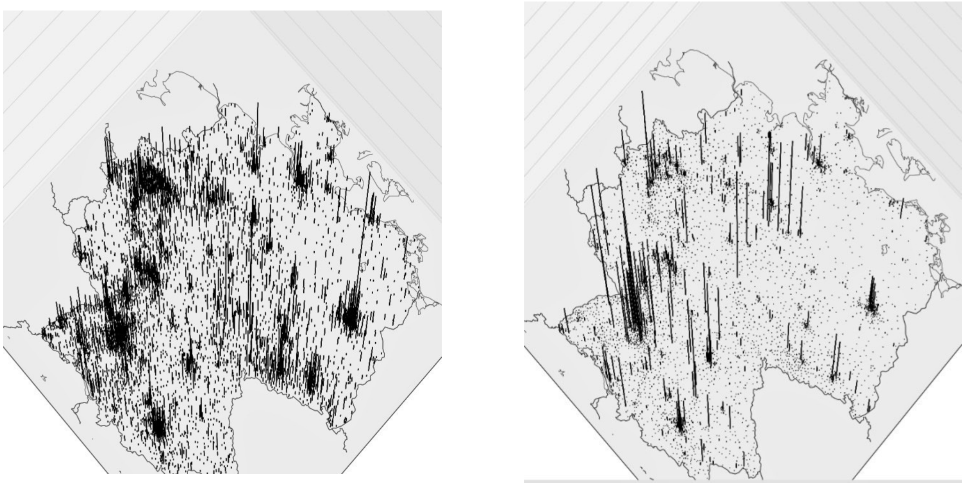
4.2.4 Researcher-weighted location patterns of corporate R&D

Up to this moment all conclusions are based on the spatial distribution of R&D establishments. In other words, when assessing the deviation from randomness we take into account current location of R&D regardless of the number of people that conduct research there. However, in order to deepen our understanding of the spatial organization of R&D, it seems reasonable not only to focus on places where people are employed in knowledge creation, but also to take into account how many of them are involved in the process. This approach shifts the unit of analysis from the individual R&D establishment to the individual researcher. The issue of R&D establishment size in terms of researchers employed is crucial as R&D establishment-size distributions, like company-size distributions, are skewed. For example, 72.7% of R&D establishments in our dataset employ less than ten researchers but account for only 11.2% of total R&D workforce.

Some previous studies concerned with spatial patterns of economic activity tackled the issue of skewed company-size distributions by censoring smallest plants in industries applying absolute or relative thresholds or by weighting according to the number of employees. The former in our case is not advisable as, given the limited size of our data in terms of R&D establishments compared to establishments reflecting general economic activities, it will lead to omitting a number of industries in the analysis. We thus choose to weight according to the number of researchers employed in R&D establishments. Following Duranton and Overman (2005) in this shift in unit of analysis from establishment to workforce, we exclude zero distances between researchers employed at the same R&D establishment in order to avoid that localization might be driven by the concentration of research personnel within a particular establishment. Formally, denoting $r(i)$ as research personnel of R&D establishment i and respectively $r(j)$ as research personnel of R&D establishment j the researcher-weighted kernel density function of industry m takes the following form:

$$\hat{K}^r_m(d) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r(i)r(j)f\left(\frac{d-d_{ij}}{h}\right)}{h\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r(i)r(j)}. \tag{4.4}$$

All other variables are defined according to *Equation 4.1.*



(a)R&D establishments

(b)Researchers

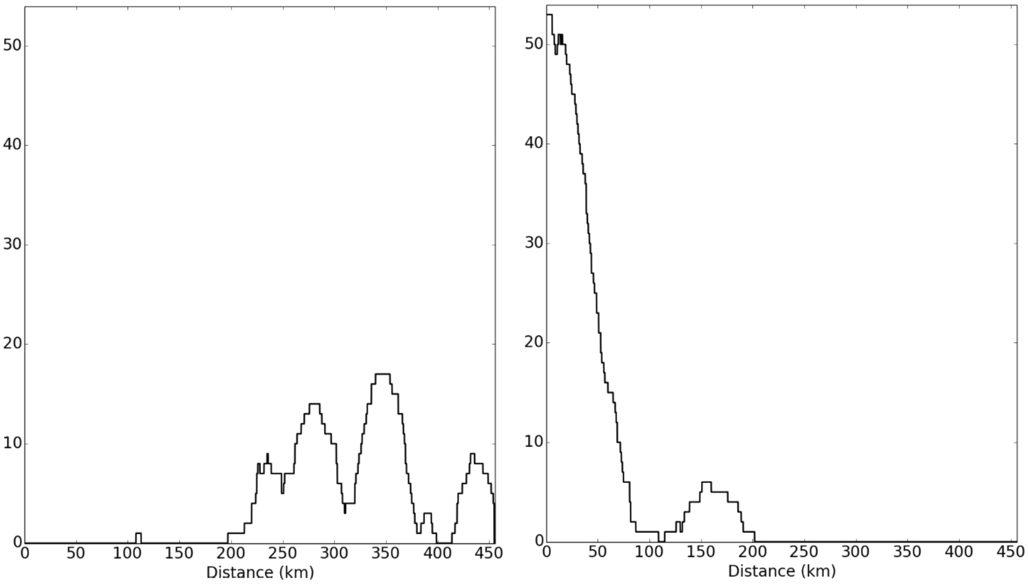
Figure 4.4: Spatial frequency distribution of R&D establishments and researchers.

Counterfactuals, confidence bands and indices are constructed following the procedure described in *Section 4.2.2.* Technically, taking into account that our spatial modeling is based on postcodes, in constructing our counterfactuals the base for simulations is a new distribution of selection probabilities for postcodes. Before we turn to the results of the researcher-weighted approach, we should devote some attention to the implications of this shift in approaches. *Figure 4.4* visualizes the resulting differences in distributions and thus selection probabilities that constitute the base for the counterfactuals. At first sight, we not only see a general shift of R&D activity towards south-western regions of Germany but also a higher concentration of R&D activities in individual postcodes implying that the distribution in the researcher-weighted approach becomes more skewed. This change is reflected by the fact that the Gini-coefficient for the frequency distribution of postcodes augments from 0.49 in the non-weighted approach to 0.81 in the researcher-weighted approach. Statistically, the selection probability of 55 postcodes increases more than tenfold, including three postcodes where it augments by more than 100. The researcher-weighted approach also induces changes regarding the weighting of industries in the reference distribution of R&D.[30] Although the Gini-coefficient for the frequency distribution of industries only increases slightly from 0.78 in the non-weighted approach to 0.86 in the researcher-weighted approach, we see one industry, *291 Manufacture of Motor Vehicles*, which accounts for 0.3% of establishments and 11.7% of R&D personnel.

[30] As we know from the descriptive data in *Section 4.2.2*, production industries represent 70.6% of R&D establishments and 82.0% of researchers.

Thus, selection probability of postcodes occupied by that industry increases by factor 41 when the number of researchers is taken into account instead of R&D establishments. Taking a closer look at industry *291 Manufacture of Motor Vehicles* reveals that 57 R&D establishments in total employ 55,702 researchers. Moreover, 94.5% of all these researchers are employed by seven automotive manufacturers: Daimler, Volkswagen, BMW, Audi, Porsche, Opel and Ford. Thus, analyzing the results for researcher-weighted postcodes, it is important to keep in mind that not only selection probabilities are distributed more unequally between postcodes but also that they are influenced considerably stronger by the location pattern of the motor vehicles industry that in turn is dominated by few big automotive manufacturers.

In total, the researcher-weighted analyses show that 58.6% of industries deviate from randomness with 17.2% being localized and 41.4% exhibiting dispersion. Looking at industry sectors, in production industries we find 63.0% deviating from randomness of which 21.0% exhibit localization and 42.0% dispersion. In services 48.6% industries are not randomly distributed with 8.1% being localized and 40.5% dispersed. Overall we note more deviation from randomness in the researcher-weighted approach than in the establishment-based approach. Especially the share of industries exhibiting dispersion increases in both services and the production industry.



(a)Localization

(b)Dispersion

Figure 4.5: Distance patterns of researcher-weighted industries exhibiting localization and dispersion in R&D.

A detailed look at the distances at which industries are localized or dispersed (*Figure*

4.5) reveals pictures similar to the unweighted approach.[31] However, if before we observed most of the localized industries at a distance of around 100km, we now see that they are concentrated around 260km and 350km. This means that industrial clusters of R&D activity from the perspective of an individual researcher occur at a higher distance. In total, our results indicate that taking into account the size of R&D establishments in terms of researchers employed there, leads either to dispersion or random distribution at close distances from 0km to 200km. This implies that at this distance interval the number of researchers in an industry either follows the general distribution of R&D workforce or is even less than one would expect from taking the general spatial distribution of researchers as a reference. Again, we need to keep in mind that these results do not contradict the notion of R&D itself being concentrated compared to economic activities. They indicate that clustering of researchers at short distances is not connected to the 3-digit industries in which they are employed.

Calculating the cross-distance indices for localization and dispersion in the researcher-weighted approach reveals major changes in both production and service industries in relation to the approach based on R&D establishments.[32] Not only that – as one might conclude from the distance distributions depicted in *Figure 4.5* – indices of localization become weaker and indices of dispersion become stronger, but also radical shifts in locational patterns occur. For example, four of the most dispersed production industries (*325 Manufacture of medical and dental instruments and supplies*, *271 Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus*, *251 Manufacture of structural metal products* and *205 Manufacture of other chemical products*) and the two most dispersed service industries (*620 Computer programming, consultancy and related activities*, *721 Research and experimental development on natural sciences and engineering*) in terms of R&D establishments become localized. However, these localization patterns occur at relatively large distances and this is why we do not explore these industrial spatial patterns in more detail.

[31] Distance patterns for production industries and service industries are provided in *Appendix C.1*, *Figure C.2*

[32] An overview on all cross-distance indices of localization and dispersion is provided in *Appendix C*.

Service Robotics

In Chapter 3 we have introduced a patent search strategy for service robotics and identified German firms patenting in the field. Here with the methodology applied above we can assess clustering activity of these firms depicted on Figure 3.8. Figure 4.6 demonstrates that the observed cluster of service robotics firms significantly deviates from the overall R&D distribution in Germany. Literature demonstrates that the general economic activity tends to be geographically concentrated while innovation-related activities – like, for example, R&D activities – are even more spatially clustered.

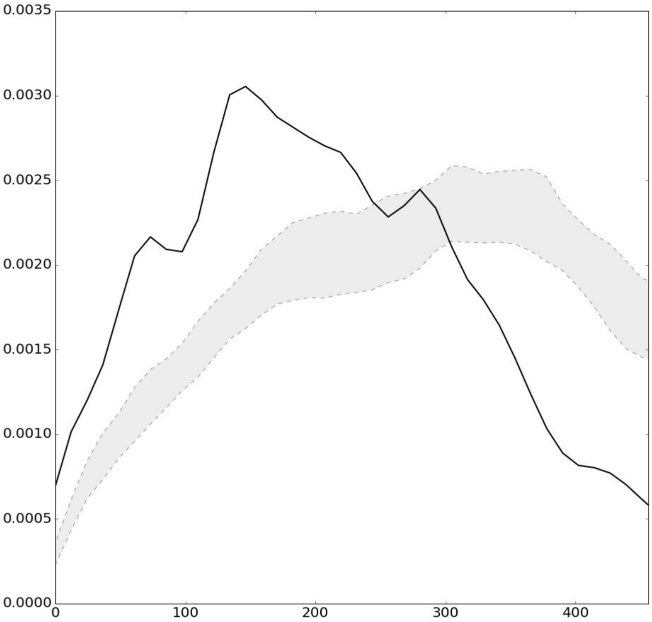


Figure 4.6: Clustering of service Robotics.

These observable spatial patterns of R&D might be related to multiple Marshallian channels, i.e. specialized inputs, labor market sharing and knowledge spillovers. However, empirical evidence indicates that they are mostly linked to knowledge spillovers which are not only limited in geographical space (Rosenthal and Strange 2004, Ellison et al. 2010) but also fostered by high densities of people (Glaeser et al. 1992, Henderson et al. 1992, Glaeser 1999, Bettencourt et al. 2007, Sedgley and Elmslie 2011) and industrial structures (Marshall 1920, Jacobs 1969, for a recent overview see: Beaudry and Schiffauerova 2009). This leads to the conclusion that even though the yield of R&D activities is influenced by multiple aspects, the exchange of ideas and thus the case of physical proximity remains a key ingredient.

4.2.5 Conclusion on the study of spatial clustering

This section empirically contributes to the literature about clustering of innovation and R&D activity by indicating if and how spatial patterns of R&D in industries deviate from the overall spatial distribution of R&D in Germany.

Analyzing the industry location patterns of R&D on a 3-digit level reveals that 40.8% of industries deviate significantly from random spatial location patterns and thus are localized or dispersed. In general, the share of localized industries in production is higher than in service industries indicating that knowledge creation in production industries tends to be more localized than in services. In service industries dispersion occurs more often than localization. Interestingly, especially research-intensive service industries exhibit strong cross-distance indices of dispersion.

Taking into account distances where localization occurs, reveals that industry-specific R&D is clustered over relatively long distances of about 100 km. Shifting the perspective from R&D establishments to the individual researcher even increases that relatively long distance of clustering to an interval from 260 km to 350 km. In total, our results indicate that taking into account the size of R&D establishments in terms of researchers employed there, either leads to dispersion or random distribution at distances from 0 km to 200 km. This implies that at this distance interval the number of researchers in an industry either follows the general distribution of R&D workforce or is even less than one would expect from taking the general spatial distribution of researchers as a reference.

Overall, the evidence on industry-specific spatial concentration of R&D is relatively weak. Our results indicate that localization of both R&D establishments and researchers, if it occurs, mainly is observable for production industries over relatively long distances. However, these results do not contradict the notion of R&D itself being concentrated. They rather indicate that clustering of R&D establishments or researchers at short distances is not or only weakly connected to the 3-digit industries in which innovative activities are performed.

While the analyses explicitly step back from theoretical concerns but aim to contribute to the empirical examination of industry-specific agglomeration patterns of innovative activity, they nevertheless implicitly relate to the continuing debate on

Marshall-Arrow-Romer-externalities and *Jacobs-externalities*. Assuming that the expected returns to R&D activities are taken into account when companies decide where to locate their R&D, as for example demonstrated in Duranton and Overman (2005), thus knowledge potential in space hints to anticipated knowledge spillover mechanisms. In the light of that reasoning, localization, as identified in *Section 4.2.3*, might be defined as industry-specific spatial specialization in R&D activities. It indicates that industries with localized R&D activities profit or expect to profit from above-average spatial proximity of their R&D activities, i.e. an R&D-related intra-industry transmission of knowledge.

Interpreting industry-specific localization as indicator for *Marshall-Arrow-Romer-externalities* suggests that they are either of minor relevance for R&D activities or occur over relatively long distances. It thus appears likely that spatial clustering of R&D establishments and researchers is only weakly connected to the industry in which the innovative activities are performed. Moreover, as we find a strong concentration of R&D activities themselves, especially in the researcher-weighted approach, R&D appears to attract R&D rather on a general than on an industry-specific level.

The results have implications for the ongoing debate on German cluster policy.[33]. Numerous cluster initiatives have been launched in Germany at both federal and state levels during the last 20 years. Evaluations of these cluster policies have revealed several positive influences. For instance, an analysis on the impact of the Leading-Edge Cluster Competition [34] on the formation of innovation networks showed a significant effect on the network structure in terms of density, centralization and geographical reach. On average, more than half of the existing linkages were either initiated or intensified by the cluster policy, leading to an increased density of the network.

However, it is crucial to know that most policies follow a definition of clusters as "geographic concentrations of interconnected companies and institutions *in a particular field*." (Porter 1998). Thus, the aim of cluster policies is to encourage the spatial agglomeration of firms and other organizations belonging to a particular sectoral or

[33] For a detailed overview of the varying implementations and effects of German cluster policies and initiatives, see EFI (2015)

[34] The Leading-Edge Cluster Competition was launched by the Federal Ministry of Education and Research in 2007 as part of the High-Tech Strategy. It addressed high-performance clusters formed by business and science.

technological field and to support cooperation primarily among those entities that are technologically close in order to generate positive network effects. However, our results indicate that spillover-caused incentives for spatial proximity to technologically related knowledge-producers – reflected by their common industry affiliation – are likely to be rather weak. The provision of various facilitating resources outside of the firm's industry specific knowledge sphere appears to be more important for the settlement of innovative activity. The implications for policy indicate that *Marshall-Arrow-Romer-externalities* – if at all – only secondarily affect localization decisions regarding R&D. Instruments aimed at stimulating R&D agglomeration need to be designed accordingly.

One immediately following question is, which combinations of industries are clustering their respective R&D activities in relation to one another. As far as it can be judged at present, this needs more thorough analysis and reflection than we can provide here without going far beyond the scope of this work. This research could moreover be continued by further thorough investigation of forces that lead to dispersion and localization of various industries. Using multivariate econometric approaches one can analyze clusters of R&D activity in Germany. One can also use as a benchmark in simulations not only the distribution of R&D activity in Germany but instead the distribution of economic activity tackling the question of whether knowledge production is concentrated or dispersed in relation to it. Answers to these further questions will bring us closer to understanding the choices made by firms in locating their R&D and, thus, lead to the development of better policies.

4.3 Firms' sustained superior job creation. Myth or reality?[35]

4.3.1 Sustained superior performance of firms

The global economic downturn that follows the Great Recession of 2007-2009 has led to a dramatic industrial decline and, with this, to prolonged unemployment in virtually all developed economies. In Europe, for example, the unemployment rate peaked at 11% in the second quarter of 2013, had slightly fallen the year after, and reached the level of 9% at the end of 2015. The most optimistic projections suggest that these negative values will persist for many years.

Not surprisingly, job creation has become a dominant theme in the policy arena worldwide. Several actions have been put in place to spur employment in existing companies and many others to offer fertile ground for the growth of new businesses. Such initiatives are aimed mainly at restoring competitiveness through innovation and productivity gains, revising the functioning of labour markets, and reducing barriers that prevent firms with growth potential from expanding (Stangler 2010, EU 2013). Most of the debate has been for long directed to a small share of the overall firm population, the so-called high-growth companies, that typically accounts for a disproportionate share of net job creation.

In the last decade, a large body of economic and management research has sought to identify the drivers of such a superior growth performance. Common practice has been to distinguish between two types of determinants. On the one hand, we find studies concerned with the identification of structural characteristics specific to the firm, such as productivity, financial constraints, innovative outcomes, management practices and organizational traits (see, among the many, Bottazzi et al. (2008), Coad and Rao (2008), Parker et al. (2010), Bloom and Van Reenen (2010), Harrison et al. (2014). On the other hand, scholars have searched for factors external to the firm that might indirectly shape its performance, such as institutional factors or characteristics of the location (see, among the many, Davidsson and Henrekson (2002), Audretsch and Dohse (2007), Acs and Mueller (2008). Although it is hard to draw "stylized facts" due the peculiarities of the samples analyzed, there is general consensus that high-growth companies tend to be in all industries (contrary to the

[35] in collaboration with Dr. Stefano Bianchini

popular belief about an overrepresentation in the high-tech sectors) and geographical areas, they are typically young but not necessarily small, and they are more innovative and productive than other firms.

Most studies produced so far have linked the occurrence of high-growth events both to macro-level and firm-specific characteristics from a static point of view, often ignoring that high growth episodes in firms are rare and most unlikely to be repeated. However, scholars, policy-makers, and practitioners have recently begun to shift their attention to longer-term growth history, putting more emphasis on "how" firm grow instead of on "how much" (McKelvie and Wiklund 2010). A substantial body of this research has focused on sustained high-growth patterns.

Different economic theories have developed explanations for persistence in superior growth performance. Contributions stem from alternative schools of thought, but despite differences in the underlying assumptions, they share a common mechanism of firm selection and growth, which is made explicit in disequilibrium models, while it is implicitly described as the convergence to the equilibrium path in equilibrium models (Jovanovic 1982, Dosi et al. 1995, Ericson and Pakes 1995, Cooley and Quadrini 2001, Asplund and Nocke 2006, Luttmer 2007). Theory predicts that an idiosyncratic shock affecting firm-specific unobserved factors leads to heterogeneous efficiency across firms; those firms with higher relative efficiency experience a reduction in prices which allows them to expand at the expenses of less efficient units. At the same time, higher profitability and sounder financial conditions grant to more productive firms the access to the resources needed to invest and fuel additional growth. The existence of growth persistence resides in the fact that either firms are assumed to choose long-run stable growth paths depending on their utility functions and on their resource and other constraints, or that inter-firm asymmetries in productivity, profitability, and financial conditions are not immediately reabsorbed, creating in turn a long lasting virtuous cycle with growth.

Complementing the economic theory, there is a long standing management literature on dynamic capabilities and resources as source of sustained superior performance. Underpinning these theories is the idea that competitive advantages are the basis of firm performance, and the presence of such advantages relies upon the possession of finest resources, routines, technological and organizational capabilities (Teece et al. 1997, Eisenhardt and Martin 2000, Teece 2007). These core competencies create value on the market, are unique, durable, and generate returns which are appropriated. The inherent firm-specific nature of capabilities as well as the way firm can adapt them to the changing environment, induce non-transitory competitive advantages which get reflected into sustained superior performance. Whilst these concepts have been mostly applied to explain profitability dynamics, they are also relevant to explain patterns of persistent growth and job creation (Dosi et al. 2001).

Despite the abundance of complementary theories, little consensus exists on the path-dependent nature of the process of high growth, not to say on the drivers enabling sustained high-growth performance. Recent contributions connote no or negative autocorrelation of high-growth rates over time (Parker et al. 2010, Hölzl 2014, Daunfeldt and Halvarsson 2015), though the magnitude can change according to the age and the size of the companies (Coad 2007, Capasso et al. 2014). Other contributions fail to detect any association between the canonical economic and financial variables and patterns of sustained high growth (Bianchini et al. 2016). As such, the growth behaviour of outperforming firms appears to be very fragile and the overall economic impact rather circumscribed to the short term. The mounting empirical evidence on the erratic and difficult to predict nature of growth rates is therefore incompatible with most theories of firm growth which have been developed over the years, and tend to support that random variation can be an important explanatory mechanism of the observed growth dynamics.

Despite the increasing availability of studies that model strategic management, organization behaviour, and corporate performance by mean of random variation, no attempts have been done in order to rule out chance in sustained superior growth performance. This is somewhat surprising since, as we have seen, the ability to create a disproportionate amount of new jobs repeatedly over time is the primary concern of policy-makers and researchers. On the one hand, if sustained higher employment growth deviated from randomness, there would be some hope for researchers to identify what micro and macro-level factors do actually spur this mode of growth. More empirical research would be needed to develop reliable and consistent answers, in turn, sounder policies could be designed to solve job crisis not only in the short term but also in the long one. On the contrary, we should resign ourselves that high-growth performance is merely a temporary phenomenon: firms create new jobs but very likely these jobs will be lost. Hence, that policies aimed at scaling up high-growth businesses in economies could be indirectly responsible for the increasing trend in firm-level volatility often advocated in the literature (Comin and Philippon 2005).

In the following we show for four European countries that there must be factors influencing firm's sustained superior job creation, however sometimes the number of sustained superior employers is not different from the one generated by a process driven purely by chance. At subsection 4.3.2 we describe the methodology and the choice of two models, than in subsection 4.3.2 we present the data and basic descriptive statistics. Subsection 4.3.3 demonstrates the results, while subsection 4.3.4 concludes with discussion of possible implications of our findings.

4.3.2 Methodology employing Markov chains

Firstly, we specify the calculation of the growth rate of employment between two subsequent years. Magnus Henrekson (2010) demonstrates different measures of firm growth in terms of employment, among which we highlight two:

- **relative growth** rate in percentage points. Given E_1 and E_2 as numbers of employees in periods 1 and 2 consecutively, the growth rate is calculated in accordance with the equation 4.5.

$$R_{gr} = (E_2 - E_1)/E_1 \tag{4.5}$$

- **absolute growth** measured by a difference in logarithms of a number of new employees in a company. Given E_1 and E_2 as numbers of employees in periods 1 and 2 consecutively, the growth rate is calculated in accordance with the equation 4.6.

$$A_{gr} = log(E_2) - log(E_1) \tag{4.6}$$

Moreover, in order to exclude influence of possible macro shocks and firms characteristics, we control for the age of the firm, it's size[36], industry specificity [37] and time dummies [38]. In particular, we estimate the least absolute deviation regression that also accounts for fat-tailed distributions of growth rates. Equation 4.7 specifies the regression and Figure 4.7 demonstrates a typical distribution of growth rates.

[36] The size of the firms is control as a dummy variable, where we see if the firm is a small and medium enterprise (SME) with less than 250 employees or not.

[37] Industries are control in a form of dummy variables specified by 2 digit NACE classification codes.

[38] Period of observation of a particular growth rate.

We than use the residuals of the regression (ε) as a measure of growth rate for a simulation model.

$$growth_i = log(age_i) + size_i + industry_i + time_i + \varepsilon_i \tag{4.7}$$

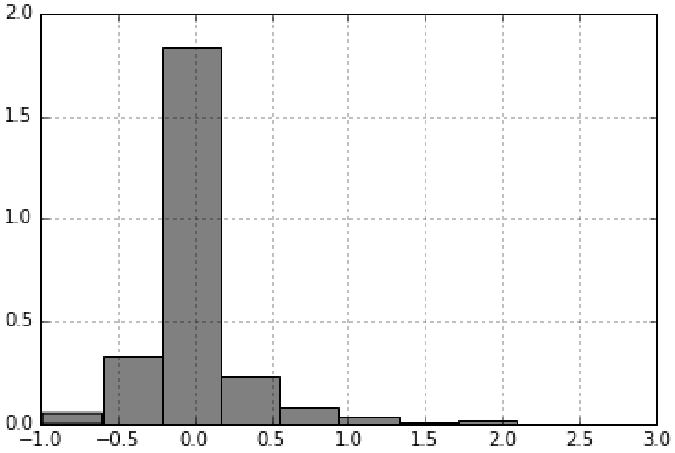


Figure 4.7: A histogram of regression residuals of the relative growth measure for a dataset including all countries.

A natural way of modeling the random process of firm growth is a law of proportionate effects (Gibrat‘s Law) which is widely discussed in the literature (Henderson et al. 2012) and can be seen in equation 4.8. This idea posits that firm’s size can be explained purely in terms of its idiosyncratic history of multiplicative growth shocks. It assumes that a firm’s performance in period t is some function of it’s performance in the period $t-1$ plus a random error that controls for other factors.

$$growth_{i,t} = f(growth_{i,t-1}) + \varepsilon_{i,t} \tag{4.8}$$

The literature provides an extensive discussion of Gibrat’s Law. Lotti et al. (2009), for example, take into account market selection and find convergence toward Gibrat-like behaviour in the long-run, implying that the growth path of surviving firms do not deviate from a random process. Coad et al. (2013) conclude that the growth pattern of a large sample of UK start-ups is largely random, and add that randomness not only can explain growth in any given year, but also that can be a good approximation for longer-term growth path over a number of years. In a similar vein, Denrell et al. (2014) conclude that randomness can often provide parsimonious explanations of several important empirical regularities in management science. However, it has to be mentioned that such an approach relies on distributional assumptions. In order to model the behavior of firms one has to assume a distribution of the error ε (i.e., Gaussian normal distribution).

In our approach we chose to abstract from these assumptions borrowing the dynamics of firm growth from the observed data. In order to model randomness, we employ a discrete-time homogeneous stochastic process characterized by a Markov chain with transition probability matrices that capture the dynamics of the process purely from the data. The Markov chain is specified by a finite set of states $S = s_1, s_2, ..., s_r$ and a transition probabilities between all states. The system starts in one state s_i and sequentially moves to another s_j with a probability p_{ij}. We model random process with first and second order Markov chains formally specified by equations 4.9 and 4.10. The first order Markov chain represents a simple hypothesis that the growth of the firm in next periods solely depends on it's current growth and is independent from other factors. However, one might object with the hypothesis that the probability that a firm will be successful in the next period, given that it is successful in the current period, might be different depending on whether it has been top performer or outsider in the previous period. This is captured by additionally modeling the random process with second order Markov chain which better reflects the idea that there is some inertia in the form of organizational learning or accumulation of experience and the process is not completely 'memoryless'.

$$\begin{aligned} Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ...X_1 = x_1) \\ = Pr(X_t = x_t | X_{t-1} = x_{t-1}) \end{aligned} \tag{4.9}$$

$$\begin{aligned} Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, ...X_1 = x_1) \\ = Pr(X_t = x_t | X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}) \end{aligned} \tag{4.10}$$

The state space of the Markov chains is determined by calculating 100 percentiles for each period and assigning every firm to its percentile/state. This rank-order statistic based on percentiles is robust and provides useful information about relative standing regardless of how a variable is distributed (Henderson et al. 2012). Having done that we can calculate probabilities of a firm to transit from a particular state based on the real transitions that we observe in the data, thus allowing us to avoid distributional assumptions.

We define a superior job creator as a firm that persistently stayed in the top 10% of employers in a given period. We complement this definition considering also top 20% which allows us to assess the robustness of our results. In order to define how many times a firm should be observed in a top ten percent of employers to be considered a superior performer, we set benchmarks for every firm grouping them by identical observed life spans. A firm that is observed for 7 periods has a higher chance of staying longer in the top than a firm with only 4 years of observations.

Benchmarks are set by simulating firms' histories 1000 times. In every simulation we let the same number of firms as in the observed panel to transit in accordance with transition matrices. Thus, if we observe 10 firms for 7 periods, there will be 10000 simulated firms with identical observation life spans. In order to be confident with $p < 0.05$ or $p < 0.01$, that an observed number of times a firm was a top employer is not a false positive, we count how many times each of the simulated firms stayed in top 10% (or 20%) and equal a benchmark for this life span to a 95th (p<0.05) or 99th (p<0.01) percentile. In doing so we abstract ourselves from a definition of persistence letting the data determine how many years a firm must be a top performer in order to be considered as superior. Thus, for every observed life span there is a benchmark.

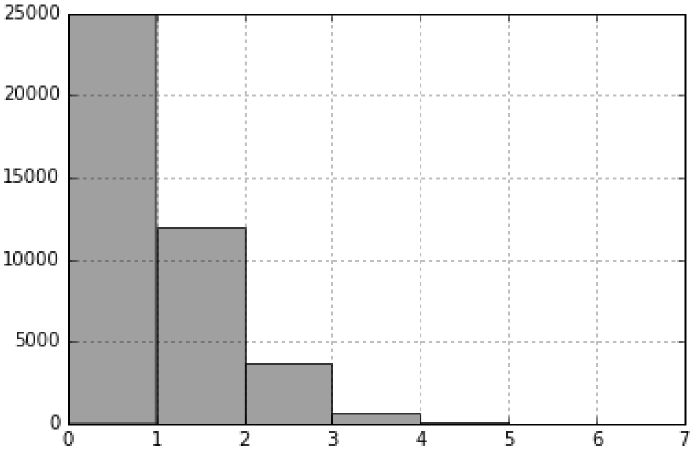


Figure 4.8: A histogram of how many times firms observed for 7 periods stayed in the top 10% of employers for the Spanish dataset measuring growth in relative terms.

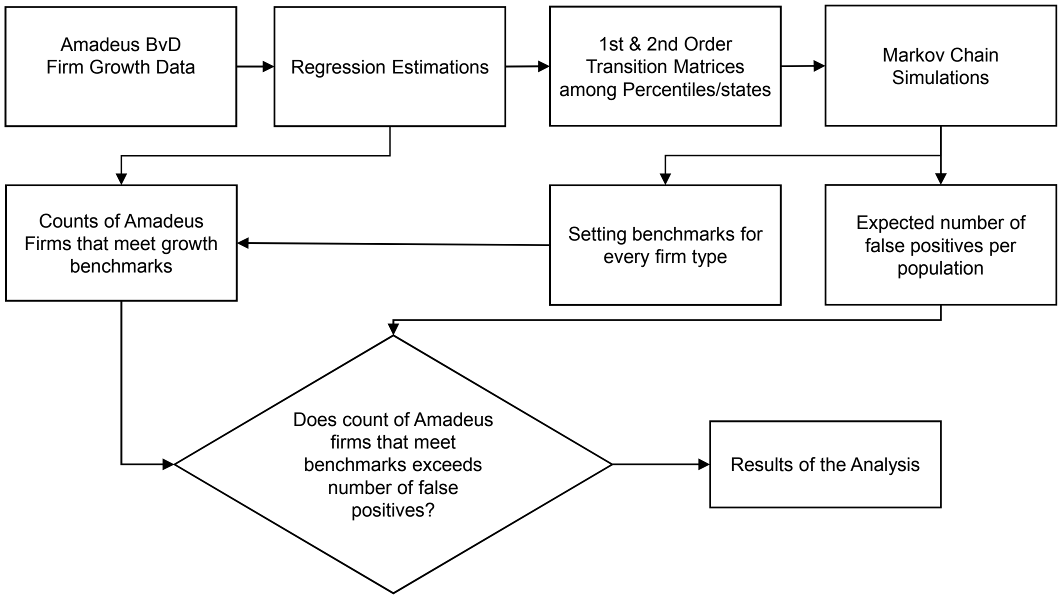


Figure 4.9: The summary of the analysis flow.

However, among those firms that met their respective benchmarks there are still some that can meet them due to chance. In order to statistically determine that the observed number of firms is higher than the expected number due to a Markov chain simulation we apply the above mentioned benchmarks to every simulated history. For every simulation we count how many times a firm stayed in the top 10% (or 20%) of employers and mark it a sustained superior performer if it met it's respective benchmark. This provides an expected number of firms that can be considered a superior job creators with a mean μ and standard deviation σ. We conclude that there is a deviation from normality and that the number of observed firms can't be explained by a random process only if this number falls outside the region of three standard deviations from the mean ($\mu \pm 3\sqrt{\sigma}$) and the distribution fulfills Kolmogorov-Smirnov test of normality. If, however, the observed number falls into that range, than we can't statistically differentiate it from an expected number of firms. Figure 4.9 summarizes all steps of our methodology.

Table 4.6: Unbalanced panel statistic. Years 2004 - 2011

	Country Data			
	UK	IT	FR	ES
Number of firms	120.690	386.123	624.878	724.848
Firms in simulation	72.064	152.075	245.146	426.816
Firms out of simulation	48.626	234.048	379.732	298.032
Balanced panel	20.246	19.696	36.332	40.549
Missing first/last year	36.743	46.212	77.133	197.922

Data

Our data set comes from the Amadeus database provided by the Bureau van Dijk and covers years from 2004 to 2011. It represents five countries: United Kingdom(UK), Italy(IT), France(FR), and Spain(ES). Table 4.6 demonstrates the number of observations for all countries. In total our analysis covered around 1.3 million firms. Since we employ a second order Markov process the firm must be observed for three subsequent years at least once imposing a restriction to the number of firms that can be taken into account in the simulation. We call such firm an active since it provides important information for the transition of firms between states in the Markov simulation. Second and third row of the Table 4.6 reflect number of inactive and active firms subsequently. Additionally, we demonstrate the sizes of balanced and almost balanced panels.

4.3.3 Mixed evidence from four European economies

The methodology explained in subsection 4.3.2 and demonstrated on the Figure 4.9 is applied to the data of every country. Firstly, we estimate transition probability matrices for our Markov chains. Figure 4.10 demonstrates the Markov chain for a relative growth measure created based on the UK dataset. One hundred nodes represent percentile states. The thickness of edges indicate probabilities of transferring from one state to another[39]. All Markov chains are ergodic (irreducible) Markov chains meaning that it is possible to go from any state to any state. This is important since it means that any firm can potentially reach any growth rate. Another

[39] Note here that this picture reflect a first order Markov chain. It is impossible to visualize in a meaningful way all transition probabilities of a second order Markov process.

interesting observation is that often there is a lot of dynamics among 1st and 2nd on the one side and 99th and 100th percentiles on the other side. This implies that in the observed data there are relatively many cases where a top employer in one year suddenly shrank in the next year and the other way around.

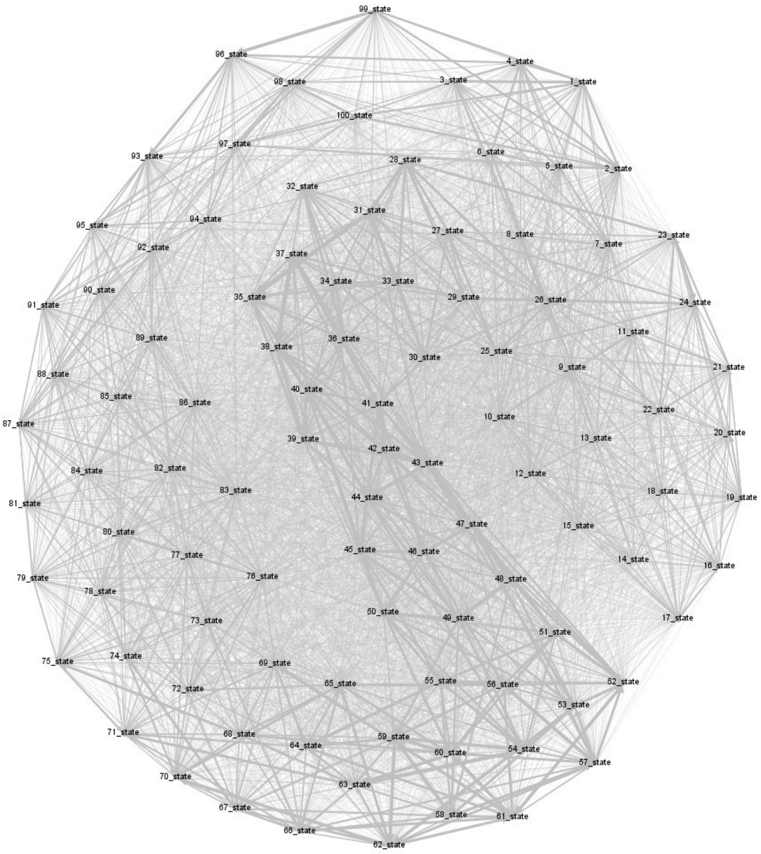


Figure 4.10: Markov Chain for a relative growth measure for UK data set. Nodes represent percentile states. Thickness of the links indicate a probability of transiting to another state as proxied by a first order Markov process.

Having calculate transition probabilities for Markov chains, we have uniformly distributed N number of firms, where N comes from a second row of the Table 4.6, and run the transition steps for the exact same number of years as in the observed data for every country.

Afterwards, we have counted how many periods each firm has stayed in the top 10% (or 20%) of job creators which results in a similar histogram as shown on the Figure 4.8 for every life span. Out of this distribution we are able to calculate how many firms are top job creators by counting ones that meet their benchmarks (see

Table 4.7: Results for the first order Markov process simulations

p		**Relative Growth Measure**			
value	Top	***United Kingdom***		***Spain***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	2287.1 (47.4)	**2887**	10502.7 (105.1)	**14968**
	20%	1950.6 (39.7)	**2422**	13128.0 (119.5)	**15427**
p<0.01	10%	412.6 (17.0)	**708**	2155.4 (50.3)	**4110**
	20%	287.9 (16.3)	**588**	2039.2 (50.9)	**2979**

p	Top	***France***		***Italy***	
value		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	7427.4 (77.3)	**10493**	3390.4 (53.5)	**4266**
	20%	10067.5 (87.6)	**12008**	5743.6 (72.3)	**7026**
p<0.01	10%	1673.8 (40.0)	**3149**	1251.0 (33.6)	**1893**
	20%	945.8 (33.0)	**1477**	694.8 (29.6)	**833**

p		**Absolute Growth Measure**			
value	Top	***United Kingdom***		***Spain***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	2311.1 (46.1)	**2909**	10502.7 (105.1)	**14968**
	20%	1959.8 (41.5)	**2494**	12961.5 (126.2)	**16060**
p<0.01	10%	426.1 (21.4)	**722**	2209.7 (41.0)	**4295**
	20%	291.4 (20.3)	**594**	1982.6 (49.5)	**3136**

p	Top	***France***		***Italy***	
value		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	7576.9 (86.6)	**10658**	4862.0 (65.5)	**6583**
	20%	10200.5 (97.7)	**12297**	5566.6 (73.8)	**7396**
p<0.01	10%	1759.2 (41.3)	**3355**	1181.5 (28.0)	**1962**
	20%	962.6 (28.6)	**1580**	674.8 (30.4)	**902**

Note: values in **bold** exceed the expected range of values.

Tables D.3, D.4, D.5, D.6 in Appendix D). This process is repeated for observed firms. Finally, having 1000 numbers reflecting simulated histories, we can calculate how many firms we expect to be sustained superior job creators and can compare this number with an observed value.

Tables 4.7 and 4.8 summarizes results for all countries for both growth measures. Each table demonstrates the results first for relative and then for absolute growth measures. Observed values highlighted in bold lie above the expected range of values.

Table 4.8: Results for the second order Markov process simulations

p value	Top	Relative Growth Measure			
		United Kingdom		***Spain***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	2091.9 (50.6)	**2349**	10352.2 (97.3)	7589
	20%	2001.6 (45.1)	**3102**	6696.0 (75.0)	**8811**
p<0.01	10%	347.2 (17.4)	**518**	995.8 (29.5)	927
	20%	311.4 (21.8)	**856**	1651.0 (42.5)	**2965**

p value	Top	***France***		***Italy***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	6787.8 (78.3)	5307	3780.6 (62.7)	2695
	20%	4328.4 (65.2)	**6394**	2963.6 (53.7)	2737
p<0.01	10%	632.1 (24.0)	693	473.1 (16.9)	341
	20%	1387.6 (30.7)	**2656**	677.5 (27.4)	**842**

p value	Top	Absolute Growth Measure			
		United Kingdom		***Spain***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	2175.8 (45.1)	**2358**	8270.9 (80.7)	7786
	20%	2359.3 (49.6)	**3125**	5624.9 (79.6)	**8961**
p<0.01	10%	362.7 (16.6)	**526**	773.0 (27.6)	**994**
	20%	200.9 (15.1)	**584**	1358.2 (31.4)	**3124**

p value	Top	***France***		***Italy***	
		expected $\mu(\sigma)$	observed #	expected $\mu(\sigma)$	observed #
p<0.05	10%	7368.6 (85.2)	5454	4209.4 (58.5)	2950
	20%	4423.5 (68.3)	**6409**	2994.9 (52.2)	2828
p<0.01	10%	793.5 (26.3)	769	443.2 (18.4)	354
	20%	1439.9 (31.9)	**2806**	697.8 (28.6)	**906**

Note: values in **bold** exceed the expected range of values.

Our conclusions are based on comprehensive analysis taking into account different growth measures, confidence bands and definitions of superiority.

As can be seen from the Table 4.7, first order Markov process is unable to generate observed number of firms regardless of the way it is calculated. In all countries, for all confidence levels and definitions of the superiority we see that the observed number of superior job creators lay above the expected range of values. With the introduction of the second order Markov process results become heterogeneous as seen in the Table 4.8. We can see that in United Kingdom the amount of observed companies is still more than expected, whereas Italy demonstrates a rather reversed pattern. In the following we closely elaborate on this issue.

4.3.4 Conclusion on sustained superior performance of firms

Job creation has become a dominant theme in the policy arena worldwide. Most of the debate has been for long directed to a small share of the overall firm population, the so-called high-growth companies, that typically accounts for a disproportionate share of net job creation. Large body of economic and management research has distinguished between two types of determinants of such a superior growth performance. These are structural characteristics specific to the firm and factors external to the firm that might indirectly influence its performance. A substantial body of this research has recently focused on sustained high-growth patterns, targeting the question of not "how much" but "how" a firm growths.

Despite the abundance of complementary theories, little consensus exists on the path-dependent nature of the process of high growth, not to say on the drivers enabling sustained high-growth performance. The underlying empirical evidence on the erratic and difficult to predict nature of growth rates is incompatible with most theories of firm growth which have been developed over the years, and tend to support the fact that random variation can be an important explanatory mechanism of the observed growth dynamics.

This section investigated whether a random variation can be an explanatory mechanism of the observed growth dynamics for four European countries. Abstracting ourselves from distributional assumptions we determine the dynamics of the model from the observed data using a first and second order Markov chain simulations. Controlling for firm age, size, industry and time dummies we find a mixed evidence of presence and absence of factors determining sustained superior growth performance.

We can not explain the observed number of firms with a simple process modeled through a first order Markov chain implying that it is obviously not enough to assume that the assumption that growth in the next period depends solely on the growth in the current period. This pattern can be seen regardless of the confidence level and definitions of superiority, which so far where top 10% and top 20%. This strongly indicates for a presence of drivers enabling sustained high-growth performance in the economy. Economic theories explain such behavior with, for example, an idiosyncratic shock that helps those firms with higher relative efficiency experience a reduction in prices which allows them to expand at the expenses of less efficient units. Together with higher profitability and sounder financial conditions more productive firms access to the resources needed to invest and fuel additional growth. In accordance with managerial literature this drivers might be firm's dynamic capabilities and resources that create value on the market, are unique, durable, and generate returns which are appropriated, and induce competitive advantages which get reflected into sustained superior performance. Accumulation of these capabilities overtime allows firms to build various routines that help them to grow.

A model using second order Markov chain captures the idea that a growth process involves some kind of 'memory' and organizations 'learn' how to grow or fail. The obtained results demonstrate heterogeneity across countries considering that the definition of superiority as being in the top 10% of job creators. Firm dynamics in the United Kingdom can't be explained with a model driven by chance. There are many more sustained performers as could be expected by a homogeneous second order Markov process for different confidence levels. A similar result is obtained for Spain which mostly shows a presence of certain factors except for one case. In contrast, data from Italy demonstrate that we can not rule out chance as an explanatory mechanism for firm behavior. French data also point out to the similar direction although being contingent on the way we measure growth process and confidence level we set. These results suggest that we should resign ourselves that high-growth performance is merely a temporary phenomenon: firms create new jobs but very likely these jobs will be lost. Hence, that policies aimed at scaling up high-growth businesses in economies could be indirectly responsible for the increasing trend in firm-level volatility often advocated in the literature (Comin and Philippon 2005). However, these results are not supported if we take into account companies that stayed not only within top 10% but within top 20% where we observe same results as for the model with first order Markov chain.

Altogether this research grants encouragements to the aforementioned economic and management theories seeking for factors of persistent job creation. It also provides a positive sign to policy makers indicating that if such factors exist they could be targeted by specific policies spurring employment. We also indirectly point out where to look for these factors. Setting benchmarks of how long a firm should grow in terms of employment in order to rule out chance, we provide evidence of which firms to focus on, while looking for possible dynamic capabilities, resources and other internal factors.

4.4 Conclusion

We started this chapter by generalizing the problem of emerging technology detection to the identification of patterns in micro data. The approach that is chosen compares the observable reality to another reality generated using simulations guided by random processes. This allows to identify statistically significant deviations pointing to the existence of forces creating those patterns. There is an advantage in comparison to a classical econometric regression approach. A regression can point to the influence of a particular factor on the dependent variable. However, it is almost impossible to prove with the regression that a process can be a byproduct of chance. Consider the case when none of the regression variable coefficients is significant. Given that enough statistical robustness checks and tricks has been applied, the only conclusion that can be made here is that these variables have no influence on the dependent one, but this does not mean that there exists no other variable that has an influence. Thus, methods of comparison to a random benchmark nicely complement classical econometrics methods. Where the former can point out on existence of forces the latter can show what are those forces.

Two application cases used in this chapter demonstrate the power of these methodologies. Firstly, we demonstrate a comprehensive assessment of German R&D allocation in private sector. A method based on distance approach and simulations allows to claim that some industries including service robotics tend to cluster and this clustering is significantly different from an average allocation of R&D. As noted in subsection 4.2.5 this result spurs some discussion with respect to *Marshall-Arrow-Romer* or *Jacobs* knowledge spillovers and leads to implications for R&D development and clustering policy. These is in a line with major literature strands, however, may be one of the most important lessons here is that there is something worth investigation going in those industries that distinguishes them from average and this type of techniques allow us to capture and detect this phenomena.

In a second application case the methodology applied was a little more sophisticated with two models of firm growth based on the Markov chain property. Here important to note is that the probabilities that drive simulations are derived from the data.

Such "let the data speak" approach coupled with exhaustive robustness tests allows to make strong conclusions with respect to the strange patterns that are observed for UK data for example where the amount of top sustained performers can not be explained by a random process. This similar to the case with clustering points to forces in our social system that lead to this behavior. Again important message is that the technique allows to detect patterns in micro data. The investigation of the mechanisms that created those patterns is a subject of another study with, perhaps, another methodology.

Chapter 5

Conclusion and contributions

Technological change brought an immense growth in welfare and quality of life in the last century. A constant flow of innovations and the appearance of new technologies are at the heart of this process. This thesis elaborates on the mechanisms of the emergence of a generally applicable knowledge - general purpose technologies - in the form of a theoretical model. It is shown how a technology can become very pervasive and emerge as a result of forces of economic agents. While the mainstream economics literature uses the concept of aggregate production function treating knowledge as a factor of production that can be accumulated, the developed model takes alternative view emphasizing heterogeneity of knowledge pieces reflecting the difference between technologies in economic structures. Abstracting the concept of knowledge as a network of interconnected technologies and simulating its discovery through the actions of agents the model demonstrates the influence of four factors on the emergence of general purpose technologies. This process of technological formation transforms into inclusion of a single technology in as many goods in the simulated economy as possible.

First socio-economic factor, influencing the process of emergence, is knowledge diffusion, given the famous public good property of knowledge (Arrow 1962) and the resulting possibility to create "complementarities among trajectories" (Dosi 1982, p. 154). The results demonstrate that this factor was a key prerequisite for the emergence of a GPT, both in terms of being used in many distinct products as well as spreading among economic agents doing R&D. Once discovered, the knowledge spills over benefiting most those technologies having multiple potential applications in combination with other intermediates in the production of different final goods. This result found support in practical studies, e.g., software security industry, highlighting the importance of external knowledge exploitation for the production of GPTs (Gambardella and Giarratana 2013). The extent of this effect, however, is contingent on the second factor considered - the exact network structure. Complex interrelationships between technologies can result in some technological links being present in numerous products or very few only. This structural property of knowledge is measured through an introduced multiplicity index reflecting the share of combinations in favor and against a potential GPT.

Literature demonstrated that the innovation process could be seen as a search in complex technology spaces "shrouded in uncertainty" (Silverberg and Verspagen 2005, p. 226) and was characterized by strong path dependence (Nelson and Winter 1982). Thus, another factor considered in the model was the choice over technological trajectories to follow while conducting R&D activities. Given the presence of knowledge spillovers, concentrating on technological trajectories with more accumulated knowledge (coordination of R&D efforts) also favors GPT, assuming a constant size of the knowledge base in the economy. However, once the technology network was modeled as a graph, growing over time, where agents become aware of new technological combinations through inventing simpler products, this positive effect of coordination transformed into an inverted U-shape form, illustrating the famous exploitation vs. exploration trade-off. Thus, it was beneficial for the knowledge discovery process in general and the GPT emergence in particular that society started favoring a certain product development after a sufficient knowledge had been accumulated as it was the case, e.g., for nuclear power plants in the 1950s (Cowan 1990) or renewable energy generation in the last two decades (Herrmann and Savin 2016). In both cases, the policy maker was providing large subsidies to discover a product with certain characteristics, while actual choice among different technological trajectories was left to innovating firms. Finally, the negative influence of frequently changing demand side was demonstrated, indicating that a society that swaps its vector of technological development too often may not benefit from general purpose knowledge.

This work depicted that the GPT formation should not necessarily be treated as a ‘black box’ where GPT comes from the outside of economic system, but could be produced (or not) by the forces within the system itself helping us to learn how to foster its emergence. Despite its many simplifying assumptions, the model successfully reproduced a wide range of stylised facts such as S-shaped curve of technology adoption, temporal clustering of innovations in time, lock in effects, as well as many structural features of the empirical product graphs (Hidalgo et al. 2007) and graphs constructed based on networks of relatedness between technological IPC classes (Boschma et al. 2014).

Though one shall be careful in drawing policy implications from the present sufficiently abstract model, some directions of thought can be outlined. It was argued, that similar to firms in the organisational theory (see, e.g., (Sidhu et al. 2007)), individual firms and whole countries should apply more differentiated technological policy, depending on their stage of development. In the ‘path-following’ catching-up process (Lee and Lim 2001) countries, aiming to discover certain product types in the knowledge base, where most of technological trajectories are known from experience of advanced economies find exploitative strategy (knowledge depth) more attractive. In contrast, if the economy is currently at the technological frontier, seeking to identify the next GPT (become ’path-creator’), it shall put more focus on exploration of new opportunities and provide incentives for sufficient knowledge breadth. For the same reason, policy maker shall avoid supporting any specific product need before the economic agents accumulate enough information on alternative production capabilities to satisfy that same economic need and payoffs to adoption of the respective technologies. Otherwise, the economy risks to be locked-in to inferior technologies due to a random choice of the technological trajectory and the increasing returns to adoption described by Arthur (1989).

Complementing theoretical work in Chapter 2 in Chapter 3 of this thesis a potential general purpose technology of our time - robotics was taken into account. Being considered in a cluster of technologies together with artificial intelligence and big data, it is seen as a potential driver of the fourth industrial revolution, that can alter modern production chains and organizational routines as well as global leadership. Robotics has all necessary characteristics of a general purpose or a structural technology. Due to its broad application it has a pervasive character entering many downstream products. It shows a significant technological dynamism demonstrating in recent decades, an increase in the number of applications (IFR 2016) and it exhibits innovation complementarity inducing significant improvements in downstream sectors.

For all those reasons a fast developing service branch of robotics was chosen as an emerging technology for an empirical study. The Chapter 3 covered the new methodology developed to detect this emerging GPT within a patent database, overcoming the initial lack of common knowledge, standards, and specifications as well as an absence of a widely agreed-upon definition of emerging technologies (Halaweh 2013). Given all these uncertainties a multiple step method was developed. Firstly, a core set of robotics patents was created using a well-established definition of robotics within a patent database based on an IPC class 'B25J'. Secondly, experts separated a sample of service robotics patents, which served as a training set for a support vector machine that was trained to classify patents. The resulting model was able to classify patents with an f1 score of 83%, allowing to retrieve patents in service robotics for further analysis.

The application of the machine learning allows avoiding human introduced bias. The experts did not choose which terms and keywords should be added to or excluded from the primal search, limiting the typical lexical bias towards preferred subfields. The developed method avoided a major drawback of citational methods, which circled around a core data set and relied on future works explicitly referring to this prior art. This is inapplicable given that citations in patents are generally rare for young emerging technologies. The procedure additionally offered strong portability and can easily be applied to scientific publications or other textual databases. Moreover, the developed step-by-step classification method can be applied to any emerging technology and not only those, that arise as an initially small subset consisting of niche applications such as service robotics out of robotics.

Taking a broader perspective of emergence and its detection Chapter 4 of the current research focused on macro pattern detection techniques in micro (firm and establishment level) data. These efforts target the question of how not to be misled by chance. In order to differentiate between a statistically significant pattern and a pattern that might emerge by chance, benchmarks were generated, using simulations driven by random processes, which allowed to reveal deviations from normality in two applications to the real-world data.

The first application analyzed the technological change and innovations from a geographical perspective, applying the "dartboard approach" (Duranton and Overman 2008) to the establishment level R&D data and service robotics patent applicants in Germany. It is shown that service robotics knowledge production is significantly clustered in southern regions of Germany. On top of that, using the data from a nationwide R&D survey the analysis of the industry location patterns on a 3-digit level (NACE) revealed that 40.8% of industries deviate significantly from random spatial location patterns. In general, knowledge creation in production industries tends to be more localized, than in services, where the dispersion occurs more often than localization. Interestingly, especially research-intensive service industries exhibited strong cross-distance indices of dispersion. Overall, the evidence on industry-specific spatial concentration of R&D was found to be relatively weak. The results indicate that localization of both R&D establishments and researchers was mainly observable for production industries over relatively long distances. However, these results do not contradict the notion of R&D itself being concentrated, but rather indicate that clustering of R&D establishments or researchers at short distances is not or only weakly connected to the 3-digit industries in which innovative activities are performed.

The second application uses simulations based on Markov chain property and demonstrated whether randomness could be ruled out when observing sustained superior job creation in Spain, United Kingdom, France and Italy. It was shown, that the observed number of firms could not be explained with a simple process modeled through a first order Markov chain. The inconsistency of the assumption was demonstrated that the employment growth of the firm tomorrow depends solely on the growth today. This pattern could be seen regardless of the confidence level and definitions of superiority (top 10 % and top 20 % levels considered in this work), strongly indicating for a presence of drivers enabling sustained high-growth performance in the analyzed economies. Economic theories explain such a behavior with an idiosyncratic shock, that helps those firms with higher relative efficiency experience a reduction in prices, allowing them to expand at the expenses of less efficient units. Together with higher profitability and sounder financial conditions more productive firms access to the resources needed to invest and fuel additional growth. In accordance with managerial literature, this drivers might be firm's dynamic capabilities and resources that are unique, durable, create value on the market, and generate returns which are appropriated, inducing competitive advantages which get reflected into sustained superior performance. Accumulation of these capabilities overtime allows firms to build various routines that help them to grow. Altogether the research grants encouragements to the economic and management theories seeking for factors of persistent high-growth performance. It also provides a positive sign to policy-makers indicating that if such factors exist they could be targeted by specific policies spurring employment. These methods of comparison to a random benchmark nicely complement classical econometrics methods. Where the former can point out on existence of forces, the latter can investigate the mechanisms that created those patterns, which is a subject of another study.

Bibliography

Abernathy, W. and Clark, C.: 1985, Innovation: mapping the winds of creative destruction, *Research Policy* **14**, 3–22.

Acs, Z. J. and Mueller, P.: 2008, Employment effects of business dynamics: Mice, gazelles and elephants, *Small Business Economics* **30**, 85–100.

Aghion, P. and Howitt, P.: 1992, A model of growth through creative destruction, *Econometrica* **60**(2), 323–51.

Aghion, P. and Howitt, P.: 1998, On the Macroeconomic Effects of Major Technological Change, *in* E. Helpman (ed.), *General Purpose Technologies and Economic Growth*, MIT Press, Cambridge, Massachussets, pp. 121–144.

Akcay, S.: 2011, Causality relationship between total R&D investment and economic growth: Evidence from United States, *The Journal of Faculty of Economics and Administrative Sciences* **16**(1), 79–92.

Albert, J., Casanoca, M. and Orts, V.: 2012, Spatial location patterns of Spanish manufacturing firms, *Papers in Regional Science* **91**(1), 107–136.

Ali, S. and Smith-Miles, K. A.: 2006, A Meta-Learning Approach to Automatic Kernel Selection for Support Vector Machines, *Neurocomputing* **70**(123), 173–186. Neural Networks Selected Papers from the 7th Brazilian Symposium on Neural Networks (SBRN 04), 7th Brazilian Symposium on Neural Networks.

Anderson, P. and Tushman, L.: 1990, Technological discontinuities and dominant designs: a cyclical model of technological change., *Administrative Science Quartely* **35**, 604–633.

Arora, S. K., Porter, A. L., Youtie, J. and Shapira, P.: 2013, Capturing new Developments in an Emerging Technology: An Updated Search Strategy for Identifying Nanotechnology Research Outputs, *Scientometrics* **95**, 351–370.

Arora, S. K., Youtie, J., Carley, S., Porter, A. L. and Shapira, P.: 2014, Measuring the Development of a Common Scientific Lexicon in Nanotechnology, *Journal of Nanoparticle Research* **16:2194**, 1–11.

Arrow, K.: 1962, Economic welfare and the allocation of resources for invention, *in* R. Nelson (ed.), *The Rate and Direction of Inventive Activity*, Princeton University Press, Princeteon, NJ, pp. 609–626.

Arthur, B.: 2015, *Complexity and the Economy*, Oxford University Press, NY.

Arthur, W. B.: 1989, Competing technologies, increasing returns, and lock-in by historical events, *The Economic Journal* **99**(394), 116–131.

Arthur, W. and Polak, W.: 2006, The evolution of technology within a simple computer model, *Complexity* **11**(5), 23–31.

Asplund, M. and Nocke, V.: 2006, Firm turnover in imperfectly competitive markets, *Review of Economic Studies* **73**(2), 295–327.

Atkinson, A. B. and Stiglitz, J. E.: 1969, A new view of technological change, *The Economic Journal* **79**(315), 573–578.

Audretsch, D. B. and Dohse, D.: 2007, Location: A neglected determinant of firm growth, *Review of World Economics* **143**(1), 79–107.

Audretsch, D. and Feldman, M.: 1996, Innovative clusters and the industry life cycle, *Review of Industrial Organisation* **11**, 253–273.

Autor, D. H., Levy, F. and Murnane, R. J.: 2003, The skill content of recent technological change: An empirical exploration, *The Quarterly Journal of Economics* **118**(4), 1279–1333.

Barlet, M., Briant, A. and Crusson, L.: 2013, Location patterns of service industries in France: A distance-based approach, *Regional Science and Urban Economics* **43**(2), 338–351.

Bassecoulard, E., Lelu, A. and Zitt, M.: 2007, Mapping Nanosciences by Citation Flows: A Preliminary Analysis, *Scientometrics* **70**, 859–880.

Baumann, J. and Kritikos, A.: 2016, The link between R&D, innovation and productivity: Are micro firms different?, *Technical Report Discussion Paper No. 9734*, Institute for the Study of Labor (IZA), Bonn.

Beaudry, C. and Schiffauerova, A.: 2009, Who's right, Marshall or Jacobs? The localization versus urbanization debate, *Research Policy* **38**, 318–337.

Bettencourt, L., Lobo, J. and Strumsky, D.: 2007, Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size, *Research Policy* **36**, 107–120.

Bianchini, S., Bottazzi, G. and Tamagni, F.: 2016, What does (not) characterize persistent corporate high-growth?, *Small Business Economics* pp. 1–24.

Bible: 1999, *The Holy Bible, King James Version*, New York: American Bible Society.

Bloom, N. and Van Reenen, J.: 2010, Why do management practices differ across firms and countries?, *The Journal of Economic Perspectives* **24**(1), 203–224.

Boschma, R., Balland, P. and Kogler, D.: 2014, Relatedness and technological change in cities: the rise and fall of technological knowledge in US metropolitan areas from 1981 to 2010, *Industrial and Corporate Change* **24**(1), 223–250.

Boser, B., Guyon, I. and Vapnik, V. (eds): 1992, *A Training Algorithm for Optimal Margin Classifiers*, Proceedings of the Fifth Annual Workshop on Computational Learning Theory – COLT '92, p. 144.

Bottazzi, G., Secchi, A. and Tamagni, F.: 2008, Productivity, profitability and financial performance, *Industrial and Corporate Change* **17**(4), 711–751.

Bresnahan, T.: 2012, Generality, recombination and re-use, *in* J. Lerner and S. Stern (eds), *The Rate and Direction of Economic Activity Revised*, University of Chicago Press, pp. 611–656.

Bresnahan, T. F.: 2010, General Purpose Technologies, *in* B. Hall and N. Rosenberg (eds), *Handbook of Economics of Innovation*, Vol. 2, Elsevier, pp. 763–791.

Bresnahan, T. F. and Trajtenberg, M.: 1995, General Purpose Technologies: 'Engines of Growth'?, *Journal of Econometrics* **65**, 83–108.

Bresnahan, T. F. and Yin, P.-L.: 2010, Reallocating Innovative Ressources around Growth Bottlenecks, *Industrial and Corporate Change* **19**(5), 1589–1627.

Briant, A., Combes, P.-P. and Lafourcade, M.: 2010, Dots to boxes: Do the size and shape of spatial units jeopardize economic geography estimations?, *Journal of Urban Economics* **67**, 287–302.

Brynjolfsson, E.: 1993, The productivity paradox of information technology: Review and assessment, *Communications of the ACM* .

Brynjolfsson, E. and McAfee, A.: 2011, *Race Against The Machine: How The Digital Revolution Is Accelerating Innovation, Driving Productivity, and Irreversibly Transforming Employment and The Economy*, Digital Frontier Press.

Bryson, J., Daniels, P. and Warf, B.: 2004, *Service Worlds – People, Organisations, Technologies*, Routledge, London.

Burges, C. J. C.: 1998, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Buzard, K. and Carlino, G.: 2009, The geography of research and development activity in the U.S., *Working Papers – Research Department, Federal Reserve Bank of Philadelphia* **Working Paper No. 09-16**.

Cantner, U. and Vannuccini, S.: 2012, A New View of General Purpose Technologies, *Jena Economic Research Papers* **2012–054**, 1–20.

Capasso, M., Cefis, E. and Frenken, K.: 2014, On the existence of persistently outperforming firms, *Industrial and Corporate Change* **23**(4), 997–1036.

Carlaw, K. and Lipsey, R.: 2006, GPT-driven, endogenous growth, *The Economic Journal* **116**(508), 155–174.

Carlaw, K. and Lipsey, R.: 2011, Sustained endogenous growth driven by structured and evolving general purpose technologies, *Journal of Evolutionary Economics* **21**, 563 – 593.

Carlino, G., Hunt, R., Carr, J. and Smith, T.: 2012, The agglomeration of R&D labs, *Working Papers – Research Department, Federal Reserve Bank of Philadelphia* **Working Paper No.12-22**.

Coad, A.: 2007, A closer look at serial growth rate correlation, *Review of Industrial Organization* **31**, 69–82.

Coad, A., Frankish, J., Roberts, R. G. and Storey, D. J.: 2013, Growth paths and survival chances: An application of gambler's ruin theory, *Journal of Business Venturing* **28**(5), 615–632.

Coad, A. and Rao, R.: 2008, Innovation and firm growth in high-tech sectors: A quantile regression approach, *Research policy* **37**(4), 633–648.

Comin, D. and Philippon, T.: 2005, The rise in firm-level volatility: Causes and consequences, *NBER macroeconomics annual* **20**, 167–201.

Cooley, T. F. and Quadrini, V.: 2001, Financial markets and firm dynamics, *American Economic Review* **91**(5), 1286–1310.

Cortes, C. and Vapnik, V.: 1995, Support-Vector Networks, *Machine Learning* **20**(3), 273–297.

Cowan, R.: 1990, Nuclear power reactors: A study in technological lock-in, *Journal of Economic History* **50**(3), 541–567.

Cowan, R. and Jonard, N.: 2007, Structural holes, innovation and the distribution of ideas, *J Econ Interac Coord* **2**, 93–110.

Crafts, N.: 2004, Steam as a General Purpose Technology: A Growth Accounting Perspective, *Economic Journal* **114**(495), 338–351.

Cunningham, P. and Delany, S. J.: 2007, k-Nearest Neighbour Classifiers, *Technical Report*, Dublin Institute of Technology.

Daunfeldt, S.-O. and Halvarsson, D.: 2015, Are high-growth firms one-hit wonders? evidence from sweden, *Small Business Economics* **44**(2), 361–383.

Davidsson, P. and Henrekson, M.: 2002, Determinants of the prevalance of start-ups and high-growth firms, *Small Business Economics* **19**(2), 81–104.

Denrell, J., Fang, C. and Liu, C.: 2014, Perspective—chance explanations in the management sciences, *Organization Science* **26**(3), 923–940.

Dosi, G.: 1982, Technological paradigms and technological trajectories: A suggested interpretation of the determinants and directions of technical change, *Research Policy* **11**(3), 147–162.

Dosi, G.: 1988a, The nature of the innovative process, *in* G. Dosi, C. Freeman, R. Nelson, G. Silverberg and L. Soete (eds), *Technical Change and Economic Theory*, Pinter London, pp. 221–238.

Dosi, G.: 1988b, Sources, procedures, and microeconomic effects of innovation, *Journal of Economic Literature* **26**(3), 1120–1171.

Dosi, G., Marsili, O., Orsenigo, L. and Salvatore, R.: 1995, Learning, market selection and the evolution of industrial structures, *Small Business Economics* **7**(6), 411–436.

Dosi, G., Nelson, R. and Winter, S.: 2001, *The nature and dynamics of organizational capabilities*, Oxford Scholarship Online, Oxford.

Duranton, G. and Overman, H.: 2005, Testing for localization using micro-geographic data, *Review of Economic Studies* **72**, 1077–1106.

Duranton, G. and Overman, H.: 2008, Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data, *Journal of Regional Science* **48**(1), 213–243.

EFI: 2015, Report on research, innovation and technological performance in germany 2015, EFI - Expert Commision on Research and Innovation, Berlin: EFI.

Eisenhardt, K. M. and Martin, J. A.: 2000, Dynamic capabilities: what are they?, *Strategic Management Journal* pp. 1105–1121.

Ellison, G. and Glaeser, E.: 1997, Geographic concentration in US manufacturing industries: A dartboard approach, *Journal of Political Economy* **105**, 889–927.

Ellison, N., Lampe, C., Steinfield, C. and Vitak, J.: 2010, With a little help from my friends: How social network sites affect social capital processes, *in* Z. Papacharissi (ed.), *The networked self: Identity, community, and culture on social network sites*, Routledge, New York, pp. 124–145.

Erdi, P., Makovi, K., Smomogyvári, Z., Strandburg, K., Tobochnik, J., Volf, P. and Zalángi, L.: 2013, Prediction of Emerging Technologies Based on Analysis of the US Patent Citation Network, *Scientometrics* **95**, 225–242.

Ericson, R. and Pakes, A.: 1995, Markov-perfect industry dynamics: A framework for empirical work, *Review of Economic Studies* **62**(1), 53–82.

EU: 2013, Innovation union competitiveness report, *Commission Staff Working Document* .

Fischer, M., Scherngell, T. and Jansenberger, E.: 2009, Geographic Localisation of Knowledge Spillovers: Evidence from High-Tech Patent Citations in Europe, *Annals of Regional Science* **43**, 839–858.

Ford, M.: 2016, *Rise of the Robots: Technology and the Threat of a Jobless Future*, Basic Books.

Frey, C. B. and Osborne, M. A.: 2013, *The Future of Employment: How susceptible are jobs to computerization?*, Oxford University Programme on the Impacts of Future Technology.

Gambardella, A. and Giarratana, M.: 2013, General technological capabilities, product market fragmentation, and markets for technology, *Research Policy* **42**, 315–325.

Garfield, E.: 1967, Primordial Concepts, Citation Indexing and Historio-Bibliography, *Journal of Library History* **2**, 235–249.

Gehrke, B., Frietsch, R., Neuhäusler, P. and Rammer, C.: 2013, Re-definition of research-intensive industries and goods – NIW/ISI/ZEW-Lists 2012, *Studien zum deutschen Innovationssystem 8-2013*, Expertenkommission Forschung und Innovation (EFI).

Gehrke, B., Rammer, C., Frietsch, R., Neuhäusler, P. and Leidmann, M.: 2010, Listen wissens- und technologieintensiver Güter und Wirtschaftszweige – Zwischenbericht zu den NIW/ISI/ZEW-Listen 2010/2011, *Studien zum deutschen Innovationssystem 19-2010.* Expertenkommission Forschung und Innovation (EFI).

Gilovich, T., Vallone, R. and Tversky, A.: 1985, The hot hand in basketball: On the misperception of random sequences, *Cognitive Psychology* **17**, 285–314.

Glaeser, E., Kallal, H., Scheinkman, J. and Shleifer, A.: 1992, Growth in cities, *Journal of Political Economy* **100**, 1126 –1152.

Glaeser, E. L.: 1999, Learning in cities, *Journal of Urban Economics* **46**(2), 254–277.

Gordon, R. J.: 2016, *The Rise and Fall of American Growth*, Princeton University Press.

Graetz, G. and Michaels, G.: 2015, Robots at work, *Center for Economic Peformance Discussion Paper* .

Griliches, Z.: 1990, Patent Statistics as Economic Indicators: A Survey, *Journal of Economic Literature* **28**, 1661–1707.

Grossman, G. M. and Helpman, E.: 1991, Trade, knowledge spillovers, and growth, *European Economic Review* **35**(2-3), 517–526.

Guyon, I., Boser, B. and Vapnik, V.: 1993, Automatic Capacity Tuning of Very Large VC-dimension Classifiers, *Advances in Neural Information Processing Systems*, Morgan Kaufmann, pp. 147–155.

Halaweh, M.: 2013, Emerging Technology: What is it?, *Journal of Technology Management and Innovation* **8**(3), 108–115.

Hall, B. H., Jaffe, A. and Trajtenberg, M.: 2005, Market Value and Patent Citations, *RAND Journal of Economics* **36**(1), 16–38.

Harrison, R., Jaumandreu, J., Mairesse, J. and Peters, B.: 2014, Does innovation stimulate employment? a firm-level analysis using comparable micro-data from four european countries, *International Journal of Industrial Organization* **35**, 29–43.

Hausmann, R. and Hidalgo, C. A.: 2011, The network structure of economic output, *Journal of Economic Growth* **16**(4), 309–342.

Helpman, E. (ed.): 1998, *General Purpose Technologies and Economic Growth*, The MIT Press, Cambridge, MA.

Henderson, A. D., Raynor, M. E. and Ahmed, M.: 2012, How long must a firm be great to rule out chance? benchmarking sustained superior performance without being fooled by randomness, *Strategic Management Journal* **33**, 387–406.

Henderson, J., Kuncoro, A. and Turner, M.: 1992, Industrial development in cities, *Journal of Political Economy* **103**(5), 1067–1090.

Henderson, R. and Clark, K.: 1990, Architectural innovation, *Administrative Science Quarterly* **35**(1), 9–30.

Herrmann, J. and Savin, I.: 2016, Optimal policy identification: Insights from the German electricity market, *Technical Report 004*, Jena Economic Research Papers.

Hidalgo, C. A., Klinger, B., Barabási, A.-L. and Hausmann, R.: 2007, The product space conditions the development of nations, *Science* **317**(5837), 482–487.

Hidalgo, C. and Hausmann, R.: 2009, The building blocks of economic complexity, *Proceedings of the National Academy of Sciences of the United States of America* **106**(26), 10570–10575.

Holland, J. H.: 1995, *Hidden Order: How Adaptation Builds Complexity*, Addison-Wesley, Michigan.

Hölzl, W.: 2014, Persistence, survival, and growth: a closer look at 20 years of fast-growing firms in austria, *Industrial and Corporate Change* **23**(1), 199–231.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J.: 2010, A Practical Guide to Support Vector Classification, *Technical Report*, Department of Computer Science and Information Engineering, National Taiwan University.

IFR: 2016, World robotics report 2016, *Technical report*, International Federation of Robotics.

Jacobs, J.: 1969, *The Economy of Cities*, Random House, New York.

Jaffe, A., Trajtenberg, M. and Henderson, R.: 1993, Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations, *The Quarterly Journal of Economics* **108**(3), 577–598.

Jovanovic, B.: 1982, Selection and the evolution of industry, *Econometrica* **50**(3), 649–70.

Kauffman, S.: 1995, *At Home in the Universe: The Search for the Laws of Self-Organization and Complexity*, Oxford University Press, Oxford.

Kerr, W. and Kominers, S.: 2015, Agglomerative forces and cluster shapes, *The Review of Economics and Statistics* **97**(4), 877–899.

Koh, H. and Riedel, N.: 2014, Assessing the localization pattern of German manufacturing and service industries: A distance-based approach, *Regional Studies* **48**(5), 823–843.

Kotsiantis, S. B.: 2007, Supervised Machine Learning: A Review of Classification Techniques, *Informatica* **31**, 249–268.

Krugman, P.: 1991, Increasing returns and economic geography, *Journal of Political Economy* **99**(3), 483–499.

Lee, K. and Lim, C.: 2001, Technological regimes, catching-up and leapfrogging: findings from the Korean industries, *Research Policy* **30**(3), 459–483.

Lee, S., Yoon, B. and Park, Y.: 2009, An Approach to Discovering New Technology Opportunities: Keyword-Based Patent Map Approach, *Technovation* **29**, 481–497.

Lee, W. H.: 2008, How to Identify Emerging Research Fields Using Scientometrics: An Example in the Field of Information Security, *Scientometrics* **76**(3), 503–525.

Li, Y.-R., Wang, L.-H. and Hong, C.-F.: 2009, Extracting the Significant-Rare Keywords for Patent Analysis, *Expert Systems with Applications* **36**, 5200–5204.

Lipsey, R., Carlaw, K. and Bekar, C.: 2005, *Economic Transformations*, Oxford University Press Inc., New York.

Lotti, F., Santarelli, E. and Vivarelli, M.: 2009, Defending gibrat's law as a long-run regularity, *Small Business Economics* **32**(1), 31–44.

Luttmer, E. G.: 2007, Selection, growth, and the size distribution of firms, *The Quarterly Journal of Economics* **122**(3), 1103–1144.

Magnus Henrekson, D. J.: 2010, Gazelles as job creators: a survey and interpretation of the evidence, *Small Business Economics* **35**, 227–244.

Manning, C., Raghavan, P. and Schütze, H.: 2008, Introduction to Information Retrieval. online, accessed October 15 2014.
URL: *http://www-nlp.stanford.edu/IR-book/*

March, J. G.: 1991, Exploration and exploitation in organizational learning, *Organization Science* **2**(1), 71–87.

Marshall, A.: 1920, *Principles of Economics*, Macmillan, London.

McKelvie, A. and Wiklund, J.: 2010, Advancing firm growth research: A focus on growth mode instead of growth rate, *Entrepreneurship theory and practice* **34**(2), 261–288.

Menz, N. and Ott, I.: 2011, On the role of general purpose technologies within a marshall-jacobs controversy: the case of nanotechnologies, *Technical report*, Working Paper Series in Economics.

Mernberger, M.: 2011, *Graph-Based Approaches to Protein Structure Comparison - From Local to Global Similarity*, PhD thesis, Philipps-University Marburg.

Miles, I.: 2007, Research and development (R&D) beyond manufacturing: The strange case of services R&D, *R&D Management* **37**(3), 249–268.

Mogoutov, A. and Kahane, B.: 2007, Data Search Strategy for Science and Technology Emergence: A Scalable and Evolutionary Query for Nanotechnology Tracking, *Research Policy* **36**, 893–903.

Moser, P. and Nicholas, T.: 2004, Was Electricity a General Purpose Technology? Evidence from Historical Patent Citations, *American Economic Review, Papers and Proceedings* **94**(2), 388–394.

Murata, Y., Nakajima, R., Okamoto, R. and Tamura, R.: 2014, Localized knowledge spillovers and patent citations: A distance-based approach, *Review of Economics and Statistics* **96**(5), 967–985.

Nakajima, K., Saito, Y. U. and Uesugi, I.: 2012, Measuring economic localization: Evidence from Japanese firm-level data, *Journal of the Japanese and International Economies* **26**(2), 201–220.

Nelson, R. and Winter, S.: 1982, *An Evolutionary Theory of Economic Change*, Cambridge, MA.

Noyons, E., Buter, R., Raan, A., Schmoch, U., Heinze, T., S., H. and Rangnow, R.: 2003, Mapping Excellence in Science and Technology Across Europe. Part 2: Nanoscience and Nanotechnology. Draft Report EC-PPN CT2002-0001 to the European Commission.

OECD: 2002, *Frascati Manual 2002 – Proposed Standard Practice for Surveys on Research and Experimental Development*, Paris.

Ott, I.: 2012, Service Robotics: An Emergent Technology Field at the Interface between Industry and Services, *Poiesis and Praxis* **9**(3–4), 219–229.

Ott, I., Papilloud, C. and Zuelsdorf, T.: 2009, What drives innovation? Causes of and consequences for nanotechnologies, *Managing Global Transitions* **7**(1), 5–26.

Parker, S. C., Storey, D. J. and Van Witteloostuijn, A.: 2010, What happens to gazelles? the importance of dynamic management strategy, *Small Business Economics* **35**(2), 203–226.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: 2011, Scikit-Learn: Machine Learning in Python, *Journal of Machine Learning Research* **12**, 2825–2830.

Pérez, F. and Granger, B. E.: 2007, IPython: a System for Interactive Scientific Computing, *Computing in Science and Engineering* **9**(3), 21–29.

Petsas, I.: 2003, The Dynamic Effects of General Purpose Technologies on Schumpeterian Growth, *Journal of Evolutionary Economics* **13**(5), 577–605.

Porter, A., Youtie, J. and Shapira, P.: 2008, Nanotechnology Publications and Citations by Leading Countries and Blocs, *Journal of Nanoparticle Research* **10**, 981–986.

Porter, M. E.: 1998, Clusters and the new economics of competition, *in* M. E. Porter (ed.), *On Competition*, Harvard Business School Press, Boston, pp. 197–287.

Rojas, R.: 1996, Neural Networks: A Systematic Introduction, Springer.

Romer, P.: 1990a, Endogenous Technological Change, *Journal of Political Economy* **98**(5), S71–S102. The Problem of Development: A Conference of the Institute for the Study of Free Enterprise Systems.

Romer, P.: 1990b, Endogenous Technological Change, *Journal of Political Economy* **98**(5), S71–S102.

Romer, P. M.: 1986, Increasing returns and long-run growth, *Journal of Political Economy* **94**(5), 1002–1037.

Roney, C. J. R. and Trick, L. M.: 2009, Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy, *THINKING & REASONING* **15**(2), 197–210.

Rosenberg, N.: 1976, *Perspectives on technology*, Cambridge University Press.

Rosenberg, N. and Trajtenberg, M.: 2004, A general-purpose technology at work: The corliss steam engine in the late-nineteenth-century united states, *The Journal of Economic History* **64**(01), 61–99.

Rosenthal, S. S. and Strange, W. C.: 2004, Evidence on the nature and sources of agglomeration economies, *in* J. V. Henderson and J.-F. Thisse (eds), *The Handbook of Regional and Urban Economics*, Cities and Geography edn, Vol. IV, North-Holland, Amsterdam, pp. 2119–2171.

Ruffaldi, E., Sani, E. and Bergamasco, M.: 2010, Visualizing Perspectives and Trends in Robotics Based on Patent Mining, IEEE International Conference on Robotics and Automation, Anchorage, Alaska.

Schmoch, U.: 2008, Concept of a Technology Classification for Country Comparisons, *Technical report*, World Intellectual Property Organisation.

Schumpeter, J.: 1934, *The Theory of Economic Development*, Harvard University Press.

Sedgley, N. and Elmslie, B.: 2011, Do we still need cities? Evidence on rates of innovation from count data models of Metropolitan Statistical Area patents, *American Journal of Economics and Sociology* **70**(1), 86–108.

Shea, C., Grinde, R. and Elmslie, B.: 2011, Nanotechnology as General-Purpose Technology: Empirical Evidence and Impliations, *Technology Analysis & Strategic Management* **23**(2), 175–192.

Sidhu, J., Commandeur, H. and Volberda, H.: 2007, The multifaceted nature of exploration and exploitation: Value of supply, demand, and spatial search for innovation, *Organization Science* **18**(1), 20–38.

Silverberg, G. and Lehnert, D.: 1993, Long waves and 'evolutionary chaos' in a simple Schumpeterian model of embodied technical change, *Structural Change and Economic Dynamics* **4**(1), 9–37.

Silverberg, G. and Verspagen, B.: 2003, Breaking the waves: a Poisson regression approach to Schumpeterian clustering of basic innovations, *Cambridge Journal of Economics* **27**(5), 671–693.

Silverberg, G. and Verspagen, B.: 2005, A percolation model of innovation in complex technology spaces, *Journal of Economic Dynamics & Control* **29**, 225–244.

Silverman, B. W.: 1986, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London – New York.

Simoes, A.: 2016, Observatory of economic complexity. Available at `http://atlas.media.mit.edu/en/resources/data/`.

Solow, R.: 1956, A Contribution to the Theory of Economic Growth, *Quarterly Journal of Economics* **70**(1), 65–94.

Solow, R.: 1957, Technical Change and the Aggregate Production Function, *Review of Economics and Statistics* **39**(3), 312–320.

Srinivasan, R.: 2008, Sources, Characteristics and Effects of Emerging Technologies: Research Opportunities in Innovation, *Industrial Marketing Management* **37**, 633–640.

Stahl, B.: 2011, *What Does the Future Hold? A Critical View of Emerging Information and Communication Technologies and their Social Consequences*, Vol. 356, Springer, Berlin, Heidelberg.

Stangler, D.: 2010, High-growth firms and the future of the american economy, *Ewing Marion Kauffman Foundation Paper* .

Stifterverband für die Deutsche Wissenschaft: 2015, Forschung und Entwicklung in der Wirtschaft 2013, Wissenschaftsstatisitk GmbH im Stifterverband für die Deutsche Wissenschaft, Essen.

Strohmaier, R. and Rainer, A.: 2016, Studying general purpose technologies in a multi-sector framework: The case of ICT in Denmark, *Structural Change and Economic Dynamics* **36**, 34–49.

Teece, D. J.: 2007, Explicating dynamic capabilities: the nature and microfoundations of (sustainable) enterprise performance, *Strategic management journal* **28**(13), 1319–1350.

Teece, D. J., Pisano, G. and Shuen, A.: 1997, Dynamic capabilities and strategic management, *Strategic management journal* pp. 509–533.

Thompson, P.: 2006, Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-added Citations, *The Review of Economics and Statistics* **88**(2), 383–388.

Tseng, Y.-H., Lin, C.-J. and Lin, Y.-I.: 2007, Text Mining Techniques for Patent Analysis, *Information Processing and Management* **43**, 1216–1247.

Tversky, A. and Kahneman, D.: 1971, Belief in the law of small numbers, *Psychological Bulletin* **76**, 105 – 110.

Vuijlsteke, M., Guerrieri, P. and Padoan, P. C.: 2007, Modelling ICT as a General Purpose Technology – Evaluation Models and Tools for Assessment of Innovation and Sustainable Development at the EU Level, *Collegium – Special Edition 35*, College of Europe.

Weitzman, M. L.: 1998, Recombinant growth, *The Quaterly Journal of Economics* **113**(2), 331–360.

Wolpert, D. and Macready, W.: 1997, No Free Lunch Theorems for Optimization, *IEEE Transactions on Evolutionary Computation* **1**(1), 67–82.

Yoon, B. and Park, Y.: 2004, A Text-Mining-Based Patent Network: Analytical Tool for High-Technology Trend, *Journal of High-Technology Management Research* **15**, 37–50.

Youtie, J., Iacopetta, M. and Graham, S.: 2008, Assessing the Nature of Nanotechnology: Can We Uncover an Emerging General Purpose Technology?, *Journal of Technology Transfer* **33**, 315–329.

Appendix

Appendix A

Appendix. General purpose technology as an emergent property

Table A.1: Probabilities of a product to become visible depending on the percentage of discovered, visible and invisible links in accordance with equation 2.6

Discovered (%)	Visible (%)	Invisible (%)	Probability
1.00000000	0.00000000	0.00000000	0.01741174
0.90000000	0.10000000	0.00000000	0.01391415
0.80000000	0.20000000	0.00000000	0.01224831
0.70000000	0.30000000	0.00000000	0.01198899
0.60000000	0.40000000	0.00000000	0.01307001
0.5000000	0.5000000	0.0000000	0.0157673
0.40000000	0.60000000	0.00000000	0.02076935
0.30000000	0.70000000	0.00000000	0.02935294
0.20000000	0.80000000	0.00000000	0.04370905
0.1000000	0.9000000	0.0000000	0.0675021
0.0000000	1.0000000	0.0000000	0.1068053
0.9000000	0.0000000	0.1000000	0.0178481
0.80000000	0.10000000	0.10000000	0.01589869
0.70000000	0.20000000	0.10000000	0.01517186
0.60000000	0.30000000	0.10000000	0.01548209
0.50000000	0.40000000	0.10000000	0.01690855
0.40000000	0.50000000	0.10000000	0.01981537
0.3000000	0.6000000	0.1000000	0.0249445
0.20000000	0.70000000	0.10000000	0.03360517
0.10000000	0.80000000	0.10000000	0.04800803
0.00000000	0.90000000	0.10000000	0.07182944
0.80000000	0.00000000	0.20000000	0.02266823
0.70000000	0.10000000	0.20000000	0.02165784
0.60000000	0.20000000	0.20000000	0.02150056
0.50000000	0.30000000	0.20000000	0.02215623
0.40000000	0.40000000	0.20000000	0.02379222
0.30000000	0.50000000	0.20000000	0.02682611
0.20000000	0.60000000	0.20000000	0.03203232
0.10000000	0.70000000	0.20000000	0.04073974
0.00000000	0.80000000	0.20000000	0.05517096
0.70000000	0.00000000	0.30000000	0.03310248
0.60000000	0.10000000	0.30000000	0.03266164
0.5000000	0.2000000	0.3000000	0.0328498
0.40000000	0.30000000	0.30000000	0.03371499
0.30000000	0.40000000	0.30000000	0.03547806
0.20000000	0.50000000	0.30000000	0.03858904
0.100000	0.600000	0.300000	0.043842
0.00000000	0.70000000	0.30000000	0.05257778
0.60000000	0.00000000	0.40000000	0.05181422
0.50000000	0.10000000	0.40000000	0.05171882
0.4000000	0.2000000	0.4000000	0.0521165
0.30000000	0.30000000	0.40000000	0.05310878
0.20000000	0.40000000	0.40000000	0.05494893
0.10000000	0.50000000	0.40000000	0.05810666
0.00000000	0.60000000	0.40000000	0.06338798
0.50000000	0.00000000	0.50000000	0.08357964
0.40000000	0.10000000	0.50000000	0.08369377
0.30000000	0.20000000	0.50000000	0.08421854
0.20000000	0.30000000	0.50000000	0.08528789
0.1000000	0.4000000	0.5000000	0.0871748
0.00000000	0.50000000	0.50000000	0.09036088
0.4000000	0.0000000	0.6000000	0.1365069
0.3000000	0.1000000	0.6000000	0.1367482
0.20000	0.20000	0.60000	0.13735
0.1000000	0.3000000	0.6000000	0.1384661
0.0000000	0.4000000	0.6000000	0.1403814
0.3000000	0.0000000	0.7000000	0.2241059
0.2000000	0.1000000	0.7000000	0.2244242
0.1000000	0.2000000	0.7000000	0.2250728
0.0000000	0.3000000	0.7000000	0.2262173
0.2000000	0.0000000	0.8000000	0.3687364
0.1000000	0.1000000	0.8000000	0.3691014
0.0000000	0.2000000	0.8000000	0.3697784
0.1000000	0.0000000	0.9000000	0.6073155
0.0000000	0.1000000	0.9000000	0.6077089
0.000000	0.000000	1.000000	1.000741

Appendix B

Appendix. A patent search strategy for service robotics

Table B.1: *Important robot definitions according to ISO 8373:2012*

	Definition
Robot:	Actuated mechanism programmable in two or more axes with a degree of autonomy, moving within its environment, to perform intended tasks. Note 1 to entry: A robot includes the control system and interface of the control system. Note 2 to entry: The classification of robot into industrial robot or service robot is done according to its intended application.
Autonomy:	Ability to perform intended tasks based on current state and sensing, without human intervention.
Control System:	Set of logic control and power functions which allows monitoring and control of the mechanical structure of the robot and communication with the environment (equipment and users).
Robotic Device:	Actuated mechanism fulfilling the characteristics of an industrial robot or a service robot, but lacking either the number of programmable axes or the degree of autonomy.

Table B.2: *SR application examples for personal / domestic use according to the IFR*

	Applications
Robots for domestic tasks	Robot butler, companion, assistants, humanoids Vacuuming, floor cleaning Lawn mowing Pool cleaning Window cleaning
Entertainment robots and Toy robots	Robot rides Pool cleaning Education and training
Handicap assistance and Robotized wheelchairs	Personal rehabilitation Other assistance functions
Personal transportation	
Home security and surveillance	

Table B.3: *SR application examples for professional / commercial use according to IFR*

	Applications
Field robotics	Agriculture Milking robots Forestry Mining systems Space robots
Professional cleaning	Floor cleaning Window and wall cleaning Tank, tube and pipe cleaning Hull cleaning
Inspection and maintenance systems	Facilities, Plants Tank, tubes and pipes and sewer Other inspection and maintenance systems
Construction and demolition	Nuclear demolition and dismantling Other demolition systems Construction support and maintenance Construction
Logistic systems	Courier/Mail systems Factory logistics Cargo handling, outdoor logistics Other logistics
Medical robotics	Diagnostic systems Robot assisted surgery or therapy Rehabilitation systems Other medical robots
Defense, rescue and security applications	Demining robots Fire and bomb fighting robots Surveillance/security robots Unmanned aerial and ground based vehicles
Underwater systems	Search and Rescue Applications Other
Mobile Platforms in general use	Wide variety of applications
Robot arms in general use	Wide variety of applications
Public relation robots	Hotel and restaurant robots Mobile guidance, information robots Robots in marketing
Special Purpose	Refueling robots
Customized robots	Customized applications for consumers
Humanoids	Variety of applications

Table B.4: *Exemplary extract of robot patents under consideration with respective titles, publication numbers (given by the patent authority issuing the patent), filing dates (on which the application was received), and expert classifcation decisions*

Title	Publication no.	Filing date	SR y/n?
Remote control manipulator	968525	1962-06-25	n (-1)
Folded robot	2061119	1979-10-24	n (-1)
In vivo accessories for minimally invasive robotic surgery	2002042620	2001-11-06	y (1)
Apparatus and method for non-destructive inspection of large structures	6907799	2001-11-13	y (1)
Surgical instrument	2002128661	2001-11-16	y (1)
Robotic vacuum cleaner	2003060928	2001-12-04	y (1)
A cleaning device	1230844	2002-01-21	n (-1)
Climbing robot for movement on smooth surfaces e.g. automatic cleaning of horizontal / vertical surfaces has chassis with crawler drive suspended and mounted turnable about vertical axis, to detect obstacles and prevent lifting-off	10212964	2002-03-22	y (1)
Single Cell Operation Supporting Robot	2004015055	2002-08-08	y (1)
Underwater Cleaning Robot	2007105303	2006-03-14	y (1)
Position determination for medical devices with redundant position measurement and weighting to prioritise measurements	1854425	2006-05-11	y (1)
Mobile Robot and Method of controlling the same	2007135736	2006-05-24	y (1)
Customizable Robotic System	2012061932	2011-11-14	y (1)
Positioning Apparatus for Biomedical Use	2012075571	2011-12-06	n (-1)
Apparatus and Method of Controlling Operation of Cleaner	2012086983	2011-12-19	n (-1)

Table B.5: List of the 1206 variables used in the SVM for classification: Part 1/4 of the 726 unigrams.

1a
abl
abnormal
accelerat
access
accommodat
accord
accordanc
accurat
achiev
acquir
act
action
activ
actual
actuat
adapt
adapter
addition
adhesiv
adjacent
adjust
adjustabl
adjustment
advanc
advantag
agricultural
aid
aim
air
algorithm
allow
amount
analysi
analyz
angl
angular
animal
annular
apertur
apparatus
appearanc
appli
applianc
applic
appropriat
architectur
arm
arrang
arrangement
arriv
articulat
assembl
assist
associat
attach
attachabl
attachment
auto
automat
automatic
autonomous
auxiliari
avoid
axe
axi
axial
backlash
balanc
barrier
base
basi
beam
bear
behavior
bend
bicycl
bipedal
blade
block
board
bodi
bore
bottom
box
brush
build
built
button
cabl
calculat
camera
capabl
capillari
captur
car
cardiac
carri
carriag
carrier
caus
cell
center
centr
central
chang
characteris
characteristic
characteriz
charg
chassi
check
circuit
claim
clamp
clean
cleaner
climb
clip
close
coat
code
collect
collision
column
combin
combinat
comfortabl
command
common
communic
compact
compar
compartment
complementari
complet
component
compos
compris
computer
condition
configur
configurat
confin
confirm
connect
connection
consequent
consist
constitut
construct
construction
contact
contain
container
continuous
control
controller
convention
convert
conveyor
coordinat
correspond
cost
coupl
cover
creat
crop
current
customizabl
cut
damag
data
decision
defin
degre
deliver
deliveri
deploy
depress
describ
design
desir
detachabl
detect
detection
detector
determin
determinat
deviat
devic
diagnosi

Table B.6: List of the 1206 variables used in the SVM for classification: Part 2/4 of the 726 unigrams.

differenc
difficult
digital
dimension
dimensional
dip
direct
direction
discharg
disclos
disconnect
dispens
displac
displaceabl
displacement
display
dispos
distal
distanc
dock
door
doubl
draw
drill
drive
driven
dust
dynamic
earth
easili
edg
effect
effectiv
effector
efficienc
elastic
electric
electronic
element
elongat
embodiment
emit
emitter
employ
employment
enabl
enclos
endoscop
endoscopic
energi
engag
enhanc
ensur
enter
entir
environment
environmental
equip
equipment
error
especial
essential
etc
exampl
exchang
exhaust
exist
expensiv
extend
extension
external
extract
extraction
extrem
facilitat
faciliti
factor
fasten
featur
feedback
field
fig
figur
fill
filter
finger
fit
fix
flang
flat
flexibl
floor
flow
fluid
forc
foreign
form
frame
free
freedom
frequenc
front
function
gear
generat
glove
grasp
grip
gripper
groov
ground
guid
guidanc
hand
handl
har
head
heat
held
help
hip
hold
holder
horizontal
hose
hous
human
hydraulic
identifi
imag
implement
improv
improvement
includ
incorporat
increas
independent
individual
industrial
informat
inner
input
insert
insertion
inspection
instal
installat
instruction
instrument
integrat
interaction
interconnect
interfac
interior
internal
invasiv
invention
involv
item
jet
join
joint
knee
laser
latter
lawn
layer
leg
length
lever
lift
light
limb
limit
line
linear
link
liquid
load
local
locat
lock
locomotion
log
longitudinal
loop
low
lower
machin
magnetic
main
maintain

Table B.7: List of the 1206 variables used in the SVM for classification: Part 3/4 of the 726 unigrams.

make	obtain	portion	referenc
manipulat	oper	position	region
manner	operabl	possibl	register
manoeuvr	operat	power	relat
manual	oppos	pre	relationship
manufactur	optic	precis	relativ
map	option	predefin	releas
marker	orient	predetermin	reliabl
master	orientat	preferabl	remot
material	orthogonal	preparat	remov
mean	outer	press	removal
measur	output	pressur	replac
measurement	overal	prevent	requir
mechanic	pair	procedur	resolution
mechanism	pallet	process	respect
medic	panel	processor	respectiv
medicin	parallel	produc	result
medium	part	product	retain
memori	partial	production	return
method	particular	program	rigid
micro	pass	project	ring
militari	path	propos	risk
milk	patient	propulsion	robot
mine	pattern	protectiv	robotic
minimal	payload	provid	rock
mobil	perform	proximal	rod
modal	performanc	purpos	roll
mode	period	quantiti	roller
model	peripheral	rack	rotari
modul	permit	radar	rotat
monitor	perpendicular	radial	rotatabl
motion	photograph	radio	rough
motor	pick	rail	run
mount	piec	rais	safeti
movabl	pipe	rang	sampl
move	pivot	rapid	save
movement	pivotabl	reach	scale
mow	place	reaction	screen
mower	plan	real	seal
mri	plane	realiti	section
multi	plant	realiz	sector
multipl	plastic	rear	secur
navigat	plate	receiv	select
network	platform	receiver	send
normal	play	reciprocat	sens
nozzl	plural	recognition	sensor
object	pneumatic	record	sent
obstacl	port	reduc	separat

Table B.8: List of the 1206 variables used in the SVM for classification: Part 4/4 of the 726 unigrams.

sequenc
seri
serv
servo
set
shaft
shape
shield
ship
short
signal
significant
simpl
simulat
simultaneous
singl
site
situat
size
skin
slave
sleev
smooth
sourc
sow
space
spatial
special
specifi
specific
speed
spiral
spray
spring
stabiliti
stabiliz
stabl
stage
station
stationari
steer
step
stop
storag
store
structur
subject
subsequent
substantial
substrat
subsystem
suction
suitabl
suppli
support
surfac
surgeon
surgeri
surgic
surround
sutur
switch
system
take
tank
target
task
techniqu
telepresenc
telescopic
terminal
terrain
test
therebi
therefrom
thereof
thereon
thereto
third
tight
tilt
time
tip
tissu
tool
tooth
top
torqu
torso
touch
toy
track
train
trajectori
transfer
translat
transmission
transmit
transmitter
transport
transportat
transvers
travel
treat
treatment
tube
type
typic
ultrasonic
underwater
uneven
unit
universal
unload
upper
use
user
utiliz
vacuum
valu
variabl
varieti
vehicl
velociti
vertic
vessel
video
view
virtual
visual
volum
walk
wall
wast
water
weed
weight
weld
wheel
wherebi
wherein
wide
winch
window
wire
wireless
workpiec
worn
wrist
zone

Table B.9: List of the 1206 variables used in the SVM for classification: Part 1/2 of the 370 bigrams.

1,2
1,compris
1,computer
1,connect
1,disclos
12,includ
12,provid
13,14
2,3
2,compris
2,move
3,4
3,compris
3,connect
4,5
43,connect
5,arrang
5,provid
accord,invention
actual,position
actuat,control
addition,equipment
adjust,position
adjustabl,surgeon
allow,surgeon
angl,adjust
apparatus,compris
apparatus,method
apparatus,perform
arm,coupl
arm,includ
arm,instrument
arm,join
assembl,method
automatic,clean
automatic,control
automatic,robot
autonomous,move
autonomous,robot
axe,rotat
balanc,control
base,informat
base,station
bodi,2
bodi,robot
bodi,surgic
button,allow
button,depress
button,effector
capabl,control
cardiac,procedur
chassi,frame
claim,includ
clean,horizontal
clean,method
clean,operat
clean,robot
cleaner,compris
cleaner,invention
comfortabl,position
component,provid
compris,base
compris,bodi
compris,main
compris,plural
compris,robot
compris,robotic
computer,program
connect,clamp
control,box
control,cabl
control,devic
control,input
control,joint
control,manipulat
control,method
control,movement
control,operat
control,panel
control,provid
control,resolution
control,robot
control,robotic
control,system
control,unit
controller,handl
correspond,movement
coupl,pair
degre,freedom
deliveri,system
depress,surgeon
detect,obstacl
detect,position
detection,mean
determin,position
determin,spatial
deviat,actual
devic,17
devic,compris
devic,control
devic,determin
devic,direct
devic,includ
devic,main
devic,position
devic,provid
devic,robot
devic,system
direction,drive
displacement,sensor
distanc,measur
door,10
drive,actuat
drive,devic
drive,forc
drive,ground
drive,mechanism
drive,system
drive,unit
drive,wheel
e,g
effector,control
effector,correspond
effector,handl
effector,manipulat
effector,move
effector,movement
effector,perform
element,5
endoscopic,imag
error,signal
factor,adjustabl
front,bodi
front,rear
front,robot
guid,mean
hand,surgeon
handl,controller
handl,move
handl,scale
har,1
hold,sutur
horizontal,vertic
imag,data
imag,process
implement,method
includ,base
includ,main
includ,pair
includ,step
independent,claim
industrial,robot
informat,relat
informat,sensor
informat,set
inner,surfac
input,button
input,data
instrument,coupl
instrument,effector
instrument,mount
invasiv,cardiac
invention,compris
invention,disclos
invention,propos
invention,provid
invention,relat
joint,provid
laser,emitter
leg,joint
longitudinal,direction
machin,tool
main,bodi
main,controller
manipulat,arm
manipulat,hold
master,handl
mean,14
mean,2
mean,detect
mean,receiv
measur,devic
mechanism,rotat
method,apparatus
method,autonomous
method,clean
method,control
method,invention
method,provid
method,system
method,thereof
method,use

Table B.10: List of the 1206 variables used in the SVM for classification: Part 2/2 of the 370 bigrams.

minimal,invasiv
mobil,robot
mobil,robotic
motion,control
motion,controller
motor,drive
motor,vehicl
mount,chassi
mount,robot
move,button
move,comfortabl
move,devic
move,effector
move,floor
move,robot
move,surgeon
movement,effector
movement,handl
movement,movement
movement,perform
movement,robotic
movement,typic
navigat,system
object,provid
operat,accord
operat,clamp
operat,devic
operat,operat
operat,perform
operat,power
operat,rang
operat,remot
operat,robot
operat,unit
output,signal
overal,structur
pair,master
pair,robotic
pair,surgic
path,robot
patient,s
patient,treat
perform,clean
perform,hand
perform,minimal
perform,surgic
position,base
position,compris
position,coordinat
position,determinat
position,devic
position,handl
position,informat
position,robot
position,robotic
position,system
power,sourc
predetermin,position
predetermin,time
procedur,system
produc,correspond
provid,mean
provid,platform
provid,robot
provid,surgic
purpos,robot
real,time
relat,automatic
relat,method
relat,mobil
relat,robot
remot,control
remot,view
resolution,effector
robot,1
robot,10
robot,arm
robot,arrang
robot,automatic
robot,bodi
robot,capabl
robot,clean
robot,cleaner
robot,communic
robot,compris
robot,control
robot,includ
robot,invention
robot,main
robot,method
robot,mobil
robot,motion
robot,move
robot,movement
robot,mower
robot,operat
robot,pick
robot,position
robot,realiz
robot,robot
robot,s
robot,system
robotic,arm
robotic,control
robotic,devic
robotic,surgeri
robotic,system
rotari,brush
rotat,axe
rotat,head
rotat,motor
rotat,movement
rotat,shaft
scale,effector
scale,factor
seal,access
send,imag
sensor,mount
servo,motor
signal,receiv
signal,robot
signal,transmitter
slave,robot
smooth,surfac
sow,weed
surfac,clean
surgeon,adjust
surgeon,control
surgeon,input
surgeon,produc
surgeon,scale
surgeri,surgic
surgic,instrument
surgic,operat
surgic,procedur
surgic,robot
surgic,site
surgic,system
surgic,tool
sutur,tissu
system,autonomous
system,compris
system,control
system,devic
system,includ
system,method
system,mobil
system,perform
system,robot
system,use
thereof,invention
time,period
tissu,robotic
travel,perform
tube,apparatus
typic,movement
uneven,terrain
unit,arrang
unit,compris
unit,control
unit,drive
unit,generat
unit,provid
upper,lower
use,robotic
use,surgic
user,operat
vacuum,clean
vacuum,cleaner
vehicl,bodi
vertic,axi
video,signal
walk,robot
water,discharg
wheel,instal
wire,wireless
x,y
y,z

Table B.11: List of the 1206 variables used in the SVM for classification: All 110 trigrams.

adjust,position,handl
adjustabl,surgeon,control
allow,surgeon,adjust
apparatus,perform,minimal
arm,coupl,pair
arm,instrument,effector
button,allow,surgeon
button,depress,surgeon
button,effector,move
cardiac,procedur,system
clean,horizontal,vertic
clean,robot,1
cleaner,invention,relat
compris,main,bodi
control,input,button
control,method,thereof
control,resolution,effector
controller,handl,move
correspond,movement,effector
correspond,movement,typic
coupl,pair,master
coupl,pair,robotic
depress,surgeon,input
devic,main,controller
devic,robot,arm
effector,control,input
effector,correspond,movement
effector,handl,move
effector,manipulat,hold
effector,move,button
effector,movement,handl
effector,movement,movement
factor,adjustabl,surgeon
front,robot,arm
hand,surgeon,scale
handl,controller,handl
handl,move,comfortabl
handl,move,effector
handl,move,surgeon
handl,scale,effector
hold,sutur,tissu
includ,pair,surgic
independent,claim,includ
input,button,allow
input,button,effector
instrument,coupl,pair
instrument,effector,manipulat
invasiv,cardiac,procedur
invention,relat,automatic
invention,relat,method
invention,relat,mobil
manipulat,hold,sutur
master,handl,controller
method,invention,relat
method,thereof,invention
minimal,invasiv,cardiac
mobil,robot,invention
mobil,robotic,devic
mount,robot,arm
move,button,depress
move,comfortabl,position
move,effector,handl
move,surgeon,produc
movement,effector,control
movement,effector,movement
movement,handl,scale
movement,movement,effector
movement,perform,hand
movement,typic,movement
pair,master,handl
pair,robotic,arm
pair,surgic,instrument
perform,clean,operat
perform,hand,surgeon
perform,minimal,invasiv
position,handl,move
position,robot,arm
procedur,system,includ
produc,correspond,movement
relat,automatic,robot
resolution,effector,movement
robot,arm,includ
robot,cleaner,compris
robot,cleaner,invention
robot,control,method
robot,control,system
robot,invention,relat
robot,system,method
robotic,arm,coupl
robotic,arm,instrument
robotic,devic,compris
scale,effector,correspond
scale,factor,adjustabl
surgeon,adjust,position
surgeon,control,resolution
surgeon,input,button
surgeon,produc,correspond
surgeon,scale,factor
surgic,instrument,coupl
surgic,instrument,mount
surgic,robot,compris
surgic,robot,system
sutur,tissu,robotic
system,control,movement
system,includ,pair
system,perform,minimal
thereof,invention,disclos
tissu,robotic,arm
typic,movement,perform
x,y,z

Appendix C

Appendix. Spatial distribution of innovative activities

Table C.1: Descriptive statistics on industry divisions: Agriculture and production industries

Industry division		**No. of R&D est.**	**Ratio R&D est. / R&D comp.**	**Researchers**	**No. of 3-digit industries**
Agriculture					
1	Crop and animal production, hunting and related service activities	95	1.22	17.6	6
2	Forestry and logging	.	.	.	2
Production industries					
5	Mining of coal and lignite	.	.	.	2
6	Extraction of crude petroleum and natural gas	.	.	.	2
7	Mining of metal ores	.	.	.	1
8	Other mining and quarrying	29	1.00	4.3	2
9	Mining support service activities	8	1.00	7.9	2
10	Manufacture of food products	307	1.09	11.1	9
11	Manufacture of beverages	35	1.00	4.0	2
12	Manufacture of tobacco products	.	.	.	1
13	Manufacture of textiles	270	1.01	5.8	4
14	Manufacture of wearing apparel	79	1.00	12.2	2
15	Manufacture of leather and related products	29	1.00	7.7	2
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	118	1.02	5.2	2
17	Manufacture of paper and paper products	122	1.05	8.7	2
18	Printing and reproduction of recorded media	64	1.00	19.0	2
19	Manufacture of coke and refined petroleum products	19	1.00	26.9	2
20	Manufacture of chemicals and chemical products	841	1.09	32.9	6
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	269	1.09	92.5	2
22	Manufacture of rubber and plastic products	779	1.07	14.7	2
23	Manufacture of other non-metallic mineral products	494	1.05	9.6	8
24	Manufacture of basic metals	298	1.12	21.4	5
25	Manufacture of fabricated metal products, except machinery and equipment	1,446	1.02	8.6	8
26	Manufacture of computer, electronic and optical products	2,436	1.06	29.7	8
27	Manufacture of electrical equipment	907	1.06	28.7	6
28	Manufacture of machinery and equipment n.e.c.	3,178	1.03	18.8	5
29	Manufacture of motor vehicles, trailers and semi-trailers	480	1.28	212.6	3
30	Manufacture of other transport equipment	198	1.14	88.8	5
31	Manufacture of furniture	136	1.02	5.7	1
32	Other manufacturing	553	1.03	12.1	6
33	Repair and installation of machinery and equipment	207	1.11	18.7	2
35	Electricity, gas, steam and air conditioning supply	85	1.04	14.1	3
36	Water collection, treatment and supply	15	1.00	6.2	1
37	Sewerage	5	1.00	3.6	2
38	Waste collection, treatment and disposal activities; materials recovery	73	1.03	3.5	3
39	Remediation activities and other waste management services	6	1.00	3.5	1
41	Construction of buildings	74	1.03	8.4	2
42	Civil engineering	64	1.02	4.1	3
43	Specialised construction activities	346	1.02	4.2	4

(.) Statistical confidentiality because of 3 or less R&D establishments in the industry

Table C.2: Descriptive statistics on industry divisions: Service industries

Industry division	**No. of R&D est.**	**Ratio R&D est. / R&D comp.**	**Researchers**	**No. of 3-digit industries**
Service industries				
45 Wholesale and retail trade and repair of motor vehicles and motorcycles	45	1.02	10.6	4
46 Wholesale trade, except of motor vehicles and motorcycles	486	1.02	7.7	8
47 Retail trade, except of motor vehicles and motorcycles	118	1.03	5.0	8
49 Land transport and transport via pipelines	28	1.08	14.0	4
50 Water transport	4	1.00	4.3	3
51 Air transport	4	1.00	17.7	2
52 Warehousing and support activities for transportation	54	1.00	11.4	1
53 Postal and courier activities	.	.	.	2
56 Food and beverage service activities	4	1.00	2.5	2
58 Publishing activities	112	1.00	5.8	2
59 Motion picture, video and television program production, sound recording and music publishing activities	11	1.00	2.6	2
60 Programming and broadcasting activities	.	.	.	2
61 Telecommunications	35	1.17	81.7	4
62 Computer programming, consultancy and related activities	1,618	1.03	16.8	1
63 Information service activities	113	1.02	15.9	2
64 Financial service activities, except insurance and pension funding	15	1.07	29.1	3
65 Insurance, reinsurance and pension funding, exc. comp. social sec.	17	1.00	24.5	1
66 Activities auxiliary to financial services and insurance activities	4	1.00	206.9	2
68 Real estate activities	16	1.07	7.1	3
69 Legal and accounting activities	.	.	.	1
70 Activities of head offices; management consultancy activities	200	1.03	10.6	2
71 Architectural and engineering activities; techn. testing and analysis	1,436	1.04	11.8	2
72 Scientific research and development	1,017	1.04	22.1	2
73 Advertising and market research	34	1.00	8.8	2
74 Other professional, scientific and technical activities	61	1.02	6.9	4
75 Veterinary activities	4	1.00	6.9	1
77 Rental and leasing activities	19	1.12	3.7	3
78 Employment activities	4	1.00	3.0	2
79 Travel agency, tour operator and oth. reservation service and rel. act.	.	.	.	1
80 Security and investigation activities	.	.	.	2
81 Services to buildings and landscape activities	16	1.00	4.0	3
82 Office administrative, office support and oth. bus. support activities	103	1.00	6.0	4
84 Public administration and defense; compulsory social security	.	.	.	1
85 Education	10	1.00	8.4	3
86 Human health activities	31	1.00	3.4	3
87 Residential care activities	.	.	.	2
88 Social work activities without accommodation	.	.	.	2
90 Creative, arts and entertainment activities	4	1.00	3.0	1
93 Sports activities and amusement and recreation activities	4	1.00	3.0	2
94 Activities of membership organizations	7	1.00	3.7	2
95 Repair of computers and personal and household goods	6	1.00	4.0	2
96 Other personal service activities	65	1.00	4.3	1
TOTAL (divisions 01 to 96)	19,804	1.05	24.1	235

(.) Statistical confidentiality because of 3 or less R&D establishments in the industry

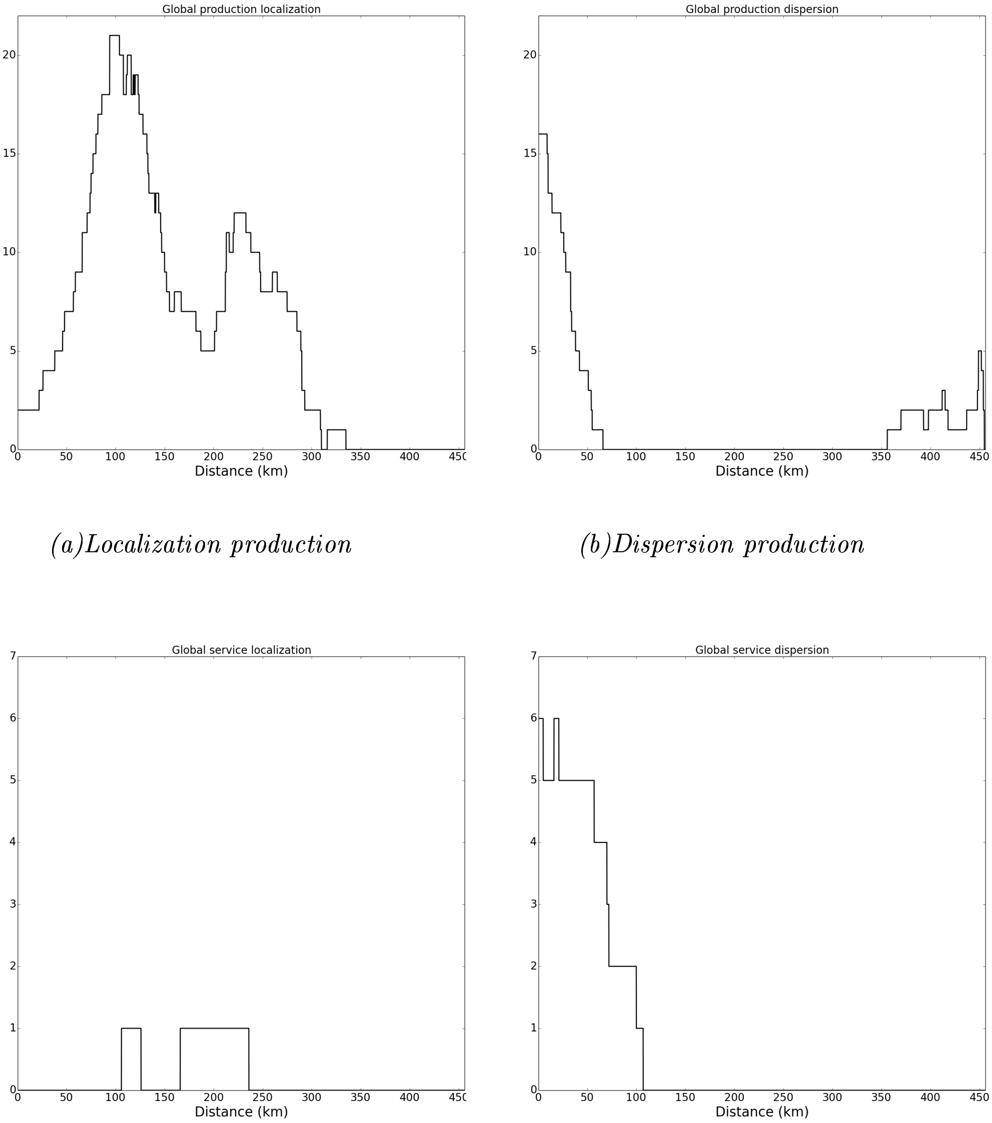
Table C.3: Industrial scope of localization patterns in agriculture and production industries

Industry division		**No. of 3-digit industries**	**Localized [%]**	**Dispersed [%]**	**Random [%]**
Agriculture					
1	Crop and animal production, hunting and related service activities	3			100.0
Production industries					
8	Other mining and quarrying	2			100.0
10	Manufacture of food products	7		28.6	71.4
11	Manufacture of beverages	1			100.0
13	Manufacture of textiles	4	25.0		75.0
14	Manufacture of wearing apparel	2	50.0		50.0
15	Manufacture of leather and related products	1			100.0
16	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials	2		50.0	50.0
17	Manufacture of paper and paper products	2			100.0
18	Printing and reproduction of recorded media	1			100.0
19	Manufacture of coke and refined petroleum products	1			100.0
20	Manufacture of chemicals and chemical products	6	33.3	33.3	33.3
21	Manufacture of basic pharmaceutical products and pharmaceutical preparations	2		50.0	50.0
22	Manufacture of rubber and plastic products	2	100.0		
23	Manufacture of other non-metallic mineral products	8	37.5	12.5	50.0
24	Manufacture of basic metals	5	60.0		40.0
25	Manufacture of fabricated metal products, except machinery and equipment	8	50.0	12.5	37.5
26	Manufacture of computer, electronic and optical products	7	42.9	57.1	
27	Manufacture of electrical equipment	6		33.3	66.7
28	Manufacture of machinery and equipment n.e.c.	5	80.0		20.0
29	Manufacture of motor vehicles, trailers and semi-trailers	3	66.7		33.3
30	Manufacture of other transport equipment	4	50.0	25.0	25.0
31	Manufacture of furniture	1	100.0		
32	Other manufacturing	5	20.0	40.0	40.0
33	Repair and installation of machinery and equipment	2		50.0	50.0
35	Electricity, gas, steam and air conditioning supply	2		50.0	50.0
36	Water collection, treatment and supply	1			100.0
38	Waste collection, treatment and disposal activities; materials recovery	2			100.0
41	Construction of buildings	1	100.0		
42	Civil engineering	3			100.0
43	Specialized construction activities	4		25.0	75.0

Table C.4: Industrial scope of localization patterns in service industries

Industry division	**No. of 3-digit industries**	**Localized [%]**	**Dispersed [%]**	**Random [%]**
Service industries				
45 Wholesale and retail trade and repair of motor vehicles and motorcycles	2			100.0
46 Wholesale trade, except of motor vehicles and motorcycles	8	12.5		87.5
47 Retail trade, except of motor vehicles and motorcycles	3			100.0
49 Land transport and transport via pipelines	1			100.0
52 Warehousing and support activities for transportation	1			100.0
58 Publishing activities	2			100.0
61 Telecommunications	1			100.0
62 Computer programming, consultancy and related activities	1		100.0	
63 Information service activities	2		50.0	50.0
64 Financial service activities, except insurance and pension funding	1			100.0
65 Insurance, reinsurance and pension funding, except compulsory social security	1			100.0
68 Real estate activities	1			100.0
70 Activities of head offices; management consultancy activities	2		50.0	50.0
71 Architectural and engineering activities; technical testing and analysis	2	50.0	50.0	
72 Scientific research and development	2		50.0	50.0
73 Advertising and market research	2			100.0
74 Other professional, scientific and technical activities	1			100.0
77 Rental and leasing activities	1			100.0
82 Office administrative, office support and other business support activities	1			100.0
86 Human health activities	1			100.0
96 Other personal service activities	1		100.0	

C.1 Distance-based sectoral location patterns



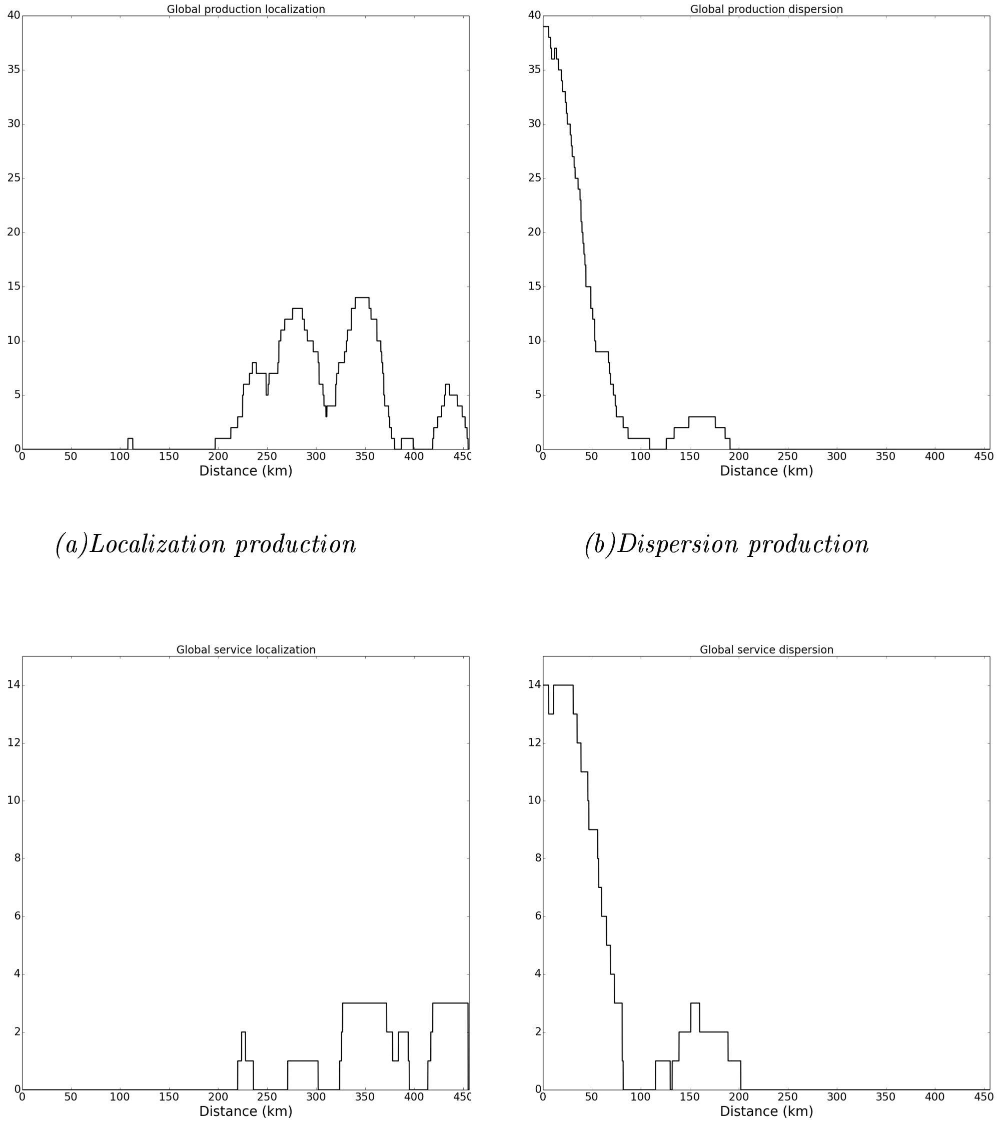
(a)Localization production

(b)Dispersion production

(c)Localization services

(d)Dispersion services

Figure C.1: Sectoral distance patterns of industries exhibiting localization and dispersion of R&D.



(a)Localization production

(b)Dispersion production

(c)Localization services

(d)Dispersion services

Figure C.2: Sectoral researcher-weighted distance patterns of industries exhibiting localization and dispersion of R&D.

C.2 Research-intensive 3-digit industries in Germany

Table C.5: List of research-intensive 3-digit production industries in Germany

3-digit production industry	
202	Manufacture of pesticides and other agrochemical products
211	Manufacture of basic pharmaceutical products
212	Manufacture of pharmaceutical preparations
254	Manufacture of weapons and ammunition
261	Manufacture of electronic components and boards
262	Manufacture of computers and peripheral equipment
263	Manufacture of communication equipment
265	Manufacture of instruments and appliances for measuring, testing and navigation
266	Manufacture of irradiation, electromedical and electrotherapeutic equipment
267	Manufacture of optical instruments and photographic equipment
303	Manufacture of air and spacecraft and related machinery
304	Manufacture of military fighting vehicles
201	Manufacture of basic chemicals, fertilizers and nitrogen compounds, plastics and synthetic rubber in primary forms
205	Manufacture of other chemical products
221	Manufacture of rubber products
264	Manufacture of consumer electronics
271	Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus
272	Manufacture of batteries and accumulators
274	Manufacture of electric lighting equipment
275	Manufacture of domestic appliances
279	Manufacture of other electrical equipment
281	Manufacture of general-purpose machinery
283	Manufacture of agricultural and forestry machinery
284	Manufacture of metal forming machinery and machine tools
289	Manufacture of other special-purpose machinery
291	Manufacture of motor vehicles
293	Manufacture of parts and accessories for motor vehicles
302	Manufacture of railway locomotives and rolling stock
325	Manufacture of medical and dental instruments and supplies

Source: Gehrke et al. (2013)

Table C.6: List of research-intensive 3-digit service industries in Germany

3-digit service industry	
620	Computer programming, consultancy and related activities
631	Data processing, hosting and related activities
712	Technical testing and analysis
721	Research and experimental development on natural sciences and engineering
722	Research and experimental development on social sciences and humanities

Source: Gehrke et al. (2010)

C.3 Cross-distance indices of localization and dispersion

Table C.7: Cross-distance indices of localization and dispersion

3-digit industry		**Location**	
		R&D establishm.	Researchers
Agriculture			
011	Growing of non-perennial crops		Ψ^r_m=0.0017
016	Support activities to agriculture and post-harvest crop activities		
Production industries			
081	Quarrying of stone, sand and clay		
089	Mining and quarrying n.e.c.		
101	Processing and preserving of meat and production of meat products		
103	Processing and preserving of fruit and vegetables	Ψ_m=0.0000	Ψ^r_m=0.0023
105	Manufacture of dairy products		Ψ^r_m=0.0013
106	Manufacture of grain, mill products starches and starch products		Ψ^r_m=0.0000
107	Manufacture of bakery and farinaceous products		
108	Manufacture of other food products	Ψ_m=0.0010	Ψ^r_m=0.0117
109	Manufacture of prepared animal feeds		
110	Manufacture of beverages		
131	Preparation and spinning of textile fibres		
132	Weaving of textiles		
133	Finishing of textiles		
139	Manufacture of other textiles	Γ_m=0.0122	Ψ^r_m=0.0007
141	Manufacture of wearing apparel, except fur apparel		Ψ^r_m=0.0003
143	Manufacture of knitted and crocheted apparel	Γ_m=0.0114	
151	Tanning and dressing of leather		
161	Sawmilling and planing of wood		
162	Manufacture of products of wood, cork, straw and plaiting materials	Ψ_m=0.0001	Ψ^r_m=0.0043
171	Manufacture of pulp, paper and paperboard		Ψ^r_m=0.0007
172	Manufacture of articles of paper and paperboard		Ψ^r_m=0.0025
181	Printing and service activities related to printing		Ψ^r_m=0.0017
192	Manufacture of refined petroleum products		
201	Manufacture of basic chemicals, fertilisers and nitrogen compounds, plastics and synthetic rubber in primary forms	Γ_m=0.0013	Γ^r_m=0.0011
202	Manufacture of pesticides and other agrochemical products		Ψ^r_m=0.0007
203	Manufacture of paints, varnishes and similar coatings, printing ink and mastics	Γ_m=0.0020	Γ^r_m=0.0008
204	Manufacture of soap and detergents, cleaning and polishing preparations, perfumes and toilet preparations	Ψ_m=0.0000	Ψ^r_m=0.0042
205	Manufacture of other chemical products	Ψ_m=0.0007	Γ^r_m=0.0009
206	Manufacture of man-made fibres		
211	Manufacture of basic pharmaceutical products		Ψ^r_m=0.0018
212	Manufacture of pharmaceutical preparations	Ψ_m=0.0023	Ψ^r_m=0.0137

Note: Γ_m=Cross-distance index of localization, Ψ_m=Cross-distance index of dispersion

Table C.8: Cross-distance indices of localization and dispersion – continued

3-digit industry		**Location**	
		R&D establishm.	Researchers
Production industries – continued			
221	Manufacture of rubber products	Γ_m=0.0010	Ψ^r_m=0.0025
222	Manufacture of plastic products	Γ_m=0.0226	Γ^r_m=0.0083
231	Manufacture of glass and glass products	Γ_m=0.0163	Γ^r_m=0.0012
232	Manufacture of refractory products		Ψ^r_m=0.0000
233	Manufacture of clay building materials		
234	Manufacture of other porcelain and ceramic products	Γ_m=0.0021	Ψ^r_m=0.0002
235	Manufacture of cement, lime and plaster		
236	Manufacture of articles of concrete, cement and plaster	Ψ_m=0.0007	Ψ^r_m=0.0059
237	Cutting, shaping and finishing of stone		
239	Manufacture of abrasive products and non-metallic mineral products n.e.c.	Γ_m=0.0000	Ψ^r_m=0.0004
241	Manufacture of basic iron and steel and of ferro-alloys	Γ_m=0.0100	
242	Manufacture of tubes, pipes, hollow profiles and related fittings of steel		
243	Manufacture of other products of first processing of steel	Γ_m=0.1044	
244	Manufacture of basic precious and other non-ferrous metals	Γ_m=0.0085	Ψ^r_m=0.0012
245	Casting of metals		Ψ^r_m=0.0008
251	Manufacture of structural metal products	Ψ_m=0.0015	Ψ^r_m=0.0004
252	Manufacture of tanks, reservoirs and containers of metal		Ψ^r_m=0.0010
253	Manufacture of steam generators, except central heating hot water boilers		
254	Manufacture of weapons and ammunition		
255	Forging, pressing, stamping and roll-forming of metal	Γ_m=0.0463	Γ^r_m=0.0001
256	Treatment and coating of metals	Γ_m=0.0003	Γ^r_m=0.0090
257	Manufacture of cutlery, tools and general hardware	Γ_m=0.0107	Γ^r_m=0.0083
259	Manufacture of other fabricated metal products	Γ_m=0.0127	Γ^r_m=0.0017
261	Manufacture of electronic components and boards	Γ_m=0.0016	Γ^r_m=0.0057
262	Manufacture of computers and peripheral equipment	Ψ_m=0.0001	Γ^r_m=0.0002
263	Manufacture of communication equipment	Ψ_m=0.0006	Ψ^r_m=0.0105
264	Manufacture of consumer electronics	Ψ_m=0.0003	Ψ^r_m=0.0038
265	Manufacture of instruments and appliances for measuring, testing and navigation	Γ_m=0.0001	Γ^r_m=0.0098
266	Manufacture of irradiation, electromedical and electrotherapeutic equipment	Ψ_m=0.0022	Ψ^r_m=0.0058
267	Manufacture of optical instruments and photographic equipment	Γ_m=0.0005	Ψ^r_m=0.0105
271	Manufacture of electric motors, generators, transformers and electricity distribution and control apparatus	Ψ_m=0.0017	Γ^r_m=0.0033
272	Manufacture of batteries and accumulators		
273	Manufacture of wiring and wiring devices		Ψ^r_m=0.0010
274	Manufacture of electric lighting equipment		Ψ^r_m=0.0004
275	Manufacture of domestic appliances		Ψ^r_m=0.0001
279	Manufacture of other electrical equipment	Ψ_m=0.0014	Ψ^r_m=0.0168
281	Manufacture of general-purpose machinery	Γ_m=0.0047	Γ^r_m=0.0012
282	Manufacture of other general-purpose machinery	Γ_m=0.0003	Γ^r_m=0.0083
283	Manufacture of agricultural and forestry machinery		Ψ^r_m=0.0014
284	Manufacture of metal forming machinery and machine tools	Γ_m=0.0148	Γ^r_m=0.0119
289	Manufacture of other special-purpose machinery	Γ_m=0.0086	Γ^r_m=0.0086

Note: Γ_m=Cross-distance index of localization, Ψ_m=Cross-distance index of dispersion

Table C.9: Cross-distance indices of localization and dispersion – continued

3-digit industry		**Location**	
		R&D establishm.	Researchers
Production industries – continued			
291	Manufacture of motor vehicles	Γ_m=0.0003	
292	Manufacture of bodies (coachwork) for motor vehicles		Ψ^r_m=0.0021
293	Manufacture of parts and accessories for motor vehicles	Γ_m=0.0395	Γ^r_m=0.0141
301	Building of ships and boats	Γ_m=0.0001	
302	Manufacture of railway locomotives and rolling stock		
303	Manufacture of air and spacecraft and related machinery	Ψ_m=0.0011	Ψ^r_m=0.0014
309	Manufacture of transport equipment n.e.c.	Γ_m=0.0027	
310	Manufacture of furniture	Γ_m=0.0010	Ψ^r_m=0.0037
322	Manufacture of musical instruments		
323	Manufacture of sports goods		
324	Manufacture of games and toys	Ψ_m=0.0001	
325	Manufacture of medical and dental instruments and supplies	Ψ_m=0.0028	Γ^r_m=0.0028
329	Manufacturing n.e.c.	Γ_m=0.0003	Ψ^r_m=0.0048
331	Repair of fabricated metal products, machinery and equipment		Ψ^r_m=0.0022
332	Installation of industrial machinery and equipment	Ψ_m=0.0003	Ψ^r_m=0.0069
351	Electric power generation, transmission and distribution	Ψ_m=0.0006	Ψ^r_m=0.0096
353	Steam and air conditioning supply		
360	Water collection, treatment and supply		
382	Waste treatment and disposal		
383	Materials recovery		Ψ^r_m=0.0004
412	Construction of residential and non-residential buildings	Γ_m=0.0005	Γ^r_m=0.0008
421	Construction of roads and railways		
422	Construction of utility projects		
429	Construction of other civil engineering projects		
431	Demolition and site preparation		
432	Electrical, plumbing and other construction installation activities	Ψ_m=0.0001	Ψ^r_m=0.0050
433	Building completion and finishing		Ψ^r_m=0.0026
439	Other specialised construction activities		Ψ^r_m=0.0032

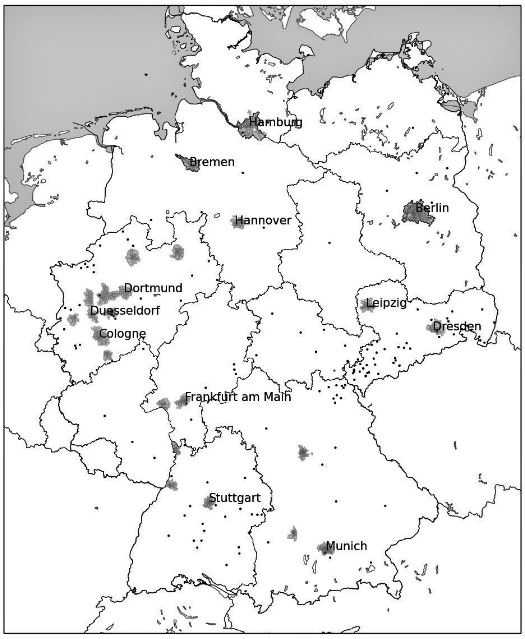
Note: Γ_m=Cross-distance index of localization, Ψ_m=Cross-distance index of dispersion

Table C.10: Cross-distance indices of localization and dispersion – continued

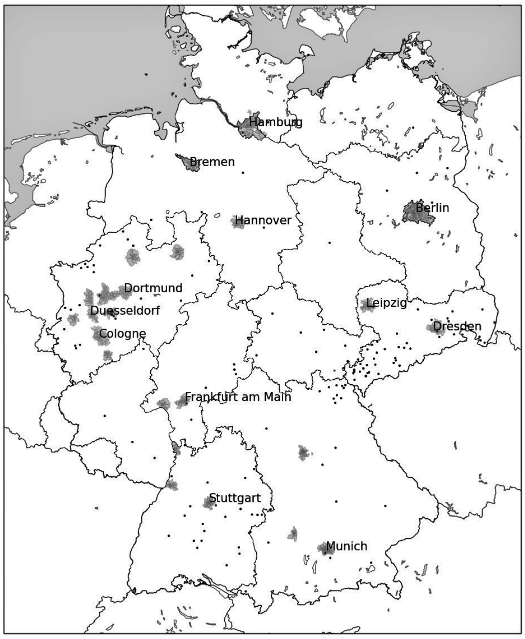
3-digit industry		**Location**	
		R&D establishm.	Researchers
Service industries			
452	Maintenance and repair of motor vehicles		
453	Sale of motor vehicle parts and accessories		
461	Wholesale on a fee or contract basis		$\Psi^r_m=0.0062$
462	Wholesale of agricultural raw materials and live animals		
463	Wholesale of food, beverages and tobacco		
464	Wholesale of household goods		$\Psi^r_m=0.0079$
465	Wholesale of information and communication equipment		$\Psi^r_m=0.0000$
466	Wholesale of other machinery, equipment and supplies		$\Psi^r_m=0.0066$
467	Other specialised wholesale	$\Gamma_m=0.0002$	$\Psi^r_m=0.0032$
469	Non-specialised wholesale trade		
474	Retail sale of information and communication equipment in specialised stores		$\Psi^r_m=0.0016$
475	Retail sale of other household equipment in specialised stores		
477	Retail sale of other goods in specialised stores		
493	Other passenger land transport		
522	Support activities for transportation		$\Psi^r_m=0.0006$
581	Publishing of books, periodicals and other publishing activities		
582	Software publishing		$\Psi^r_m=0.0079$
619	Other telecommunications activities		
620	Computer programming, consultancy and related activities	$\Psi_m=0.0061$	$\Gamma^r_m=0.0095$
631	Data processing, hosting and related activities	$\Psi_m=0.0021$	$\Psi^r_m=0.0111$
639	Other information service activities		
641	Monetary intermediation		
651	Insurance		
682	Rental and operating of own or leased real estate		
701	Activities of head offices		$\Psi^r_m=0.0015$
702	Management consultancy activities	$\Psi_m=0.0008$	$\Psi^r_m=0.0178$
711	Architectural and engineering activities and related technical consultancy	$\Gamma_m=0.0018$	$\Gamma^r_m=0.0096$
712	Technical testing and analysis	$\Psi_m=0.0022$	$\Psi^r_m=0.0142$
721	Research and experimental development on natural sciences and engineering	$\Psi_m=0.0057$	$\Gamma^r_m=0.0102$
722	Research and experimental development on social sciences and humanities		
731	Advertising		
732	Market research and public opinion polling		
749	Other professional, scientific and technical activities n.e.c.		$\Psi^r_m=0.0034$
773	Rental and leasing of other machinery, equipment and tangible goods		
829	Business support service activities n.e.c.		$\Psi^r_m=0.0052$
869	Other human health activities		
960	Other personal service activities		$\Psi^r_m=0.0014$

Note: Γ_m=Cross-distance index of localization, Ψ_m=Cross-distance index of dispersion

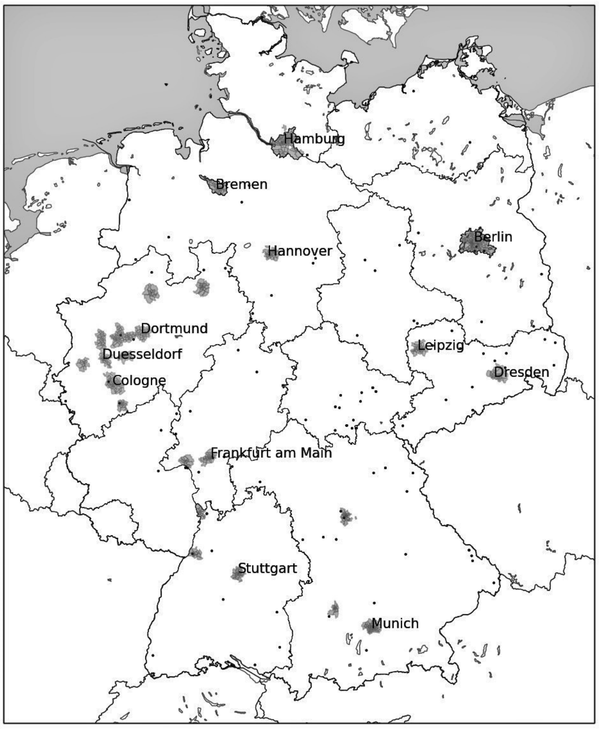
C.4 Selected maps of localized 3-digit industries



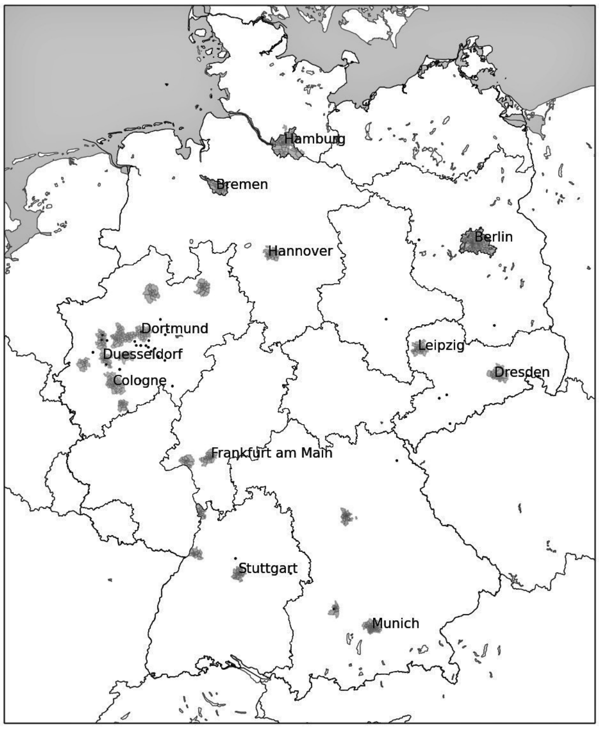
139 Manufacture of other textiles



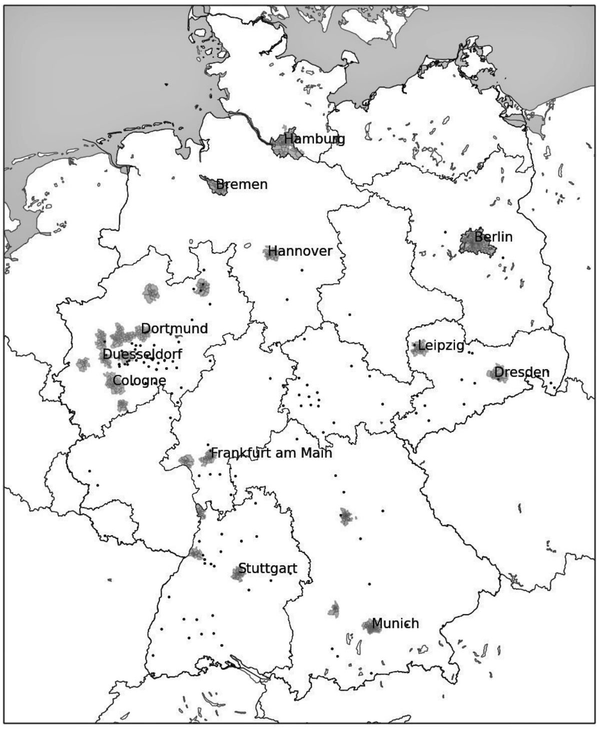
143 Manufacture of knitted and crocheted apparel



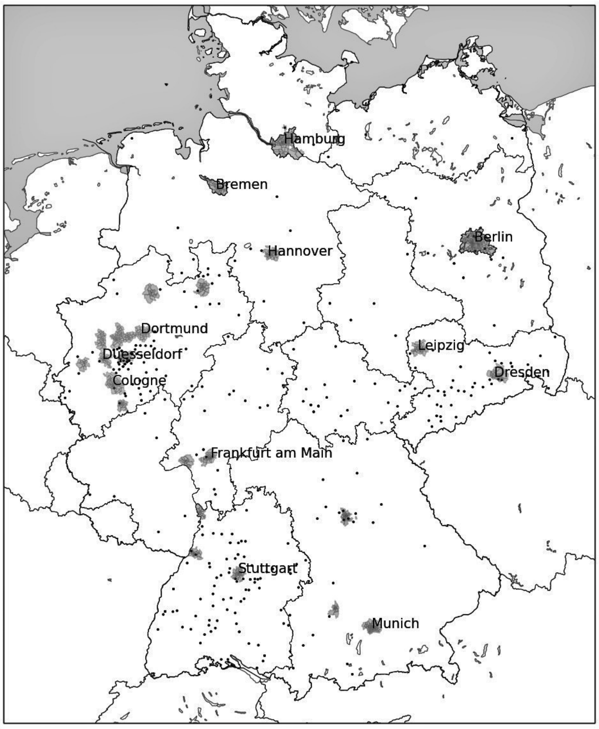
231 Manufacture of glass and glass products



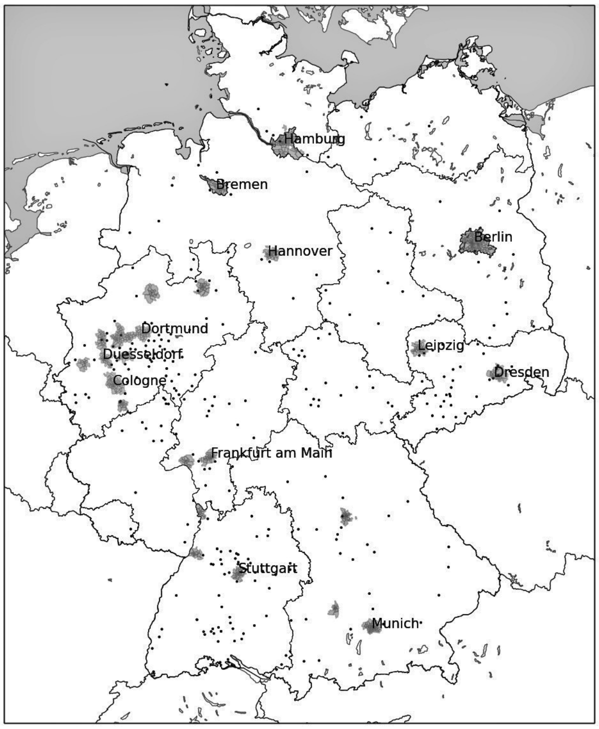
243 Manufacture of other products of first processing of steel



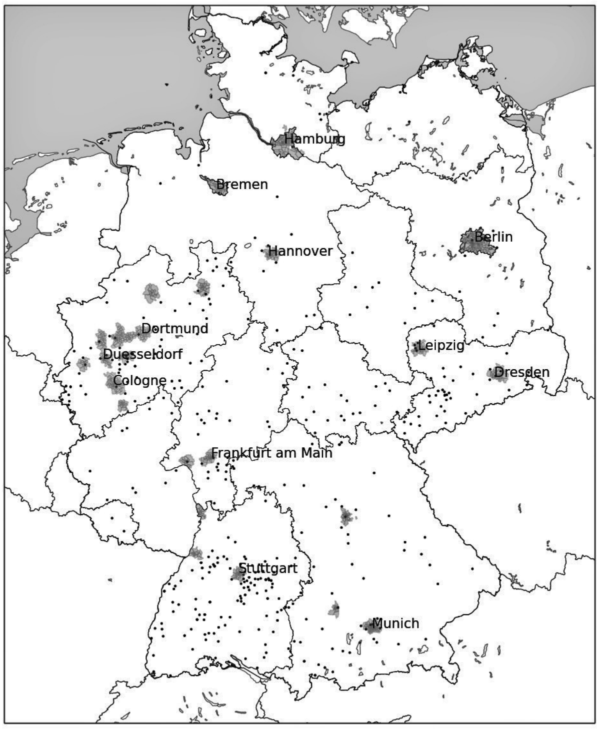
255 Forging, pressing, stamping and roll-forming of metal



257 Manufacture of cutlery, tools and general hardware



259 Manufacture of other fabricated metal products



284 Manufacture of metal forming machinery and machine tools

Appendix D

Appendix. Firms' sustained superior job creation

Table D.1: Results of simulations for ***top 10 %***

Markov order	p	Relative Growth Measure UK		ES		FR		IT		DE	
I	p<0.05	2287.1 (47.4)	**2887**	10502.7 (105.1)	**14968**	7427.4 (77.3)	**10493**	3390.4 (53.5)	**4266**	2560.5 (53.4)	**3065**
I	p<0.01	412.6 (17.0)	**708**	2155.4 (50.3)	**4110**	1673.8 (40.0)	**3149**	1251.0 (33.6)	**1893**	352.6 (21.5)	**586**
I	p<0.001	25.4 (5.0)	**81**	118.4 (11.9)	**322**	97.4 (7.9)	**231**	19.9 (4.4)	18	17.2 (3.7)	**84**
II	p<0.05	2091.9 (50.6)	**2349**	10352.2 (97.3)	7589	6787.8 (78.3)	5307	3780.6 (62.7)	2695	2603.1 (50.0)	2149
II	p<0.01	347.2 (17.4)	**518**	995.8 (29.5)	927	632.1 (24.0)	693	473.1 (16.9)	341	339.7 (19.7)	323
II	p<0.001	24.5 (5.1)	**78**	109.6 (10.0)	**277**	110.6 (10.3)	**232**	9.3 (3.3)	7	8.9 (3.5)	15
Markov order	p	Absolute Growth Measure UK		ES		FR		IT		DE	
I	p<0.05	2311.1 (46.1)	**2909**	10502.7 (105.1)	**14968**	7576.9 (86.6)	**10658**	4862.0 (65.5)	**6583**	2252.2 (53.0)	**2830**
I	p<0.01	426.1 (21.4)	**722**	2209.7 (41.0)	**4295**	1759.2 (41.3)	**3355**	1181.5 (28.0)	**1962**	286.2 (18.9)	**493**
I	p<0.001	26.3 (4.9)	**85**	119.8 (8.7)	**321**	34.3 (5.8)	**65**	65.4 (8.2)	**103**	13.4 (3.4)	**82**
II	p<0.05	2175.8 (45.1)	**2358**	8270.9 (80.7)	7786	7368.6 (85.2)	5454	4209.4 (58.5)	2950	2193.5 (51.6)	1941
II	p<0.01	362.7 (16.6)	**526**	773.0 (27.6)	**994**	793.5 (26.3)	769	443.2 (18.4)	354	490.8 (20.3)	485
II	p<0.001	23.4 (4.9)	**82**	130.1 (12.9)	**327**	26.9 (5.6)	36	26.6 (5.0)	16	40.8 (7.0)	**75**

Table D.2: Results of simulations for ***top 20 %***

Markov order	p	Relative Growth Measure									
		UK		ES		FR		IT		DE	
I	p<0.05	1950.6 (39.7)	**2422**	13128.0 (119.5)	**15427**	10067.5 (87.6)	**12008**	5743.6 (72.3)	**7026**	1951.5 (43.0)	**2549**
I	p<0.01	287.9 (16.3)	**588**	2039.2 (50.9)	**2979**	945.8 (33.0)	**1477**	694.8 (29.6)	**833**	398.2 (17.7)	**906**
I	p<0.001	20.6 (4.4)	**73**	65.5 (6.5)	**257**	65.8 (6.5)	**185**	14.7 (3.7)	19	0.6 (0.7)	**21**
II	p<0.05	2001.6 (45.1)	**3102**	6696.0 (75.0)	**8811**	4328.4 (65.2)	**6394**	2963.6 (53.7)	2737	2088.7 (45.8)	**2309**
II	p<0.01	311.4 (21.8)	**856**	1651.0 (42.5)	**2965**	1387.6 (30.7)	**2656**	677.5 (27.4)	**842**	128.3 (10.1)	**279**
II	p<0.001	22.0 (4.3)	**181**	44.8 (6.5)	**248**	38.8 (6.3)	**168**	9.6 (3.5)	18	0.0 (0.0)	**21**
Markov order	p	Absolute Growth Measure									
		UK		ES		FR		IT		DE	
I	p<0.05	1959.8 (41.5)	**2494**	12961.5 (126.2)	**16060**	10200.5 (97.7)	**12297**	5566.6 (73.8)	7396	1781.3 (44.2)	**2379**
I	p<0.01	291.4 (20.3)	**594**	1982.6 (49.5)	**3136**	962.6 (28.6)	**1580**	674.8 (30.4)	**902**	354.1 (19.6)	**827**
I	p<0.001	20.2 (4.6)	**78**	61.3 (7.2)	**193**	66.7 (9.3)	**200**	14.2 (4.0)	18	9.4 (3.1)	**68**
II	p<0.05	2359.3 (49.6)	**3125**	5624.9 (79.6)	**8961**	4423.5 (68.3)	**6409**	2994.9 (52.2)	2828	2133.5 (47.3)	**2406**
II	p<0.01	200.9 (15.1)	**584**	1358.2 (31.4)	**3124**	1439.9 (31.9)	**2806**	697.8 (28.6)	**906**	178.5 (13.6)	**363**
II	p<0.001	3.5 (2.1)	**78**	156.1 (13.2)	**618**	39.9 (5.5)	**185**	10.7 (3.3)	17	0.0 (0.0)	**19**

Table D.3: Benchmarks for various life spans of the firm. ***First order Markov. Top 10%***

Observed life span	p	Relative Growth				Absolute Growth			
		UK	ES	FR	IT	UK	ES	FR	IT
2	0.05	1	1	1	1	1	1	1	1
	0.01	2	1	1	1	2	1	1	1
3	0.05	1	1	1	1	1	1	1	1
	0.01	2	2	2	2	2	2	2	2
4	0.05	2	1	1	2	2	1	1	1
	0.01	2	2	2	2	2	2	2	2
5	0.05	2	2	2	2	2	2	2	2
	0.01	3	2	2	2	3	2	2	2
6	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
7	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3

Table D.4: Benchmarks for various life spans of the firm. ***First order Markov. Top 20%***

Observed life span	p value	Relative Growth				Absolute Growth			
		UK	ES	FR	IT	UK	ES	FR	IT
2	0.05	2	1	1	1	2	1	1	1
	0.01	2	2	2	2	2	2	2	2
3	0.05	2	2	2	2	2	2	2	2
	0.01	3	2	3	2	3	2	3	2
4	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
5	0.05	3	3	3	3	3	3	3	3
	0.01	4	3	3	3	4	3	3	3
6	0.05	3	3	3	3	3	3	3	3
	0.01	4	4	4	4	4	4	4	4
7	0.05	4	3	3	3	4	3	3	3
	0.01	5	4	4	4	5	4	4	4

Table D.5: Benchmarks for various life spans of the firm. ***Second order Markov. Top 10%***

Observed life span	p value	Relative Growth				Absolute Growth			
		UK	ES	FR	IT	UK	ES	FR	IT
3	0.05	1	1	1	1	1	1	1	1
	0.01	2	2	2	2	2	2	2	2
	0.001	3	2	2	3	3	2	3	3
4	0.05	2	2	2	2	2	2	2	2
	0.01	3	2	2	2	3	2	2	2
	0.001	3	3	3	3	3	3	3	3
5	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
	0.001	4	4	3	4	4	3	4	3
6	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
	0.001	4	4	4	4	4	4	4	4
7	0.05	2	2	2	3	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
	0.001	5	4	4	4	5	4	4	4

Table D.6: Benchmarks for various life spans of the firm. ***Second order Markov. Top 20%***

Observed life span	p value	Relative Growth				Absolute Growth			
		UK	ES	FR	IT	UK	ES	FR	IT
3	0.05	2	2	2	2	2	2	2	2
	0.01	3	2	2	2	3	2	2	2
	0.001	3	3	3	3	3	3	3	3
4	0.05	2	2	2	2	2	2	2	2
	0.01	3	3	3	3	3	3	3	3
	0.001	4	4	4	4	4	4	4	4
5	0.05	3	3	3	3	3	3	3	3
	0.01	4	3	3	3	4	3	3	3
	0.001	5	4	4	4	5	4	4	4
6	0.05	3	3	3	3	3	3	3	3
	0.01	4	4	4	4	4	4	4	4
	0.001	5	5	5	5	5	4	5	5
7	0.05	3	3	3	3	3	3	3	3
	0.01	4	4	4	4	5	4	4	4
	0.001	5	5	5	5	6	5	5	5