M.Sc. Thesis

Master of Science in Engineering

DTU Management Engineering
Department of Management Engineering

Measuring the uniqueness of technological capabilities

A data-driven network exploration

Duarte Oliveira e Carmo (s160951)

Kongens Lyngby 2018

DTU Management Engineering

Department of Management Enginnering

Technical University of Denmark

Produktionstorvet

Building 426

DK-2800 Kgs. Lyngby, Denmark

Phone +45 45 25 48 00

Briefing

What?

This project is a master's thesis written by Duarte O.Carmo, an Engineering Management Student at the Technical University of Denmark. In summary, the project aims to measure the technological capabilities of countries, time periods, or organizations with the goal of understanding their intricacies and help policy makers and stakeholders make better informed decisions.

Why?

The idea for this master's thesis was born from two other initiatives:

- **EURITO**: this European Commission Project brings together a consortium of organizations with the goal of utilizing new data sources and methods to improve Research and Innovation. In summary, the European Commission believes that there is a need to improve R&I policies and that science data can play an important role in helping resolve this need.

- **AMICa**: AMICa stands for Advanced Mapping of Industrial Capabilities. The project aims at mapping worldwide industrial capabilities to support the development of new technologies, products, and services with a positive climate change impact. Its first proof of concept uses the development of sustainable biofuels as a proof of concept for capability mapping.

It is believed that this MSc. Project will help push these two initiatives forward by providing a proof of concept on how the data collected can be used to help entities develop better policies, and answer questions that might arise during the decision

making process.

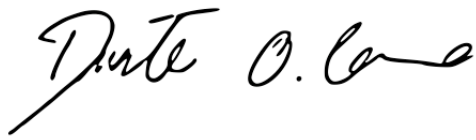
Who?

- Project Responsible: Duarte Ribeiro Oliveira e Carmo
- Supervisor: Pedro Parraguez Ruiz
- Supervisor: Anja Maier
- Institution: Technical University of Denmark
- Department: DTU Management Engineering

Preface

This Master thesis was prepared at the department of Engineering Management at the Technical University of Denmark in fulfillment of the requirements for acquiring a master degree in Engineering Management.

Kongens Lyngby, June 7, 2018



Duarte Oliveira e Carmo (s160951)

Acknowledgements

I would first like to thank my thesis supervisor Pedro Parraguez Ruiz, a postdoctoral researcher at the Engineering Systems Division of the Management Engineering department at the Technical University of Denmark. When I first contacted Pedro in December 2017, I had no idea of the great choice I had just made. Throughout the semester he was always available to help me structure my thoughts, to enlighten me about possible research paths, and to discuss the endless possibilities that such research can have. Moreover, his passion for this research field and his complete availability at all times, made this experience one of the most fulfilling throughout my academic journey at DTU.

I would also like to thank my main supervisor Prof. Anja Maier, Head of Division at DTU Management Engineering for her support and guidance. This allowed me to develop what I strongly believe is my own work.

Moreover, I would also like to thank the Technical University of Denmark. It seems like yesterday that I was browsing possible Master's programmes to apply for, since then, not one day has passed where I have regretted to participate in the Engineering Master's degree at DTU. I strongly believe that the incredible academic staff, the talented students with whom I have interacted, and the out-of-this-world conditions have all contributed equally to prepare me for the future.

Finally, I have to thank my family and my close friends for listening to me geek out about topics that have absolutely nothing to do with their interests but also for endless years of unconditional support in all its dimensions.

This project would not have been possible without any of the above mentioned. Thank you.

Contents

CHAPTER 1

Introduction

In the first chapter of the report, the project will be introduced. This introduction is essential for the comprehension of the scope of the work that is going to be developed. Several parts shape this chapter:

- Purpose: the reason why this project took place under the academic context.
- Context: the previous work that has been developed by the Engineering Systems department at the Technical University of Denmark.
- Research Approach: the overall approach that was taken for researching the project.
- Report Structure: How the report is structured and how the results are presented.

1.1 Purpose

This project was developed as a Master's Thesis project for the Engineering Management MSc degree at DTU. The Master's program in Engineering management seeks to shape *"engineers capable of working at the interface between business and engineering"*(DTU 2018). This reinforces the idea that both qualitative and quantitative competencies are relevant for a project of this scope. This research should not only satisfy certain learning objectives but also touch upon more specific fields that were studied during the program. It is believed that the subject of this thesis is highly connected to at least 4 topics that were studied during the above-mentioned degree.

Sustainability

Sustainability is an intricate part of DTU's strategy, as mentioned in its mission statement: *"The University will apply significant research strength to laying the foundations for technologies and processes that promote innovation and sustainability, and which address major societal challenges"*(*DTU's strategy 2014-2019 - DTU*).

Not only as an important part of this particular institution, Sustainability is also a global concern that should be addressed with a high degree of importance. As shown in several research works such as, for example, the work of Karel Mulder where he mentions that various civilizations throughout the world have collapsed as a result of unsustainable practices in different areas such as agriculture (Mulder 2006).

For this reason, this project focuses on sustainability as a core part of eco-innovation by exploring the knowledge network behind biofuel research and how it can help stakeholders make more transparent and environmentally aware decisions.

Management and Innovation

Management is another core pillar of the master's degree. It constitutes the qualitative field of research and is made up of several different areas. This project is highly related to management by focusing on areas such as:

- **Innovation**: By studying the knowledge network behind biofuels, there is an opportunity for understanding not only how it works, but also how companies create and retain value in the context of eco-innovation. For example, patents and publications are quantitative measures of innovation.

- **Levels**: Complex systems can be studied in levels. For example, one can distinguish the world, the national, and the organizational, levels. Throughout this research, all of these levels will be taken into consideration, as well as the complex networks that characterize them and their development through time.

- **Differentiation**: Technology is one of the ways of strategically differentiating a country or an organization. In this research, the factors that differentiate entities will try to be understood, as well as the intricacies of the relationships between countries, organizations and universities.

Data Science

The data revolution is upon the world - by data it is meant "lose information". For this reason, it is hard to understand why this revolution should not serve as a catalyzer for solving other problems that may lie in other areas such as management, sustainability, or even innovation.

Preparing, studying and analyzing large data sets is another important characteristic of this project. The dataset that is going to be studied is composed of 4585 patents and 5313 scientific publications; moreover, the complex dataset is stored in a graph database composed of more than 100 countries and 10,000 distinct organizations.

Data science per se is not enough to solve problems, in order to make sense of the vast amount of information that such a database carries, one must apply efficient and effective techniques for the analysis.

One of the reasons why this project exists is to help organizations, policymakers, and researchers to make sense of this type of information. For this reason, the project also focuses on the application of algorithms that would be classified as Unsupervised Learning, as a way of trying to understand the hidden patterns that might lie in the data.

In the next section, more on this will be discussed.

1.2 Context

The following project is not the first effort in this direction, in reality, two other projects have carried out efforts in connecting the fields in the previous section. On the broader side, the EURITO project, funded by the European Commission and on the narrower, the more specific side, the AMICa-Pathfinder project at the Technical University of Denmark. In this part of the report, a more in-depth explanation will be given on each one of these initiatives.

EURITO - EU Relevant, Inclusive, Timely, Trusted, and Open Research Innovation Indicators

This project created and funded by the Community Research and Development Information Service of the European Commission has the high-level goal of: "better

integration of evidence on the impact of research and innovation in policy making". Furthermore, it carries a total contribution of around €1.5M. The project consists of the participation of three institutions: the Fraunhofer research organization (GER), the COTEC innovation foundation (ESP), and the Technical University of Denmark (DK).

This project's main goal is to bring big data and data analytics to the heart of Research and Innovation(R&I) policy. By first defining the R&I policy user needs and then turning those needs into analytics data pilots in an exploration stage. Moreover, by creating new R&I indicators, and communicating them through interactive visualizations, the project expects to make R&I policy-making more transparent and democratic. Also, the project description includes several considerations about big data and machine learning but argues that there are concerns about representativity, accuracy, and interpretability in what concerns the sources of data.

The success of this project would mean that R&I policies are better informed, better targeted, and that new innovation opportunities could surface. This is because of the open data, code, and knowledge developed alongside the project. (*EURITO, 2018*)

AMICa - Advanced Mapping of Industrial Capabilities for Climate

AMICa is a project led and executed by the Engineering System Division at DTU Management Engineering and funded by Climate-KIC. The participating members in this project are Chalmers University, MASH-Biotech, The Nordic Initiative for Sustainable Aviation (NISA) and Novozymes.

The project's main goal is to facilitate better data-driven decision making, and providing assistance for designing, developing and implementing more sustainable pro-

duction systems using pre-existing capabilities. With a specific technological target in scope, the project hopes to answer questions such as: are there untapped research gaps? Are there hotspots of unexploited but complementary capabilities? What organizations are unique? An innovative approach is applied to this research problem: AMICa focuses on technological capabilities instead of flows of material and makes use of a complex system view. This complex system view ultimately results in an input-process-output model.

Success for the AMICa project would translate into a fruitful mapping of worldwide industrial capabilities that can support the development of new technologies, products, and services with a positive climate change impact. The first proof of concept utilizes biofuel research as a starting point for this mapping.

More on the technical specifications for this project will be given in Chapter 3. (*AMICa, 2018*)

Thesis - Measuring the uniqueness of technological capabilities

AMICa and EURITO are two complementary projects in technical terms. In fact, EURITO proposes a theoretical possibility (or idea), and the AMICa project seeks to get closer to the practical applications and implications of such an idea.

By taking biofuel research and trying to map the capabilities of such a field, this thesis aims to provide a proof of concept that is highly modular. By modular, it is meant that the procedures applied to biofuel research can possibly be applied to virtually any field of research, following the complex system approach. Some visualizations and practical applications have already been tested by AMICa, such as the Data Exploration Dashboard or the Sankey Diagram visualizations. However, there is a need to:

- Further understand the how this data can be explored and used by industry and government.
- Provide more visualizations with other dimensions.
- Show possible applications of Big Data processing tools.

In this context, this project appears as a natural extension of the AMICa pathfinder project, by exploring this biofuel-related data, but also as a source of potential an-

alytical approaches for the EURITO project, given the importance of data-driven insights for policymakers and other institutions.

1.3 Research Approach

Research Motivation

There is a need of companies, countries, and organizations to improve Research and Innovation policies. Most approaches rely on methods that are not reusable, applicable to other fields and produce hard to prove results.

Research Goal

Helping decision makers understand highly interconnected technological landscapes and as a result make better R&I decisions using knowledge data. For this purpose biofuels will be used as testing ground.

Based on this research goal, several research questions were developed. These research questions offer a more concrete and quantifiable way of evaluating the research and more particularly, how these stakeholders would be helped. On section 2.1, the reasoning and the more practical aspects of these research questions will be described.

Research Questions

The research is structured following the scale of the examined levels of analysis, from macro to micro. The first level shows the dynamic analysis of the studied system as a whole (worldwide biofuel research and innovation), and serves as an introduction to the complex system model. The second level of analysis focuses on a meso level focusing on the country level. After the meso level, the study narrows with the analysis of the system at the micro level, this corresponds to organizations such as universities and businesses. Finally, two complementary analyses will be made. The first focuses on the intricacies of the system itself and the elements that compose it - patents and publication in this case - . The second is a validation exercise where, as a baseline, a more basic approach will be used and compared the general approach taken throughout the thesis.

The research questions follow:

Macro Level

General:

- How does the research of different terms develop over time?

Differences in research:

- How to characterize the similarity between two periods in time?

Clustering:

- What differences exist when comparing different years in scientific research?
- Are all years linearly related or do research gaps exist?

Context:

- Is biofuel research correlated with the price of oil? What exact terms are the most related to it?
- Is the price of a consumer good related with the research volume? What terms are most affected by it?

Meso Level

General:

- How to characterize the scientific capability of a country?
- How are countries related from a capability perspective?

Clustering:

- Do international capability clusters exist?

Context:

- Does the amount of money of you have (GDP per capita) determine the space of possibilities or your technological freedom?
- Does the GDP per capita play a part in the capability similarity of countries?

Collaboration:

- Is capability similarity related to international collaboration?

Differences in research:

- How to characterize the similarity between two countries?

Country Spectrums:

- Is there value in characterizing a country's capability as a spectrum?

Micro Level

General:

- How to characterize the capability of a given organization?

Clustering:

- Do organizational clusters exist?

Comparing:

- How to compare two organizations from a capability perspective?

Collaboration:

- Do universities collaborate more with universities or businesses? Do businesses collaborate more with universities?

Organizational Spectrums:

- Is there value in characterizing an organization's capability as a spectrum?

Patents and Publications

Evolution:

- How does the volume of patenting and publishing of a certain scientific term evolve over time?

Differences:

- Is there a bias towards patenting or publishing? An analysis of feedstocks, processing technologies and outputs.

Basic Approach, term frequency

General:

- Can we apply unsupervised learning techniques to visualize scientific clusters?
- Term frequency vs term pair frequency? What characterizes better the scientific capability of a country or organization?

Research Approach

As a guiding approach to these research problems, the Design Research Methodology (Blessing and Chakrabarti 2009)was adopted. Figure 1.1 provides an outline of the elements that characterize this method: which include criteria definition, descriptive study 1, prescriptive study, and descriptive study 2. These correspond to criteria definition, literature analysis, method development, and application/success evaluation.

Although the DRM approach makes for the richest analysis possible, this framework was used as more of a guideline than a step by step methodology. In fact, some elements of the framework require extensive validation and cycling throughout all of the levels that compose it. For this reason, the thesis focuses on two main elements of the DRM.

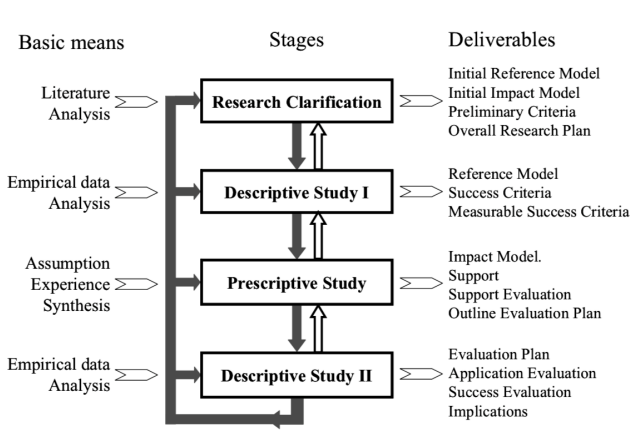


Figure 1.1: Design research methodology representation.

The first element used is the descriptive study 1, where, taking as a starting point the criteria (e.g. research questions), the analysis and pattern detection in the dataset was made. The second element used, to a less extensive extent, is the prescriptive study 2b where using the experience and some basic assumptions about the behavior of the research landscape, conclusions were reached. It should be pointed out that throughout the analysis; an effort was made to continuously validate the results, through investigation of underlying behavior.

Finally, as this thesis seeks to assist a group of stakeholders in making better decisions through data, these individuals, utilizing their experience and superior domain knowledge, can draw on their own conclusions using the data provided, which by itself leads to a type of prescriptive study.

1.4 Report Structure

As previously mentioned , this project uses the DRM framework as a guideline for the research. Based on this same methodology, the project was structured in the following manner:

- **Chapter 2** includes the statement of the practical motivation for the project. Following this statement, a literature review is provided. This literature review details the necessary studies across the macro, micro, and systems described above. Finally, a description of the analytical and conceptual approach will be provided.

- **Chapter 3** describes the in-depth methodology. It starts with a description of the data model and dataset. Following this description, the graph model is introduced as a way of relating the data. This chapter concludes by providing an overview of the quantitative analysis and techniques used to achieve it.

- **Chapter 4** presents the results of the application of the methodology to the different dimensions of the dataset. This includes the chronological, macro, micro, and complex system dimension.

- **Chapter 5** develops a discussion of the different topics developed in the previous chapter. This includes an interpretation of the results but also a statement

of possible limitations that arise during the analysis. It is worth noting that only relevant results will be discussed.

- **Chapter 6** closes the thesis by providing a statement of the theoretical and practical implications of the research, including connections to the literature review and recommendations for industry and political stakeholders. This chapter ends with considerations about further research.

This project can also be read online or downloaded as a .pdf. All of the code for this project can be consulted in the GitHub repository.

CHAPTER 2

Challenge and related literature

In the second chapter of this report, two important preliminary themes will be addressed:

- Motivation: the motivation part of the chapter will consist in outlining the starting point of this project as well as the practical implications that the project might have.
- Literature Review: in the second part of the chapter, previous work on these topics will be covered. This coverage consists of a three part literature review that
- Analytical and Conceptual Approach: in the final part of this chapter considerations about the viability of a system approach to both the macro, meso and micro levels.

2.1 Motivation

"If data had mass, the earth would be a black hole"- Stephen Marsland

We live in the age of information, or, as many refer to, the age of data. Every day, 2.5 quintillion bytes of data are created - 90 percent of it in the last two years (*IBM Analytics*). This information comes from a wide range of different fields and in a wide range of shapes. One of the biggest challenges in the forthcoming years will consist of successfully leveraging it. People, companies, countries, and even international

organizations have started to realize it. The European Union, for example, with the funding of the EURITO project (*EURITO, 2018*) has already started to tackle this challenge. The project wishes to help countries and organizations make better informed R&I policies, create new innovation opportunities and enhance the understanding of these systems, all by leveraging vast amounts of information, bringing benefits to both policymakers, researchers, and businesses.

As the scope tightens, the Technical University of Denmark is one of the organizations that wants to help advance this initiative by taking part in the EURITO project. In partnership with companies and institutions, AMICa, seeks to help solve a number of challenges related to the positive impact that industry can have on climate change. Such impact includes leveraging data to support management of complex value chains, identification of new research opportunities, and development of a new system layout design. As a general rule, this master thesis project wants to provide stakeholders with better tools to make data-driven decisions in what regards their business or operational logic. These stakeholders, just like in the project EURITO, include technology developers, organizations, and policymakers.

To prove the concept, researchers at DTU started by creating a database of technological assets (patents, scientific publications and projects) which will be further described in chapter 3. These are all related to biofuel research. After cleaning, preparing and analyzing the data, some results have already started to surface. Data exploration dashboards (*AMICa, 2018*), which help understanding how technological assets are distributed, and sankey diagram implementations that help to understand in what country or organization focuses on in terms of biofuels research.

However, another need soon surfaced, what more insights lie in this vast amount of data? How can this information help decision makers and other stakeholders understand the technological landscape and make better R&I decisions by further exploiting this knowledge data? The possibilities of the analysis of this knowledge data are almost endless, but the problems that need solving can easily be quantifiable (see chapter 2). With the goal of not losing the focus, this project benefits participants in three different levels.

The **macro level**, wishes to help policy makers, as the first stakeholder, in assessing and understanding the "big picture" biofuel research landscape at a global level, and how this system might evolve over time. One problem, for example, can be the need for understanding the differences and similarities between two periods in time:

How has biofuel research changed from 2000 to 2010? Another important problem, which is addressed in this work, might be the need for comprehending how external factors might impact research. For example, does the price of oil affect the research in any important way? After the macro level, countries become the object of focus.

The **meso level** of the analysis serves as a tool that assists in understanding the technological landscape by taking countries as units of research. This is particularly interesting for policy makers trying to make sense of, for example, what characterises countries in the research landscape. Here, questions such as: "How can a country be characterized in terms of its scientific capability?" or, "Are any countries related from a research perspective?" will be addressed. Other questions related to collaboration and similarity between countries might be critical in policy making, and will be addressed in this project.

The **micro level**, seeks to assist organizations. Between countries and organizations, even if the scope is different, the types of questions are highly connected. For example, as a president of a university one might want to understand what are the best options in business terms, the university should collaborate with.Moreover, one might want to comprehend the clusters of universities and organizations that exist throughout the world. This can help organizations and universities understand the underlying landscape of innovation and research strategies, and accurately design their own strategy.

In the **system level**, the understanding of the particularities of the system can benefit stakeholders at the national, organizational or even research level. These problems, although being more general, are highly relevant. For example, is sugarcane more researched or more patented? Is wood more patented or researched? Moreover, what are the external factors that affect research and do the price of goods influence it in any way?

Helping stakeholders make decisions at different levels and understand the research landscape is the primary goal of this work. It is believed that data, in its most pure form, has advantages not only in terms of the quality of the decisions but most importantly in the transparency of those same decisions.

2.2 Literature review

There are multiple possible ways and angles of studying all of the underlying factors that appear when studying knowledge as a system. In order to stay in scope of the goal of the project, three main areas will be covered in this literature review. The first area, corresponding to the macro and meso levels focuses on **(1) the study of innovation and innovation measurement at the global and national level**, which is, taking countries as the units or entities of the system and studying the global system over time. One should point out the intense relationship between the macro and meso level in terms of literature. Most macro level analysis also focus on the evolution of the global system, and for this reason, these levels will be merged for the sake of the literature review.

The second area of the literature review chapter focuses on **(2) open innovation** as a core strategy in organisations. The third and final part of the analysis focuses on the **(3) engineering systems perspective**, as a core part of the analysis, this perspective could be applicable to both organisations and countries. These three areas have some intersecting concepts, for example, many regional innovation studies focus both in countries and organisations. These areas are illustrated in Figure 2.1.

2.2.1 National Innovation Systems

The macro level of this literature review starts by focusing on the field of research of National Innovation Systems (NIS) as the more classical approach of understanding innovation as a quantifiable characteristic of a certain region, in this case, a country. Although a lot of research has been developed on this topic, one might argue that its pillars come from three distinct areas of research: growth theory (Romer 1990) which states that an increase in productivity (due to innovation) leads to growth of a country's economy; the cluster based theory (Porter 1998) which argues that in a global economy *"one would expect location to diminish in importance. But the opposite is true."* and finally, research on the innovation systems (Nelson 1993) of nations carried by Richard.R.Nelson where he famously describes the innovation systems of 15 both high and low income countries. Nelson argues, for example, that differences between countries in terms of innovation systems are closely related to the differences in economic and political circumstances of those same countries.

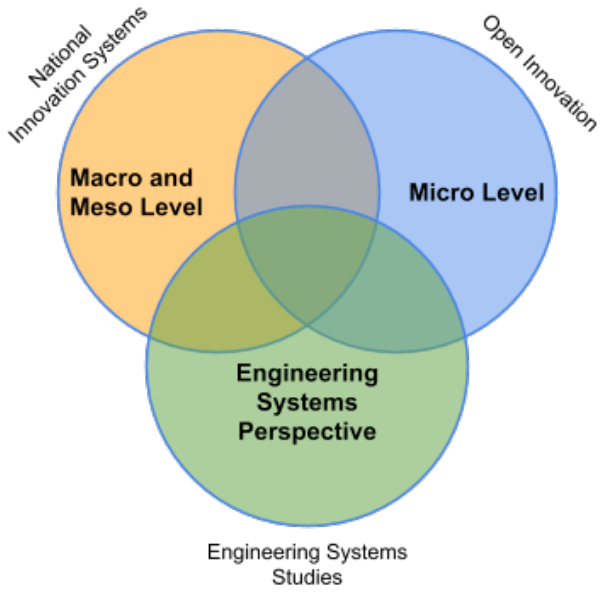


Figure 2.1: Literature review Venn diagram.

Taking these three perspectives, a new field emerges that wishes to further understand the underlying factors between divergences in innovation across countries and at the same time its quantification it in a satisfactory way. This is done by the introduction of the concept and framework of National Innovative Capacity (NIC) (Furman, Porter, and Stern 2002) as *"the ability of a country to produce and commercialize a flow of innovative technology over the long term"*. By using international patenting data from 17 countries members of the Organisation for Economic Co-operation and Development (OECD) the authors managed to determine that the patenting volume was well characterized by a set of indicators. First, that while the discrepancy between countries is due to the difference between research and development (R&D) resources and spending, another very important factor is the R&D productivity. By R&D productivity it is meant policy choices, share of academic and private research, as well as specialization of a certain country. Moreover, other general indicators such as population, gross domestic product (GDP) per capita, and openness to in-

ternational trade and investment also proved to be highly related to this volume of patenting. This paper and research is considered to be the core of the NIC area and gave birth to a number of derived studies on this same topic.

The dynamic innovation system framework (Castellacci and Natera 2011) (Castellacci and Natera 2013) showed that absorptive (human capital, trade and infrastructure) and innovative capabilities drive the NIC of a certain country. Studies such as "Innovation capabilities of European nations" (Faber and Hesen 2004) test the framework on 14 European Union countries and show that patents depend on the sales of product innovations but that at the same time, while some innovation indicators depend on the same economic indicators, governmental regulation and firm specific conditions may also affect innovative output. Some other studies (Filippetti and Peyrache 2011) work on developing a composite indicator that consists of patents, R&D resources, personal computers, internet users and others to find that there is a convergence in technological capabilities is occurring.

Although the classical view has certainly produced more than enough interesting studies on how to measure, characterize, and quantify the innovation capacity of a certain country or group of countries, one should not discard the critical role that the economic dimension might play in the understanding of a country's productivity.

Economic complexity (Hidalgo and Hausmann 2009) is a framework that interprets economic growth and development that stresses the importance of the complexity of the country's economy. It does so by utilizing a network structure and connecting countries to products they export. The characteristics of this network are highly correlated with the GDP of a certain country or even its potential economic growth. Moreover, the authors also argue that the more economically complex, or product diverse, a country is, the more economically complex it becomes. This view is particularly interesting not because of the economic growth per se, but because of the different methodology in the quantification of the assets of a country. For example, instead of taking the volume of patents (such as in the classical view in the beginning of the segment), this view emphasizes the idea of a network structure.

In what regards cluster based theory and the proximity of countries, recent studies have also taken advantage of a network based view. Recent work (Bahar, Hausmann, and Hidalgo 2014) has shown for example that a country is 65% more likely to export a certain product if it possesses a neighbour country that exports that same product, or that the growth of exports is 1.5% per year if its neighbour has a comparative

advantage in this product. However, as distance increases, this probability tends to decrease as well. Hence the title 'Evidence of knowledge diffusion?'. This study also shows the importance of interpreting countries and products as pairs and not as isolated entities, and the value of network interpretation.

Finally, one can conclude that the research field of regional studies and national innovations systems is very heterogeneous. On the one side, the classical view tries to approach a perfect indicator that can perfectly describe the innovation capability of a certain country and all of the external and internal factors that might influence it. On the other side, the economic view, where there seems to be an increasing importance in the network view of the international innovation system, and indicators that economic development and innovation are highly related to the characteristics of this network.

2.2.2 Open innovation

Innovation is not a new field, particularly in what regards scientific research. The dictionary defines it as simply being the "introduction of something new" (Merriam-Webster Dictionary 1828), others (Freeman, C. and Soete 2000) define it as the design manufacturing, management and commercialization of a new process or equipment. However, when observing the extensive list of possible definitions on classical literature, a pattern emerges: they all give equal importance to not only invention, but to the exploitation and development of unseen knowledge.

One particular field of innovation that quickly became of interest to a wider audience and that is also a special focus of this project is the field of so-called "open innovation". It started by being described as a model where companies commercialize both their own innovations and innovations from other firms. Moreover, in this model, companies should find ways of bringing in-house ideas to market by building bridges outside their current business (Gassmann 2006). This field evolved over time into what can be described as 8 main areas of research (Giannopoulou et al. 2010): the concept of open innovation, organizational design and boundaries of the firm, open strategy, the human factor in open innovation, communities and distributed co-creation, patenting and appropriation, the innovation intermediaries' model, and the triple helix model. In the scope of this project, it was deemed appropriate to focus on three of these fields, which will be reviewed below.

2.2.2.1 Open strategy

Open strategy is one of the main fields of research on open innovation which has highly impacted the way firms and organizations operate. For example, companies are increasingly adopting innovation ecosystems across countries in order to match the escalating demand for innovation from customers (OECD 2008). Other findings in this OECD report come to find some characteristics of firms that are highly connected to the scope of this project such as the fact that "large companies are four times more likely to collaborate", that geographic proximity plays a crucial role in the development of these networks, or that firms usually rely on external sources to develop technologies and processes outside their core competencies. Furthermore, it was also discovered that universities and public research institutions have an increasingly important role in this strategy, that staff mobility is crucial, and that national R&D programmes should be as open as possible.

In this scope, some work (Morgan and Finnegan 2008) in the field shows the importance of collaboration with universities, research institutes, communities and governments in order to create and exploit knowledge. Other works (Bessant 2008) demonstrate other strategies firms use such as scout sending, web usage, and work with customers (in this case active users), to achieve a higher level of innovation.

2.2.2.2 Patenting and appropriation

In the same OECD report cited earlier, research shows that industries where intellectual property rights (IPR) are highly protected, companies mainly look outside of the business to keep up to date with research. On the other hand, industries where IPR are "softer", companies mainly employ collaboration as a way of attaining that same objective.

Although normally one possibly think of patents as something that might throttle open innovation and the sharing of knowledge, and that they directly contrast with something like open source software development, research (Pénin and Wack 2008) has shown that adequate use of this system of IPR protection can positively impact the preservation of freedom of access to research tools. Pénin and Wack do this by giving examples on how a certain patent might require a certain technology to be used by all but only under certain special conditions, thus leading to the creation of an environment where: *"a material or invention can be improved by the ideas of many,*

but access is maintained for all who agree to the terms, without exclusive capture by anyone".

2.2.2.3 Triple helix: industry, academia and government policy

The triple helix is a non-linear innovation model that describes an industrial society as something that has shifted from industry-government to an industry-government-university relation (Lewontin 2000). It was first developed by Etzkowitz in the 1990s. Etzkowitz states that knowledge and universities play an increasingly important role in the development of technology and technology based firms (Etzkowitz and Leydesdorff 2000). Moreover, he argues that this interaction is crucial in order to improve conditions for innovation. With the emergence of globalization and its decentralization, regional university networks will fuel innovation by creating "discrete pieces of intellectual property".

Perman and Walsh described the university-industry relationships by establishing a list of possible links between them that include research partnerships, research services, academic entrepreneurship, human resource transfer, informal interaction, commercialization of property rights and scientific publications (Perkmann and Walsh 2007). These are also organized into a hierarchy of relational involvement that ranges from high to low. The researchers also point to the crucial importance of these relationships in the context of open innovation.

On a final note, there is evidence (Léon 2007) that countries and international organizations such as the European Union have already realized the importance of these relations, by creating specialized knowledge transfer structures such as research centers to foster the exchange of knowledge and catalyse innovation. Léon also illustrates the benefits of government-industry-relations by taking the example of grid service deployment and the long and short term instruments developed by the EU to foster these same relationships.

2.2.3 Engineering systems perspective

Engineering Systems is undeniably growing as a research field, the term was reportedly born in the Bell Laboratories in the 1940s, ten years later, G.W.Gilman - then director of systems engineering at Bell - made the first attempt at teaching it in

the Massachusetts Institute of Technology (Hall 1962). During the next decades, systems engineering, its tools, definition and implementation "continued to evolve" (Brill 1998). Among a vast number of definitions, Olivier L. de Weck (Weck et al. 2011) defines an engineering system as:

"A class of systems characterized by a high degree of technical complexity, social intricacy, and elaborate processes, aimed at fulfilling important functions in society."

Piaszczyk (Piaszczyk 2011) developed a conceptualization for engineering systems, where he argued that some domains are common to almost all engineering projects:

- **Environmental**: the external drivers or consequences of the engineering system.
- **Social**: the human components of the system.
- **Functional**: the objectives and goals that the system wishes to achieve.
- **Technical**: the non-human components of the system (assets, information, infrastructure).
- **Process**: the processes that take part in the core of the system.
- **Temporal**: how the system develops or changes over time.

These five domains are critical to be understood in the presence of any engineering system and served as a point of departure that the author used to develop his own **conceptual model of engineering systems**, known as the Engineering Systems Multiple Domain Matrix (ES-MDM). This is not the first conceptual model that was created to categorize engineering systems, other frameworks include the Design Structure Matrix (Browning 2001), the popular House of Quality (Moran 1994), and CLIOS (Complex, Large-Scale, Interconnected, Open, Sociotechnical System) (Dodder, Sussman, and McConnell 2004).

Diving deeper, the notion of complex system emerges, de Weck (Weck et al. 2011) defines a **complex system** as a system where the components, interconnections, interactions or interdependencies are particularly difficult to describe, understand, predict, manage, design or change. In his work, the author also states that a complex

system has not only a technical dimension but also a management and social dimension. Furthermore, he establishes two types of complexity: **behavioral complexity** - where the difficulty lies in the prediction, analysis, description and management - and **structural complexity** where the number of elements and the nature of their relationship are intricate.

However, engineering systems are not only objects of study, engineering systems are also an approach that is used to solve or understand complex problems; it is a "technique for the application of a scientific approach to complex problems" (Miles 1973) that takes a holistic view. This perspective is highly related to the work that will be developed in the forthcoming project. One should stress the different methodologies that have been used to study these systems, in particular, two popular ones.

The first approach that is widely used is the **graph-based or network approach** where system components and relationships are described as networks would be. While the network approach to a system analysis can be strong in its "ability to visualize and perform statistical analysis on the properties of the network and isolate particularly interesting or important system elements or clusters of elements that may be present" (Weck et al. 2011), this approach can also become quickly overwhelming. For example, Eppinger states that "A boxes-and-arrows depiction of the design process for a car's suspension, for example, would run to more than 30 pages." (Eppinger 2001). This approach allows the study of systems as diagrams, flows, and essential visual representations but also allows the study of a system through network analysis in some systems, through different network indicators such as degree centrality, clustering coefficient, degree, and others (Albert-László Barabási 2016). One popular application of this approach is the Program evaluation and review technique (PERT) diagrams, which can be used for example, to visualize a power system restoration (Assis Mota, Mota, and Morelato 2007).

The second approach is the **matrix-based approach**. One can generally represent a network as a matrix using the adjacency matrix of a network, or its co-occurrence matrix, where A_{ij} is equal to 1 if node i and j are connected. One popular application of this approach is the Design Structure Matrix (DSM) where each task or element is laid out in rows and columns, and one can visualize the information and sequential dependencies of the entire project (Eppinger 2001). Others such as the derived domain mapping matrix (DMM) can combine domains to show interdependencies across domains and synchronize several inter domain dependencies

(Danilovic and Browning 2007). This matrix representation is convenient because of the easy "manipulation by the tools of linear algebra" (Weck et al. 2011) and allows the understanding of the more general characteristics of the engineering system.

2.3 Analytical and conceptual approach

The final part of this chapter seeks to understand the viability of using an engineering systems approach as a methodology of describing the macro, meso and micro levels of research. As previously stated, one of the most important challenges of this project is the understanding of the dynamics inherent to both levels of analysis, an interesting question is: Is it a viable strategy to analyse both the macro, meso and micro levels with the same engineering systems conceptual approach?

Let us revisit the macro, meso and micro level literature with an engineering systems state of mind.

2.3.1 A systems approach to the world and country levels

When analyzing the macro (world) or meso (national) levels, in the roots of the National Innovations Systems research area, lies the work of Richard R.Nelson (Nelson 1993), here he justifies the term "system" as:

> *"[...] The concept is of a set of institutions whose interactions determine the innovative performance, in the sense above, of national firms. There is no presumption that the system was, in some sense, consciously designed, or even that the set of institutions involved works together smoothly and coherently. Rather, the "systems" concept is that of a set of institutional actors that, together, play the major role in influencing innovative performance. [...]"*

Several elements allow a direct connection between this approach and a pure engineering systems approach: there are several elements (institutions), which have a relationship between each other (interactions). Moreover, this citation might even lead us to a certain complex system due to the inherent randomness of its design.

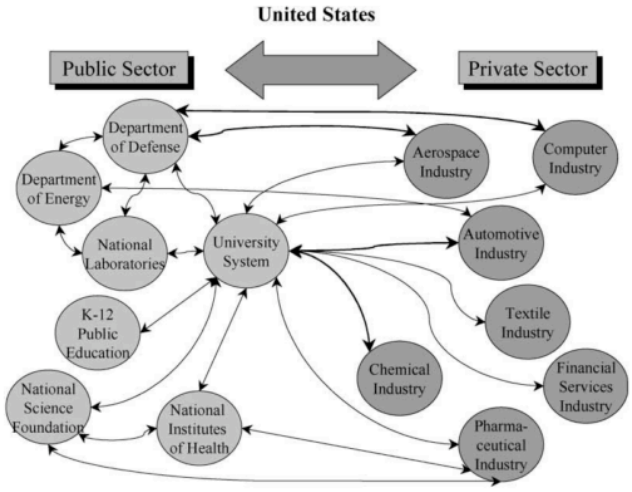


Figure 2.2: Important Elements of national innovation systems of the United States (Furman, Porter, and Stern 2002).

Another clear example is in the work of Jeffrey L. Furman, which describes (Furman, Porter, and Stern 2002) the important relationships and connections between the institutions of a country as a base to the National Innovative Capacity framework. An example of this concept can be seen in Figure 2.2.

2.3.2 A Systems Approach to the organizational level

When analyzing the micro, or organizational level, the compatible system oriented literature is not as obvious. Open innovation research is horizontal to an important number of areas that focus on management, research and development, product development, strategic thinking and more. Perhaps the most system-compatible area of open innovation is the Triple Helix framework (Etzkowitz 2003). In this area, universities, organizations and the government are all involved in an open innovation system where their interactions are crucial to advance the technological field; in fact these are described as “tri-lateral networks”. This is illustrated in Figure 2.3.

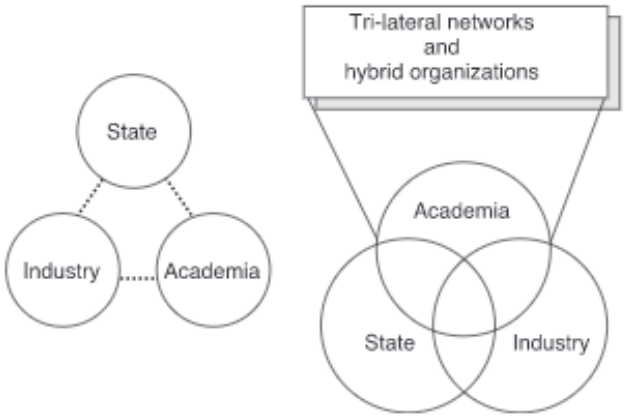


Figure 2.3: Triple-Helix illustration (Etzkowitz 2003).

The work of Perman and Walsh (Perkmann and Walsh 2007) takes the triple helix further by describing the different nature that these relationships, or edges that these elements, or nodes, might have between them.

2.3.3 System compatibility

After dissecting the possibility of applying an engineering system conceptual approach to both levels of the analysis, one should analyse if they in fact constitute an engineering system. This would validate the possibility of applying the methodology to both levels. Let us take the definition of a more complex engineering system, known as complex systems (Sheard and Mostashari 2009). Complex Systems:

- Have many autonomous components.
- Are self-organizing.
- Display emergent macro-level behaviour.
- Adapt to surroundings.

Let us understand how the macro, meso and micro level fit in this definition, taking innovation by the use of technological assets as an example in Table 2.1.

As one might note from the table above that these systems appear to be highly related, one could even say they are different scales of the same system, or even a system-of-systems (Sheard and Mostashari 2009). It is also apparent that the micro level system highly influences the macro level system and vice-versa. For example: a very strong relationship between a university of country A and a university of country B, will lead to a strong relationship between country A and country B, it is simply a matter of scale. One could even go further and try to understand the nano-level, which would consist of analysing a particular university or organization as its own system.

Finally, generally speaking it seems possible to apply the same engineering systems perspective to the macro, meso, and micro level. However, this is highly dependent on the type of analysis; for example, some particularities of the organizational level system might not be reproducible in the meso level system and vice versa.

Complex Systems Definition	Meso/National Level	Micro Level
Autonomous components	Countries, policies, organizations, universities, technological assets (patents, publications, projects), researchers	Universities, businesses, organizations, technological assets
Self-organizing	Countries form partnerships and international organizations, such as the EU or OECD.	University-University partnerships, University-Organization partnerships, Organization-Organization partnerships.
Macro-level behaviour	The types of relationships are not determinable from the type of technological asset being analyzed	The types of relationships are not determinable from the type of technological asset being analyzed
Adapt to surroundings	Research direction might inherently be related to the global sustainability concerns at that particular time.	Research direction might inherently be related to the global sustainability concerns at that particular time.

Table 2.1: Application of the complex system definition to Macro, Meso, and Micro level.

CHAPTER 3

Method

In the third chapter of this project, the background related to the datasets, the data models used, and the tools and methods used to analyze them will be covered. The high-level contents of this chapter are:

- Description of data set and data model: Diving deeper into the data model this part of the chapter will introduce the database used, its architecture and intricacies.
- Network/Graph model: The description of the graph model will give an insight into the transformation of the original database into a network representation of an engineering system.
- Quantitative analysis: The final part of this chapter will give insights into the quantitative tools used throughout the analysis.
- Tools and systematic approach: description of high level tools.

3.1 Description of data set and data model

The first part of the chapter describes the original data model used, provided by the AMICa pathfinder project. This explanation is an essential starting point of the analysis since the representation of the engineering system under analysis is a reflection of the way the original data was modeled.

Original Data Sources

In what regards the original sources of data, the AMICa pathfinder project focused on gathering the following types of data as a representation of the biofuels research

ecosystem:

- Research Projects: Through the European Commission's community research and development information service (CORDIS) (CORDIS 2018), several research projects were retrieved.
- Patents: the OECD REGPAT (OECD 2018) Database provides patent data linked to geographical regions.
- Scientific Publications: the Crossref (Crossref 2018) database was used as a source of scientific publications.
- Industry Facilities and Organizations: with the goal of accessing the names and specifications of organizations and facilities, several sources were used such as ETIP Bioenergy, Biofuels Digest, reegle and more.
- Knowledge graph and data reconciliation services: the Global Research Identifier Database and DBpedia were used to enrich and reconcile the original data

After compiling information from the sources above, all of the data had to be preprocessed using the open source software OpenRefine (OpenRefine 2018) with the goal of finally building the graph database in Neo4j. An overview of the sources can be seen in Figure 3.1.

From a theoretical model to a data model

Originally, the data extracted from the above sources came in a format that was, to say the least, hard to work with. A mix of explicit information and expert knowledge related to biofuels research. With the goal of extracting the knowledge structure behind such data, it was decided that each knowledge asset (patent, publication, project) can be looked at as a combination of the following categories of terms:

- **Feedstocks**: the raw material used to fuel a machine or industrial process (eg. wood, sugar, corn, waste, etc.)
- **Processing Technologies**: The processes that feedstocks undergo in order to achieve a desired output, which in the case of biofuels are mainly chemical processes (fermentation, gasification, catalysis, etc.).

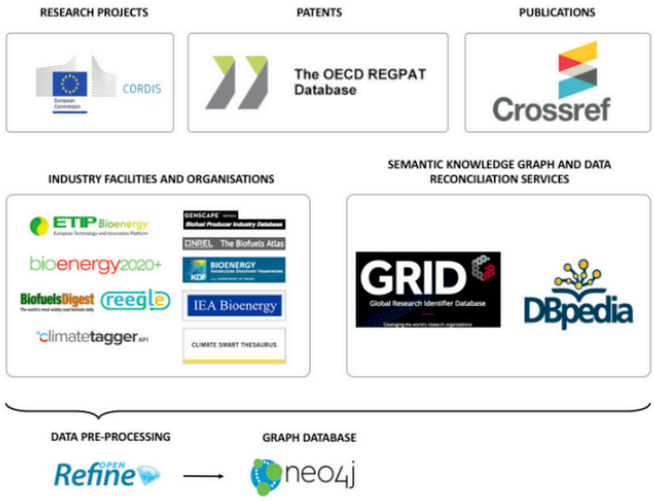


Figure 3.1: Overview of data sources from the AMICa pathfinder technological briefing (*AMICa, 2018*).

- **Outputs**: The result of feedstocks undergoing a certain process, or in other words, the result. (Biogas, ethanol, biodiesel, etc.)

It is quite clear that each piece of knowledge extracted from the database will very likely contain one of the above types of terms. For example, there might be 4 patents in the database that contain the term "waste" and 17 scientific publications that contain the term "fermentation".

However, a bag-of-words approach to this data might be too simplistic. At the end of the day, feedstocks, technologies and outputs possess an intricate relationship between them. And these relationships is information which should be preserved. Through the development of the project, the project participants understood that the interesting aspect of these terms was to study the combinations of these terms:

"A key finding of our meetings and literature review is that exploring all relevant combinatorial possibilities between potential feedstocks (e.g. microalgae), processing technologies (e.g. microwave-assisted transesterification) and outputs (e.g. biodiesel), is a difficult but crucial task in the development of new sustainable biofuels." - Amica technical brief (*AMICa, 2018*)

This 'alternative' approach is more useful in the project context, remember, one of the goals of the project was to understand if there might be untapped areas of research. Let us take the following as an example: we are in the presence of feedstock F1 and feedstock F2 and processing technology PT. There is a high frequency of assets with the F1/PT pair, but a low frequency of the F2/PT pair. This might mean that F2/PT is an untapped area of research. Such conclusions would be much harder to reach if the focus was not in the combinations.

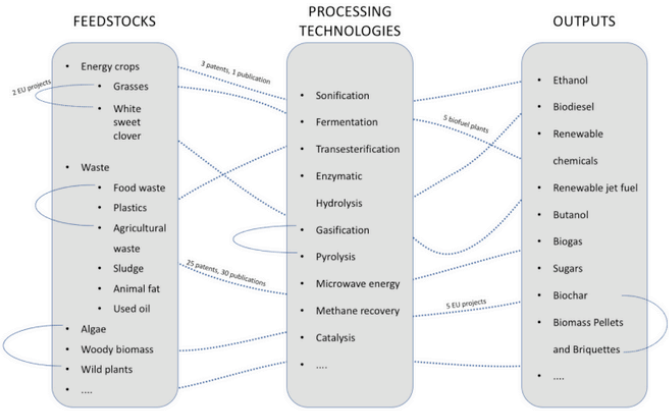


Figure 3.2: Combinatorial model of the AMICa pathfinder model.

3.1.1 Database Architecture

One of the main reasons why a graph database was adopted for this project was its capacity of preserving these important relationships while, at the same time, maintaining an acceptable response time and scalability. Some of the elements that make for this database follow.

- The first group of elements, or the nodes are the **asset nodes**. These come directly from the database; there are 5 types of asset nodes: patents, projects, facilities, organizations, and publications. Each one of these types of nodes contains attributes such as year, owner or abstract.

- The second group of elements are **asset attributes**, these are loose elements that possess an intricate relationship with asset nodes. Some of these are coun-

tries, locations, types or asset owners. These attributes can be understood as metadata for the asset nodes.

- The final group of elements, and perhaps the most relevant to for the project in scope, is the group of **terms**. Process terms are the various outputs, processing technologies and feedstock terms that appear in assets.

All of the above groups have relationships to each other, for example, a patent might contain several feedstock terms and output terms. Since an extensive description of these relationships would be out of the scope of this thesis, Figure A.1 provides a simplistic overview of the different relationships these nodes possess with each other, as well as the attributes of each node.

3.2 Network/Graph model

Model Criteria

In the third part of the chapter, the focus will lie in the description of the engineering systems representation that was developed. The developed model would have to respect a number of important criteria:

- Accurately **portray the biofuel research landscape** dynamics.
- Preserve the **complexity of term combinations**. Simply describing the frequency of a certain term throughout documents would not be enough.
- Be **scalable to different dimensions**, particularly the global level, the country level and the organizational level.
- **Statistical analysis and pattern detection should be possible** in order to detect patterns and answer quantitative questions.

One possible way of respecting these constraints with the available data goes back to the graph/network system approach. As previously stated, one of the **advantages of this approach relies in the visualization and statistical analysis of a particular system**. But the graph system approach is not per se a mathematical definition of how a system should be expressed; however, one of the possibilities that

appear in the literature is the creation of a matrix that expresses the frequency of pair combinations.

Adjacency Matrix

An adjacency matrix is a mathematical term that seeks to express a network through a matrix. Let us mathematically define it as:

$$A_{ij} = x \tag{3.1}$$

- The value of x is equal to 0 if node i and j are not connected.
- The value of x is proportional to the weight of the edge between node i and node j.

The adjacency matrix is purely symmetrical:

$$A_{ij} = A_{ji} : \forall_{i,j} \tag{3.2}$$

In this case, the connection of a node with itself is discarded, therefore the diagonal of the adjacency matrix is equal to zero if a:

$$A_{ij} = 0 : \forall_{i,j:j=i} \tag{3.3}$$

Model

The model proposed is based on the adjacency matrix; it treats every term (feedstock, output, processing technology) as a node in the network. If two terms appear in the same document, then these nodes are connected. The weight of the edge connecting these nodes is proportional to the number of times that these appear together throughout the database. For example, let us say that “fermentation” and “corn” appear together in assets 45 times in a database, then, the edge that connects them will have a weight of 45.

In the model, the rows and columns of the matrix are all of the process variable terms that exist in the database (in total 342 distinct terms). Which will create a 342x342 adjacency matrix, with a total of 116964 values, of which 58 482 are unique (due to symmetry). This is represented in Figure 3.3.

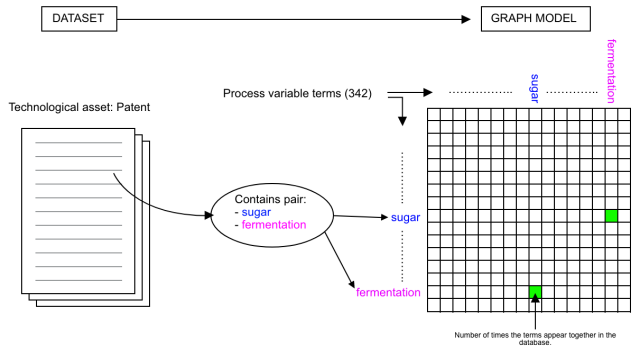


Figure 3.3: Graphical Representation of the graph model.

This graph model is highly powerful not only because of its capacity to capture all of the pairwise combinations of terms but also because of its scalability. In fact, **one can apply this model to several levels of the analysis by pre-filtering the technological assets used to build the adjacency matrix**. This process is illustraded in Figure 3.4.

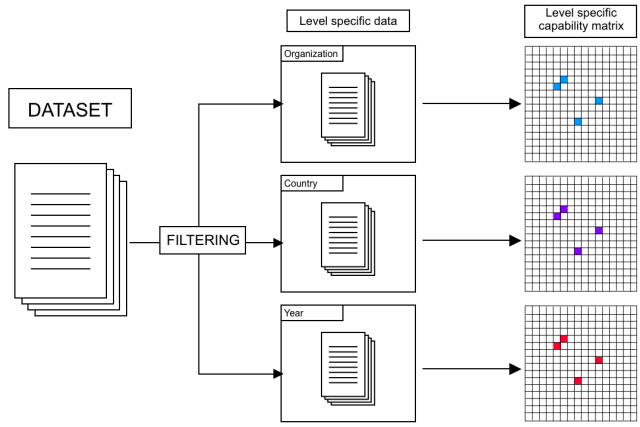


Figure 3.4: Illustration of filtering and scalability of graph model.

If only documents from a particular year, country, or organization serve as input, then the subsequent model will represent the technological capability of that partic-

ular unit, or its capability matrix. It is important to note that whatever the filtering (per organization, per country) applied to the data, the capability matrix will always have the same dimensions; this is because the extraction of all of the process variable terms was made beforehand. This allows the application of the same quantitative analysis techniques to all of the levels of the project.

Finally, looking back at the graph model constraints established at the beginning of the section, it can be said that this model appears to respect them entirely. By representing the entirety of term pairs used across documents, it not only represents the research landscape, but also preserves the scientific complexity of such research. Moreover, the pre-filtering of documents allows for scalability of the model, and the fixed matrix structure allows for the quantitative analysis.

3.3 Quantitative Analysis

During the analysis, several measures had to be normalized in order to focus on patterns rather than volume of occurrence. For example, the United States of America is the country that produces the most patents in the world and therefore, if one wants to compare its capability matrix to the capability matrix of another country, the results should be normalized:

Let us define the **normalized capability matrix** as:

$$B_{ij}^{k} = \frac{A_{ij}^{k}}{N_k} \tag{3.4}$$

Where, A_{ij}^{k} is the capability matrix of a certain category k, B_{ij}^{k} the normalized capability matrix of that same category, and N_k the total number of technological assets owned by or located in that category.

Another important normalization is related to the chronological evolution of records. There is a tendency for an increase in the number of technological assets over time, and therefore if the study is focused on the proportion of technological assets of a particular type in a year, a normalization should also be applied:

$$T_{Norm}^{k} = \frac{T_{Abs}^{k}}{N_k} \tag{3.5}$$

Where T_{Norm}^{k} is the **normalized number of records** in k, T_{Abs}^{k} is the absolute number of records in k, and N_k the total number of records in that period k.

Capability Matrix Operations

The second important quantitative analysis operation is related to comparing capability matrices of different entities. In order to compare two or more capability matrices, a python script was developed to transform a capability matrix into a list:

```
def get_list_from(matrix):
    only_valuable = []
    extension = 1
    for row_number in range(matrix.shape[0]):
        only_valuable.append(matrix[row_number, extension:matrix.shape[0]].
            tolist())
        extension += 1
    return [element for column in only_valuable for element in column ]
```

Listing 3.1: Python script to transform capability matrix into list.

This script takes all of the unique values of a capability matrix (58482) and places them in a vector with a total of 58482 entries which will be referred to as a capability list.

In order to compare two capability matrices, the first step is to create two capability lists, after doing this, the **Pearson correlation index** was chosen as an indicative of the similarity between them:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \tag{3.6}$$

Where r is the Pearson correlation index, a value between 0 and 1. The program built also returns the p-value of the correlation as a measure of the confidence of the correlation between two capability matrixes.

Correlation Matrix

Another important concept when comparing capability matrixes is the concept of a correlation matrix.

When comparing units (countries, years, organizations), in order to understand how there capability matrices differ, the Pearson correlation index was used (as above described). Let us say that there are a total of 100 countries to compare. By comparing every country to each other, a correlation matrix between all of the countries can be created. Figure 3.5 illustrates this concept.

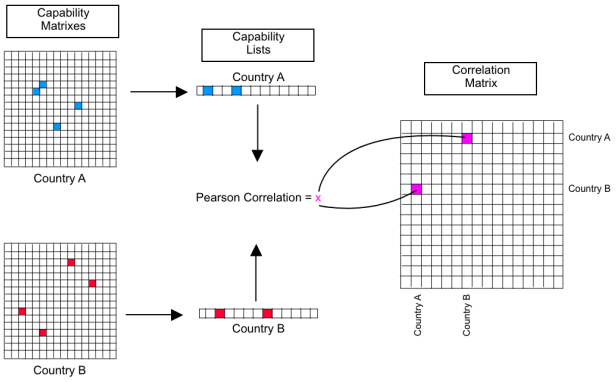


Figure 3.5: Correlation matrix creation workflow.

Hierarchical Clustering

In order to find the clusters of entities whose capability matrices are more similar and consequently find "capability groups" the main technique used derives from unsupervised learning: hierarchical clustering. This method seeks to form a hierarchy of clusters from data points and is usually expressed through a dendrogram of connections.

This method was applied by the use of the scipy python library which also allows for the determination of a particular hierarchical clustering method. A hierarchical clustering method is usually defined by the distance measurement that helps in building clustering, such measures include: ward, single, complete, average. Through trial and error, it was determined that the average distance method was the most appropriate for the analysis.

Collaboration

Another important measure that was used throughout the analysis is the relationship between countries, universities or organizations. To quantify this measure, as previously described in the triple-helix framework in the literature review, the properties of the graph database were used.

```
colabQuery = """ MATCH (a:Asset)-[:LOCATED_IN]->(ac:Country)
            MATCH (b:Asset)-[:LOCATED_IN]->(bc:Country)
```

```
3             WHERE a.id = b.id AND ac.name <> bc.name
4             RETURN ac.name, bc.name ,count(a.id)
5             ORDER BY count(a.id) DESC"""
```

Listing 3.2: Collaboration query.

Using the cypher language, it is possible to retrieve technological assets that are located (in the case of countries), or owned by (in the case of organizations), by two different entities. For instants, when comparing France and Denmark, if the query above returns 45 documents, this means that France and Denmark collaborated in 45 different technological assets. The same logic applies to organizations.

External data and considerations

Throughout the analysis, several external data was used, mainly for considerations regarding the relationship between these externalities and the patterns detected in the data treated.

The first data was the **Gross Domestic Product** (GDP), extracted from the World Data Bank (World Bank 2017). This indicator was extensively used in national innovation studies previously mentioned in the literature review, and serves as comparative index of economic development between two countries. However, in order to preserve the scale of the analysis, the GDP per capita was used. Moreover, the GDP per capita average and difference of country pairs was also used. Since only considering the GDP per capita difference could provide a too "simplistic" analysis, the GDP per capita average gives information about the relative richness of the country pair, and not only how different the two countries are.

On a second note, the analysis also sought to relate chronological development of certain technological assets with the evolution of external factors, in order to understand their intricate relationship (if there is one). As a proof of concept, two indexes were used, a more general one and a more specific one. The more general index, **price per gallon of oil** in $US, and the more specific index, **price per kg of sugar**. In the introductory part of the analysis, their evolution over time is compared to the usage of specific terms as a possible way of understanding the behavior of the biofuel research landscape.

3.4 Tools and data management

In this part of the thesis, some of the more technical high level tools are introduced, these are mainly made of programming languages, libraries and protocols that were essential in reaching the goals for the analysis.

Data Storage and Retrieval

The first tool that is worth mentioning is the graph database where all of the patents, scientific publications and projects were stored, work developed in the context of the AMICa pathfinder project at DTU. Graph databases are a special type of database that use network structures composed of nodes and edges, to represent and store data. This is opposed to the relational database model where data is separated into different tables. The advantage of "graph databases" lies in the relationships, and the fact that these are explicit.

The database used is managed in Neo4j, an open source graph database management system that allows not only to easily run a database server in a machine, but also allows the direct interaction and querying of the data.

To query the data, Neo4j requires the usage of a programming language know as Cypher. Cypher can be understood as the graph database equivalent of SQL, and allows for fast relational queries to the data. As a simple example to the cypher language, let us retrieve all of the technological assets located in Denmark:

```
1 MATCH (a:Asset)-[:LOCATED_IN]->(c:Country)
2 WHERE c.name = "Denmark"
3 RETURN a.type, a.owner, a.title
```

Listing 3.3: Using the Cypher language.

A snippet of the response is represented in Figure 3.6.

"a.type"	"a.owner"	"a.title"
"PATENT"	"MAN DIESEL & TURBO AF MAN DIESEL & TURBO;MAN B & W DIESEL AS;MAN DIESEL & TURBO AF MAN DIESEL & TURBO"	"Fuel valve for injecting liquid fuel into combustion chamber of self-igniting internal combustion engine used in propulsion system of ship, has axially displaceable shaft provided with closure unit partially placed in chamber in nozzle"
"PATENT"	"BONDE T A"	"Bioenergy production e.g. fermentation liquids separation technique consists of addition of lime and water containing salts for steam stripping of ammonia"
"FACILITY"	"DAKA ECOMOTION"	"Daka ecoMotion"
"ORGANIZATION"	"BIOGASOL -BORNBIOFUEL PROJECT"	"BioGasol -BornBioFuel project"

Figure 3.6: Resulting table from Cypher query.

Data Handling and Analysis

The main programming language used to analyze and handle the data was **Python**, more particularly its 2.7 version. This was not only because of the familiarity of the author with it but also because of its power and flexibility with data handling.

As a way of interfacing with the Neo4j database, the py2neo open source library was used. This is a particularly convenient way of interacting with the original database through the python environment, since it allows one to write cypher queries directly in the python console and to extract data in a convenient format (numpy matrix, pandas dataframe).

For handling the data, an extensive list of python libraries was used, the most important ones follow:

- `Numpy`: An essential package for scientific computing that includes a very good matrix object implementation known as the numpy matrix.
- `Pandas`: Provides an easy to read and handle data structure that is especially useful for table visualizations.
- `Math`: Some mathematical functions are not available out-of-the-box with the python language.
- `Itertools`: A library that allows for efficient looping cycles. For visualizations, three toolkits were recurrently used:
- `Matplotlib`: An easy to use tool for producing visualizations such as graphs, bar plots, and others.
- `Seaborn`: Based on matplotlib, seaborn provides a more visually appealing and statistics focused visualization library.

- `Plotly`: Used to create dynamic - web based - visualizations.

If you wish to download a comprehensive list of the libraries used please visit this link.

Code Management and Presentation

To present the data in a narrative format jupyter notebooks. These notebooks constitute an interesting way of presenting not only the code, but also allow a narrative from for the analysis, which is written in markdown. Notebooks are a popular tool in data science and an easy way of presenting data science procedures.

To store the code for the analysis, **GitHub** was used as a tool for keeping everything in cloud storage. Moreover, by creating a repository for this project, this means that the code is available 24/7 for anyone that wants to consult it or request any modifications.

Repository link: `www.github.com/duarteocarmo/technological_capabilities`

CHAPTER 4

Results

In the fourth section of this master´s thesis, the results of the analysis will be presented. This section follows the structure of the macro, meso, and micro levels, and contains the following sections:

- Macro Level: In the first subsection, the result of the analysis at the macro level will be presented. Here, contents such as the characterization of years, the relationship between them, and the influence of external factors will be studied.

- Meso Level: In the country level analysis, the focus will lie in the country as the unit of scientific capability. Coverage will start from the capability matrix of a country, and will reach topics such as collaboration, similarity between countries, and the role of the GDP per capita in the national innovation landscape.

- Micro Level: At the micro level, the results of the application of the meso level methods to organizations will be addressed. With a special focus on the collaboration types of different organizations.

- Patents and Publications: In the first part of the complementary analysis, results of the comparison between terms used in publications and patents will be shown.

- An alternative approach: Finally, as the closing part of the results section, and as validation of the methodology chosen, results of an analysis focusing on term frequency instead for term-pair frequency will be addressed.

A model of this approach is represented in Figure 4.1.

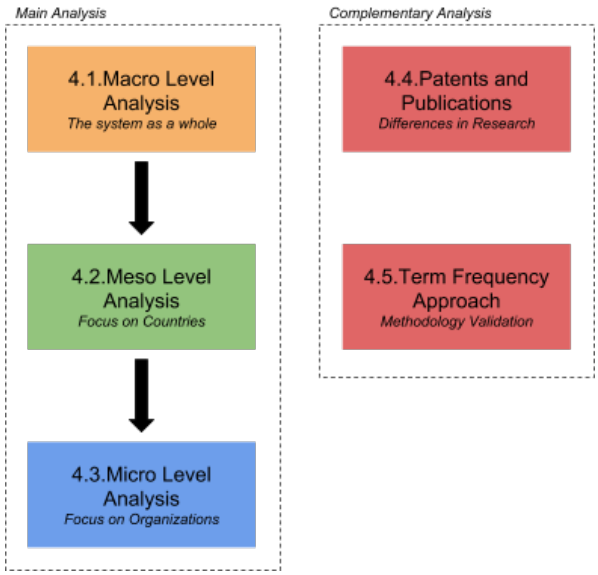


Figure 4.1: Structure of result presentation.

4.1 Macro Level: The biofuel research system

4.1.1 Characterization of Years

Capability Matrices

As a first result, the whole database will be considered without any sort of filtering. The functions built using python, allow the rendering of two different capability matrices: a normalized version and an absolute version (Figure 4.2). Here, one can notice some columns and rows of the matrices that have a higher number of technological assets. These correspond to biofuel related terms that are more frequently used. In general, the structures of both matrices are similar, which is expected since the normalized version is proportional to the absolute one.

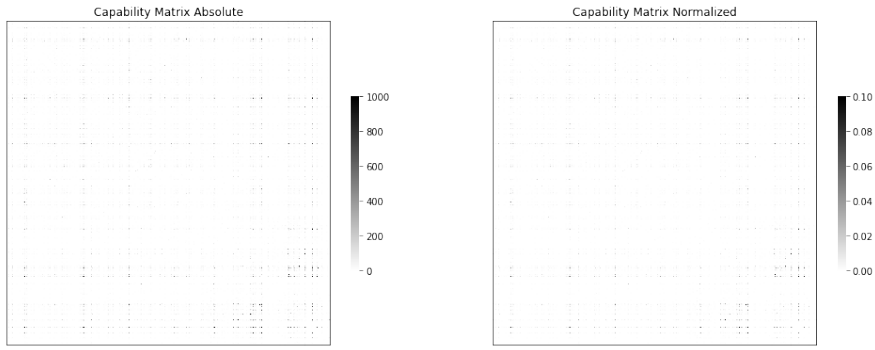


Figure 4.2: Capability matrices of the AMICa database.

As a validation of this approach, the clustered version of this capability matrix was also produced (Figure A.5), although visualization is tough, one can notice two different areas of higher density particularly in the top left and bottom left part of the matrix. While investigating the clustered term pairs, related terms from a scientific perspective stand out, some examples of clusters follow:

- Pal waste, food waste, organic waste, municipal solid waste, industrial waste.
- Sawdust, woody biomass, wood waste.
- Beverage waste, garden waste, brewery waste, biodegradable waste.
- Mixed prairie grass, cereals/sugar, corn/barley, grain, agriculture, agricultural waste.
- Sugarcane, cellulosic ethanol, corn, cellulosic biomass, yeast.
- Rice straw, wheat straw.

From the general clustering, the algorithm tends to accurately separate terms that are somehow related. These relations are a result of the composition of certain feedstocks, the type of derivatives of certain raw materials, or the proximity of certain outputs.

Capability Matrices of years

Taking a year as a unit of analysis, the first step produced is the characterization of a certain year in terms of its technological capability. This methodology was first described in section 3. In this matrix, each row and column represent a term of the dictionary of biofuel relevant terms, each value of that matrix takes the number of documents queried that possess both the term in the column and the term in the row. As a point of departure, all of the documents related to the year 2017 were queried. As a result, the matrix in Figure 4.3 was produced.

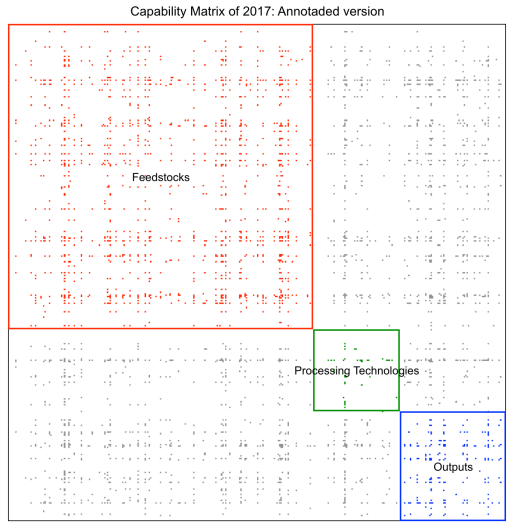


Figure 4.3: Annotated version of the capability matrix of 2017. (Binary).

Here, the visualization is rather challenging because of the large size of matrix, remember, there are a total of 352 rows and columns. Another interesting matrix to produce, because it allows for a fairer comparison of years between themselves is the production of a normalized capability matrix, where the above matrix is divided by the total number of documents present in the database for that year. In the case of 2017, where there are a total of 670 documents (patents, publications, projects), the normalized version of the matrix has the properties described in Table 4.1 and is illustrated in Figure 4.4. One important thing to point out is the number of zero

values in the matrices above. Moreover, there appear to be “areas” where the value is higher.

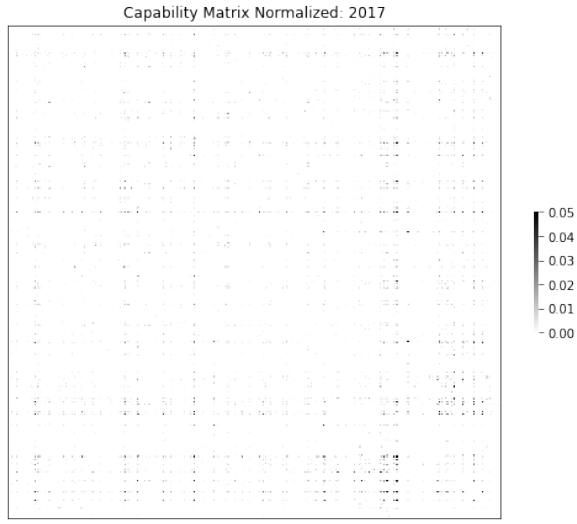


Figure 4.4: Capability Matrix of the year 2017, Normalized.

Capability Lists

After being capable of reproducing a capability matrix for each year, this same matrix was transformed into a vector, or list, utilizing the function previously described in the methodology. Below, a visualization of two years, particularly 2012 and 2013 is provided. In Figure 4.5, each row corresponds to the normalized capability lists of each one of the years.

Property	**Value**
Shape	352x352
Max	0.23
Min	0.00
Mean	2.47e-04

Table 4.1: Normalized capability matrix of 2017.

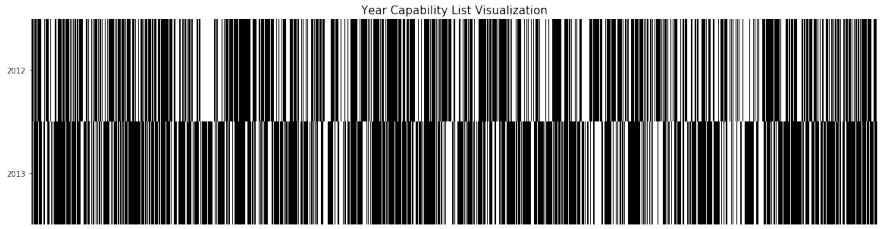


Figure 4.5: Normalized capability lists for 2012 and 2013.

4.1.2 Year Correlation Matrix

After successfully visualizing the capability matrix of a certain year, and characterizing it in terms of an adjacency matrix, or list, the next goal lies in successfully comparing different years between themselves, to do this, we must first choose exactly which years to include in this comparison. In the database, the number of documents over time is not regular, on the contrary, there is a large amount of years with low-to-no documents. Figure 4.6 represents the number of documents per year in the database. Until 1997, the number of documents is almost or even equal to zero, and therefore, it was chosen that this year-to-year comparison would focus on the year of 1997 until 2018.

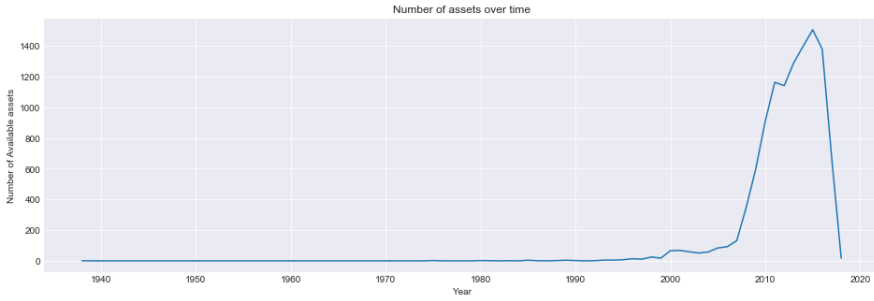


Figure 4.6: Number of technological assets over time.

To compare two years between themselves, the Pearson correlation index was used (as described in the methodology), between their capability lists. For example, the years 2012 and 2013 (as seen in the figure above) have a Pearson correlation of 0.90, which indicates a very high relation between them. Applying the same methodology

to every year between 1997 and 2018, a year correlation matrix can be produced, as seen in the figure below. In Figure 4.7 the lighter the color, the higher the similarity, in terms of research. For example, the year 2003 is more similar to 2000 than to 2002. It can be observed that recent years are on average more similar than less recent years. However, there seems to be some exceptions.

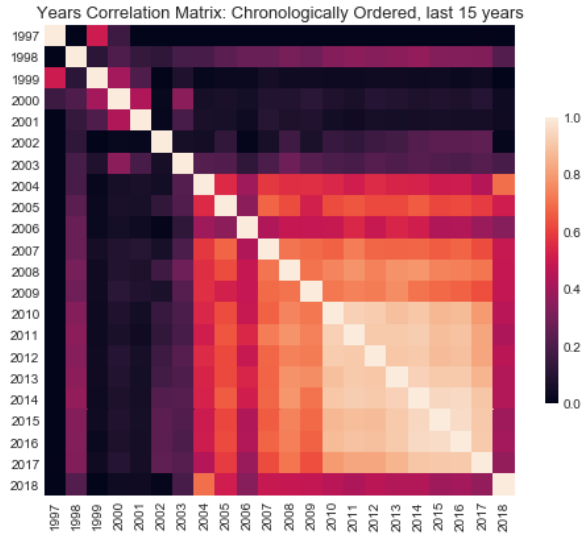


Figure 4.7: Year Correlation Matrix of the last 15 years.

After producing the year correlation matrix, clustering was applied to the matrix as a way of identifying "clusters" of years that are more related between themselves (Figure 4.8). To do this, hierarchical clustering with average distance was the chosen methodology. The application of this methodology led to the figure below where the clustering algorithm orders the matrix in a way that years that are more similar appear closer together. Moreover, one can also notice a dendrogram as a visual aid to that same algorithm. In general, more recent years (2010 onwards) form a cluster of their own. The results of this clustering confirm that from a year-to-year correlation perspective the period 2010-2017 shows few changes, the period 2005-2009 shows a medium level of changes and all the rest of the years are characterized by relatively large year-to-year changes.

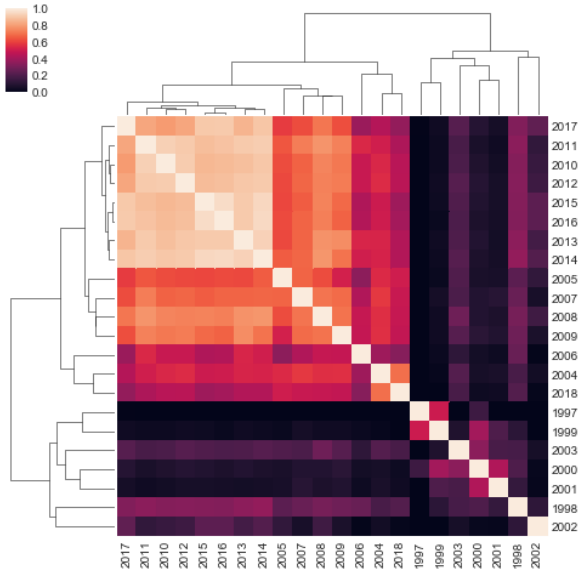


Figure 4.8: Clustered Year Correlation Matrix of the last 15 years.

4.1.3 Correlation of years over time

Following the comparison of all of the years between themselves, it is interesting to understand if the relationships and similarities between those years are at all connected to a chronological timeline e.g. are consecutive years more connected between themselves? To help in assess this question the correlation of consecutive years was also studied. Figure 4.9 represents how one year is correlated with the previous year. For example, 2005 has a 0.5 correlation with 2004, on the other hand, 2006 has a 0.35 correlation with 2005. One can observe a tendency of rise of the correlation of years over time. In other words, more recent years are more related to each other. On the other side, before 2007, the correlation between years follows a less obvious pattern.

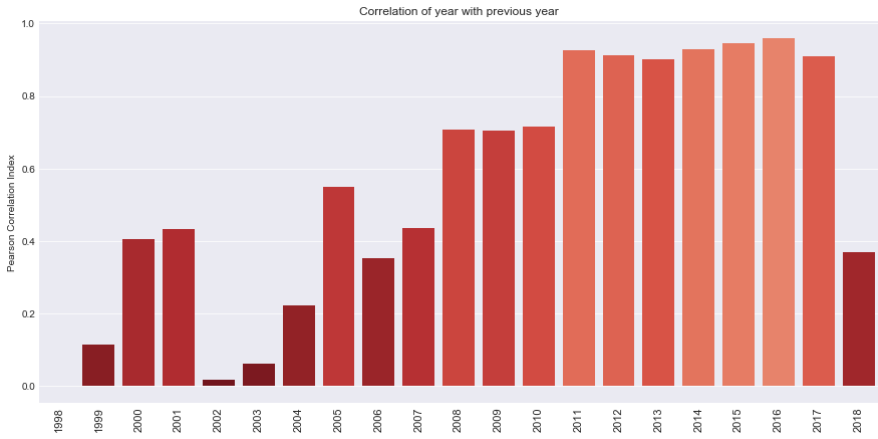


Figure 4.9: Pearson Correlation with previous year.

4.1.4 Comparing Years

The final result of the Macro level analysis concerns the comparison of two years, and particularly the understanding of the intrinsic differences that may result in a high or low similarity between years. Taking as a point of departure, and as a proof of concept, the capability matrices of the years 2017 and 2010, the first step was to build their capability matrices. Due again, to the high number of terms and term pairs, the visualization and understanding just from the visualization of the matrices side by side is rather poor (Figure 4.10).

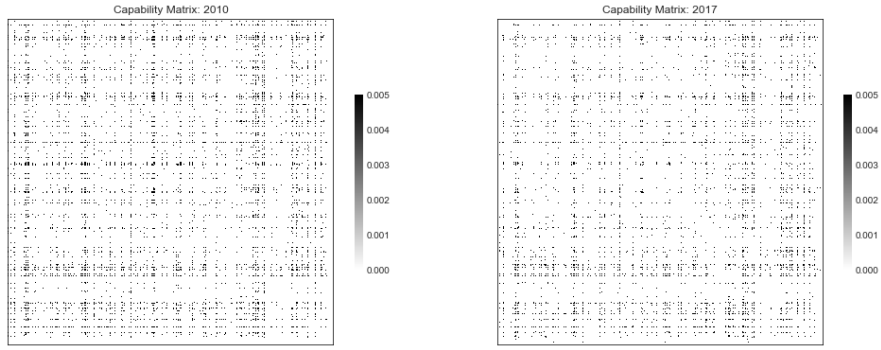


Figure 4.10: Capability matrices of the years 2017 and 2010.

In order to try to visualize the differences and areas of the matrix that differ from one year to the other, taking the normalized capability matrices of both years, the difference between these two was also produced. Knowing that these years have a relatively high Pearson correlation (e.g. they are similar in terms of capability), the matrix of differences serves as a simple way of directly comparing two years.

To understand and compare the years at a term-pair level, the tables of the most frequent terms pairs for the years of 2010 and 2017 were produced (Table 4.2 and 4.3). When observing these tables some factors stand out:

- The number of documents in 2010 is far superior to the number of documents for the year of 2017.
- The pair ethanol-fermentation is the top term pair on both years appearing in at least 17% of all of the technological assets of both years.
- In general, the most used term-pairs are made of output-processing technology terms (e.g. ethanol-hydrolysis) and output-feedstock terms (e.g. waste-ethanol or sugar-ethanol).
- Some pairs seem to diminish in importance, for example sugar-fermentation is important in both years but sugar-ethanol is not present in the top 10 for the year 2017.

First Term	Second Term	Documents	Percentage
fermentation	ethanol	319	0.350935
hydrolysis	ethanol	225	0.247525
transesterification	biodiesel	168	0.184818
anaerobic digestion	biogas	152	0.167217
catalysis	biodiesel	131	0.144114
fermentation	bioethanol	120	0.132013
sugar	ethanol	106	0.116612
sugar	fermentation	102	0.112211
hydrolysis	bioethanol	95	0.104510
enzymatic hydrolysis	ethanol	85	0.093509

Table 4.2: Top term pairs for 2010.

First Term	Second Term	Documents	Percentage
fermentation	ethanol	154	0.229851
anaerobic digestion	biogas	137	0.204478
pyrolysis	bio-oil	101	0.150746
hydrolysis	ethanol	76	0.113433
fermentation	bioethanol	76	0.113433
sugar	ethanol	60	0.089552
waste	ethanol	58	0.086567
sugar	fermentation	57	0.085075
waste	biogas	53	0.079104
fermentation	biogas	53	0.079104

Table 4.3: Top term pairs for 2017.

Finally, as a way of directly comparing these years, a table of the term pairs and their evolution from 2010 to 2017 was created. Here, the main question that was sought to answer was: If term pair A-B was in x% of assets in year X, what was that same percentage in the year X+Y? Moreover, what were the term pairs that most differed in terms of usage between these two years? Table 4.4 is a possible representation of that same question, in it, one can see the term pairs that differed the most greatly between these two years in terms of usage. For example, the term pair "bio-oil-pyrolysis" appears in only 0.26% of the documents in the year of 2010, against 1.15% in 2017:

- The term pairs with the most important differences in usage are not necessarily the term pairs with the most usage in each of the years. With the exception of the pair "ethanol-fermentation".
- Most term pairs contain output terms such as "ethanol", "biodiesel", "biogas" etc.
- There is a relative balance between the number of term pairs that decreased in usage and the number of term pairs that increased in usage.

First Term	Second Term	2010	2017	Difference in %
hydrolysis	ethanol	0.247525	0.113433	0.134092
transesterification	biodiesel	0.184818	0.050746	0.134072
fermentation	ethanol	0.350935	0.229851	0.121084
catalysis	biodiesel	0.144114	0.023881	0.120234
pyrolysis	bio-oil	0.041804	0.150746	0.108942
vegetable oil	transesterification	0.080308	0.007463	0.072845
catalysis	ethanol	0.079208	0.010448	0.068760
transesterification	ethanol	0.089109	0.023881	0.065228
catalysis	methanol	0.061606	0.001493	0.060114
anaerobic digestion	ethanol	0.008801	0.067164	0.058363
gasification	syng	0.005501	0.059701	0.054201
vegetable oil	biodiesel	0.084708	0.032836	0.051873
vegetable oil	catalysis	0.063806	0.011940	0.051866
transesterification	methanol	0.073707	0.022388	0.051319
fermentation	gasoline	0.055006	0.004478	0.050528
pressing	ethanol	0.045105	0.002985	0.042119
solvents	biodiesel	0.033003	0.074627	0.041624
seed oil	biodiesel	0.038504	0.000000	0.038504
anaerobic digestion	bioethanol	0.001100	0.038806	0.037706
anaerobic digestion	biogas	0.167217	0.204478	0.037261

Table 4.4: Top term-pairs with the most important differences in usage between 2010 and 2017.

4.1.5 Biofuel-related terms over time

After analyzing the evolution of the correlation of years over time and providing visualizations that allow the comparing of two years, the study re-focuses on seeing how the usage of different biofuel-related terms evolve over time. To this, the same framework of term division is respected: biofuel related terms can be feedstocks, processing technologies, or outputs. To study their evolution over time and due to the inconsistency in terms of the volume of documents over time, one should focus on the normalized quantity of terms rather than their absolute values. Three different graphs were produced, each related to one type of term (Figure A.2, Figure A.3 and Figure A.4). The terms chosen to represent each group were selected due to their high occurrence in each group.

The same behavior can be generally observed across the three types of terms, until

the year 2000, the normalized quantity of terms is rather "turbulent", while after that year there seems to be a more regularized behavior across the normalized usage of different terms. Moreover, there seem to be some spikes in terms such as "sugar" or "ethanol", which means that all of the documentation related to that particular year in the database contains that same term.

4.1.6 Contextual Relationships

The following section shows the result of the comparison of external factors with the usage of certain terms. To do this, two external factors were used as a proof of concept.

Price of Oil

Taking the evolution of the price of the barrel of oil in $US from `inflationdata.com` (Data 2018), which is inflation adjusted, it was decided to compare how its evolution compared with the relative presence of terms over the years in the database of assets. As a first visual tool, a double axis plot with the normalized usage of three example outputs (biogas, bioplastic, and butanol), and the price of oil from 1990 to 2017 was produced, and can be seen in Figure 4.11.

Figure 4.11: Normalized term usage and evolution of the price of oil from 1990 to 2017.

One can first observe that there seems to be a rise in both the term usage over the years, and at the same time a regular augmentation of the price of oil until about 2014. Moreover, there are patterns that appear in the evolution of the price of oil that seem to repeat themselves in the usage of terms, such as the period of the usage of the term “biogas” between 2000 and 2005, and the evolution of the price of oil between 2003 and 2008.

But a chronological visualization has drawbacks; there is no way of consistently comparing all of the terms and their correlation with the price of oil. To achieve this comparison, the evolution of the price of oil was compared to all of the different terms in the database. For each term, the Pearson correlation between its usage in every year and the evolution of the price of oil was calculated. As a result, a ranking of the terms with the highest positive correlation with the price of oil can be observed in the following table. In Table 4.5, the top 10 terms with the highest correlation with the price of oil are presented. For example, the evolution of the usage of the term “butanol” has an 85% correlation with the evolution of the price of oil, the term “bioplastic” a correlation of 80%. Moreover, a table with the terms with the most important negative correlations was also produced and can be consulted in the jupyter notebook.

Output Name	P-value	Pearson Correlation Index
butanol	1.603189e-08	0.844547
bioplastic	2.463734e-07	0.804599
biodiesel	7.978637e-07	0.784034
fatty acid ethyl ester	1.427601e-06	0.772960
adipic acid	1.048009e-05	0.729790
bioethanol	2.862862e-05	0.704439
syng	3.649295e-05	0.697899
biobutanol	5.140385e-05	0.688369
cellulosic ethanol	1.301892e-04	0.660616
biopolymers	3.263515e-04	0.630094

Table 4.5: Top 10 terms with the highest positive correlation with the price of oil from 1990 to 2017.

Price of Sugar

The second example was based on a more traditional asset, the price of sugar. To do this, the exact same approach as the price of oil was applied but now taking the evolution of the price of sugar over time from the DataBank (World Bank 2017). As previously noted, the general behavior over time in the double axed chart in Figure 4.12 is rather poor in expressing the relationship between the price of the kilo of sugar in $US and terms such as “sugar”, “sugarcane”, or “wood”. For this same reason, a table with the top 10 terms with the most important Pearson correlation index was produced (Table 4.6).

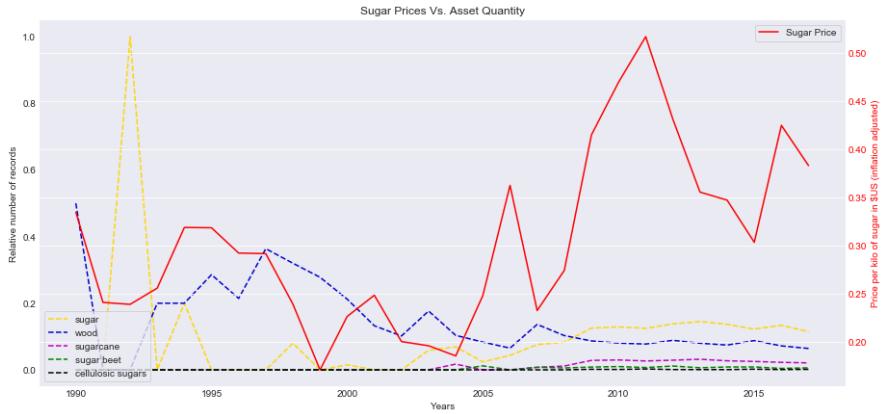


Figure 4.12: Normalized term usage and evolution of the price of the kilo of sugar from 1990 to 2017.

When observing the term ranking in Table 4.6, some interesting observations can be made. Firstly, almost all of the terms in this top ranking are in fact feedstock terms, raw materials used for biofuel production. Secondly, the two terms with the highest correlation with the price of sugar are highly related to it, particularly “sugarcane”, and “cellulosic sugars”. Finally, there is a presence of flowering plants such as “jatropha”, and “sorghum” that also have an important Pearson correlation index with the price of the kilo of sugar.

Feedstock Name	P-value	Pearson Correlation Index
sugarcane	6.074365e-07	0.789014
cellulosic sugars	1.263220e-06	0.775341
jatropha	1.521222e-06	0.771713
sorghum	3.083429e-06	0.757299
dry biomass	3.454736e-06	0.754884
beets	4.105286e-06	0.751165
dedicated energy crops	6.915371e-06	0.739525
algae	1.103076e-05	0.728562
hybrid poplar	1.490683e-05	0.721206
soy	2.631077e-05	0.706675

Table 4.6: Top 10 terms with the highest correlation with the price of the kile of sugar from 1990 to 2017.

4.2 Meso Level: Technological Capabilities of Nations

In the second part of the fourth section, the focus will lie in the study of the technological capabilities of countries. The methods used in this part of the analysis are closely related to the methodologies presented in the third section and the analyses made in the first part of this section, except for some small differences and additional analysis.

4.2.1 Characterisation of Countries

Capability Matrices

The first result produced relied on the representation of the biofuel research ecosystem of a country as a capability matrix. This capability matrix is the result of the same application of the term-pair methodology as previously presented at the macro level, but instead of filtering the documentation by year, the documentation was filtered by its location. For example, when creating the capability matrix for Denmark, the term-pair matrix is related to technological assets (patents, publications, projects) located in Denmark or owned (or even co-owned) by Danish organizations. To introduce this concept, taking Sweden and Denmark as examples, the normalized capability matrix for each one of these institutions was produced. Figure 4.13 and Table 4.7, a visualization of both matrices side by side, as well as some indicative properties are

shown.

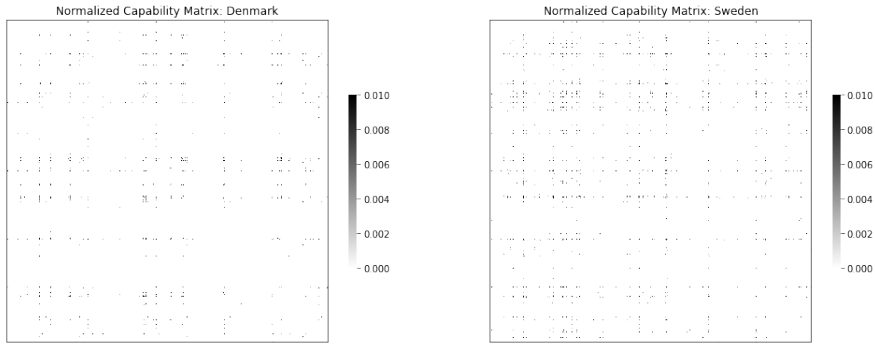


Figure 4.13: Normalized Capability Matrices of Denmark and Sweden.

When looking at the properties of both matrixes, some observations can be made:

- Both matrices are symmetrical and equal in dimensions which is expected given the same dictionary of biofuel related terms.

- The maximum, minimum and mean values of both matrices are generally similar.

- The standard deviation of the capability matrix of Denmark is 40% higher which would mean that Denmark has a wider usage of different term pairs. On the other hand, Sweden's capabilities are more "focused".

Property	Denmark	Sweden
Dimensions	352x352	352x352
Mean	1.71e-05	1.71e-05
Standard Deviation	3.95e-04	2.80e-04
Max	3.90e-02	3.06e-02
Min	0.00	0.00
Symmetry	Yes	Yes

Table 4.7: Denmark and Sweden capability matrix properties.

Capability Lists

Following the same approach as the macro analysis, a capability matrix can be transformed into a capability list by taking its upper triangle and adding each entry to a vector. In this level of the analysis, and as a proof of concept, the capability list of the United States of America and the capability list of China are presented side by side in Figure 4.14. However, due to the large number of entries in these lists (58482) the visualization of the differences becomes rather difficult. Consequently, in the final part of this subsection, this concept will be revisited in a more detailed manner.

Figure 4.14: Country capability lists of China and the USA.

4.2.2 Country Correlation Matrix and Profiles

Country Correlation Matrices

With the goal of applying the same engineering systems approach to the meso level, as was applied to the macro level, the Pearson correlation index was used as an indicator of the similarity between the capability lists of two countries. For example, the Pearson correlation index between the US and China lists has the value of approximately 0.65, or 65%. This could mean that the biofuel research between these two countries is 65% similar. To visualize this, the country correlation matrix was created in Figure 4.15.

After creating this matrix, and just like it was done for each year, a hierarchical clustering algorithm was applied to the matrix as a way of possibly identifying clusters of countries that are more similar between themselves. Moreover, this clustering technique also produced a dendogram as a way of quickly identifying the countries that are more related to another (Figure 4.16). For example, if this dendrogram was to be cut in the level n=2, forming clusters of two countries, Denmark would

be connected to Portugal, and the United States would be clustered with Taiwan. Interestingly, one can observe three main cluster areas in the ordered matrix:

- On the top left side of the matrix, an area of highly related countries that range from France to Serbia. (see axis of second figure)
- On the bottom right side of the matrix, an area of related countries, which on average are less related than the top left but separated. (Belgium, Hong Kong, Hungary Tunisia...)
- In the middle, a cross like area of countries which are not particularly related to each other or any other country. (El Salvador, UAE, Scotland...)

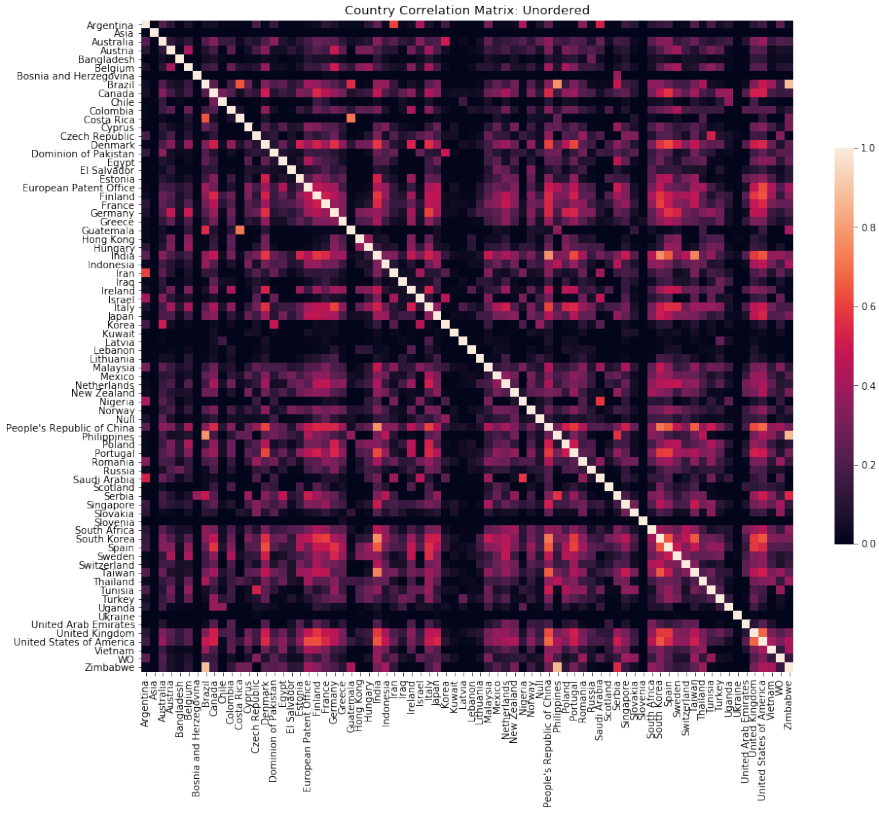


Figure 4.15: Country capability correlation matrix. Interactive version: https://plot.ly/ duarteocarmo/24.

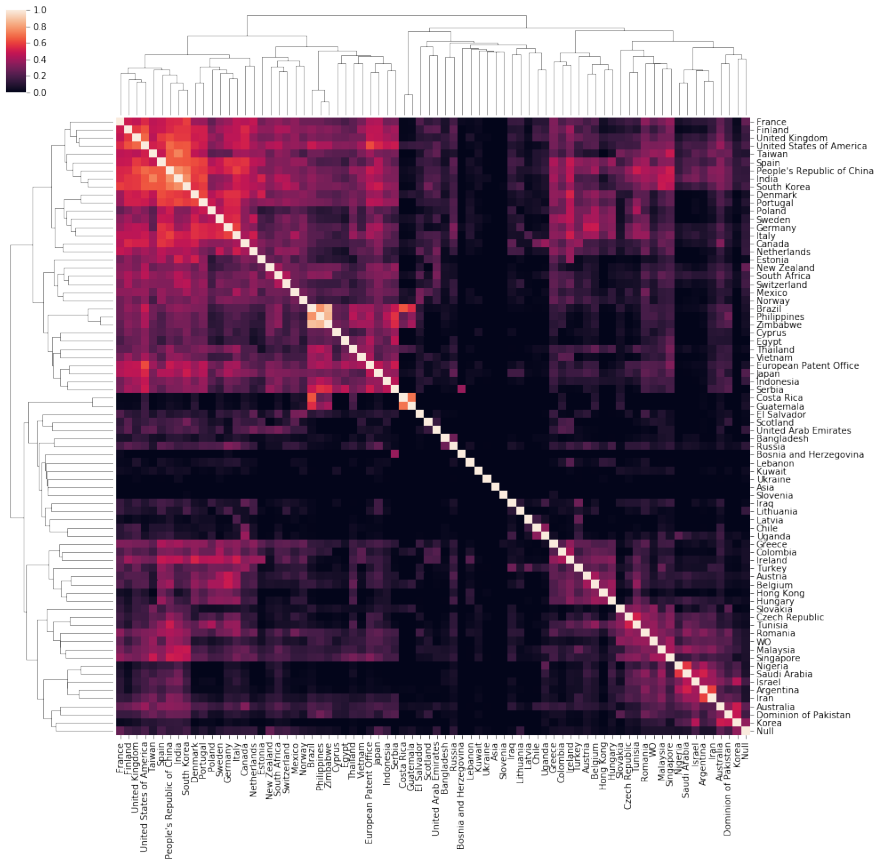


Figure 4.16: Clustered country capability correlation matrix.

Country Correlation Profiles

While clustering is an interesting way of visualizing the general trends that would possibly occur between countries, it does not explicitly show what countries are more related to each other. To visualize this, country profiles were created. A country profile is built by "cutting" the capability matrix for a particular row (country) and ordering the results. In Figure 4.17 below, the country profile of Denmark is presented. On the y axis, the Pearson correlation index (x100) is used as a measure of similarity between countries.

This graph is a simple way of quickly visualizing the most similar countries to

Denmark in terms of biofuel research. For example, the most related country to Denmark is Spain, with an index of about 60%, following Portugal with an index of 58%, etc. Interestingly, the most similar countries are not necessarily close in geography to Denmark, but close to themselves. Sweden for example is related to Denmark by a factor of 50%, and Norway by only 30%. Following this method one could say that in terms of biofuel-related capability matrices Norway is as similar to Denmark as Colombia is.

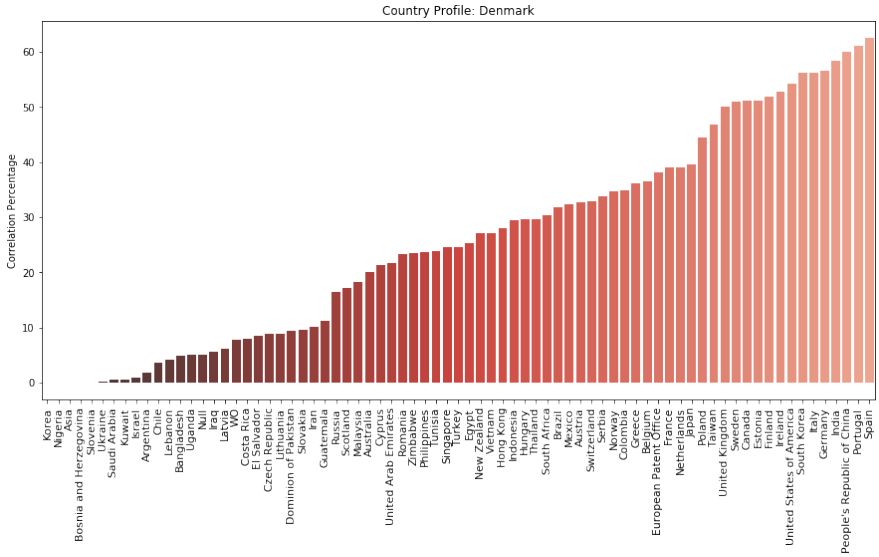


Figure 4.17: Country profile of Denmark.

4.2.3 Contextual Relationships

4.2.3.1 GDP per capita

Using the world bank as a source of data to get the values in $US of the Growth Domestic Product per capita, the GDP per capita difference for every country pair was calculated. After calculating this, the goal is to understand if the GDP per capita of two countries is telling of the technological similarity of those two countries. In Figure 4.18, presented below, each data point is a pair of countries. In the x axis, the Pearson correlation (0-1) between the country pairs, and in the y axis, the absolute

GDP per capita difference of those same country pairs. For readability purposes, if two countries have a capability similarity of less than 10%, or 0.1, this pair would be excluded from the graph.

When observing the graph, one can notice that most country pairs have less than a 40% capability correlation and less than 40000$US GDP per capita difference. On the other hand, when looking at the dashed guidelines in the graph, the further from the origin of the graph a guideline is, the less country pairs appear. Moreover, generally, countries that are more related (higher capability correlation), have a more similar GDP per capita. For example, Brazil and Zimbabwe, have a capability correlation of 88.60% (0.88), and a GDP per capita difference of 7620.87$US, which is rather low.

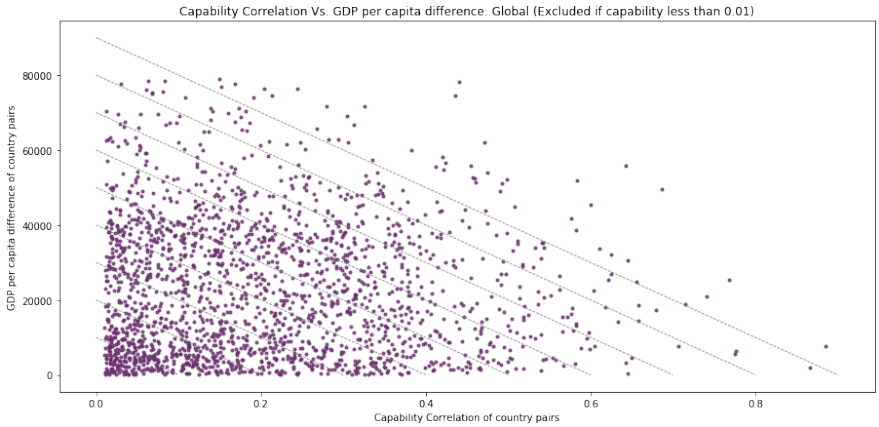


Figure 4.18: Capability correlation and GDP per capita difference of country pairs.

However, the graph above loses an important dimension: it is hard to distinguish country pairs just from the GDP per capita difference. For example, let us take as an example the country pairs Sweden-Singapore, and Romania-Brazil. These two country pairs have a low GDP per capita difference; however, the first pair is made of economically developed countries, and the second, generally underdeveloped countries. The graph above treats them equally.

In order to add an extra dimension to this visualization, Figure 4.19 produced. Here, one can also see the average GDP per capita of each country pair as a color scale. For instants, Sweden-Singapore is light blue, and Romania-Brazil is red.

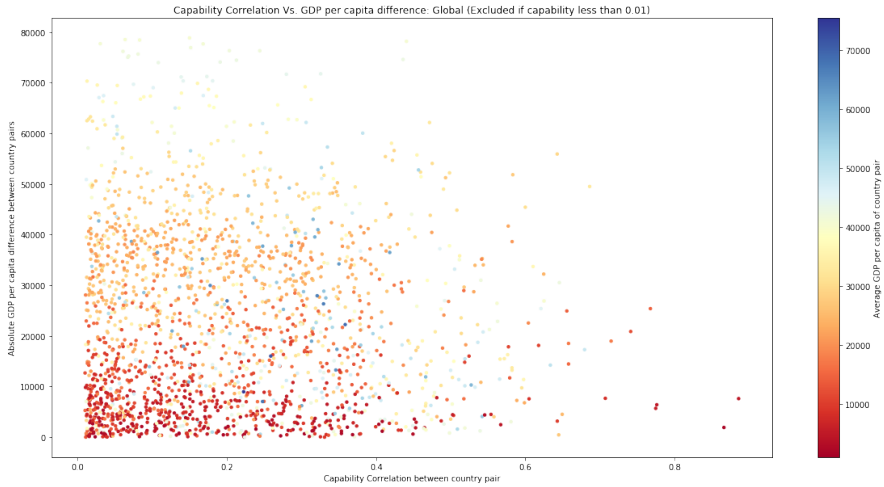


Figure 4.19: Capability correlation and GDP per capita difference of country pairs. Interactive version: https://plot.ly/ duarteocarmo/32/.

4.2.3.2 Collaboration

The second contextualization is not necessarily from an external data source, instead, it was obtained from the database itself. By querying the database, it was possible to retrieve, for each country pair, the number of technological assets where these countries collaborated.

By taking the number of shared assets between a country pair and the capability correlation between that same country pair, Figure 4.20 was produced. In it, 4 different areas can be observed (in italic, example pairs):

On one hand, most country-pairs are located in the "Different and not collaborating" quadrant. On the other hand, there is a high number of country pairs that are similar in terms of capability but are not collaborating.

-	Low number of shared assets	High number of shared assets
High capability correlation	**Similar and not collaborating** *Argentina - Iran* *Belgium - China* *Brazil - Costa Rica* *Canada - Denmark*	**Similar and collaborating** *Austria - Germany* *Belgium - Germany* *Denmark - Germany* *Finland - Germany*
Low capability correlation	**Different and not collaborating** *Argentina - Australia* *France - Servia* *Indonesia - Malaysia* *Brazil - Portugal*	**Different and collaborating** *Austria - France* *Belgium - France* *Denmark - Netherlands* *El Salvador - Germany*

Table 4.8: Examples of collaborations and categories.

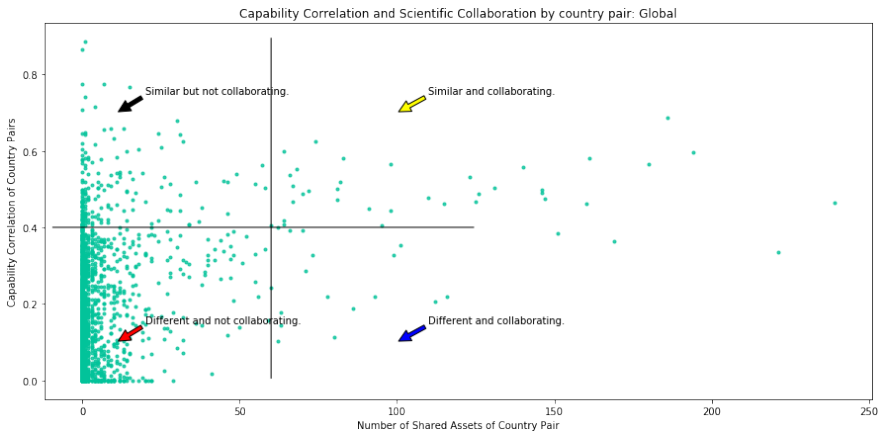


Figure 4.20: Capability correlation and collaboration between country pairs.

When looking at Figure 4.20, one can consider the number of shared assets indicator as an unfair index. This because not all countries possess the same number of assets. For example, the US has an extremely high number of documents, while other countries such as Costa Rica or Lebanon have a very low number of documents. For this reason, a new index, the normalized number of shared assets was created, as a way of valuing collaborations as a percentage of total documents produced by the country pair, its definition follows:

- Old collaboration definition: Country i and country j have z assets that have both their name as location.

- New normalized collaboration definition: normalized collaboration = (number of shared assets between country i and j)/(number of total possible collaborations between i and j)

For example, for the country pair Portugal-Denmark:

- Number of assets Denmark: 351

- Number of assets Portugal: 180

- Number of shared assets: 25.0

- Number of normalized shared assets: 0.13 (=25/180)

In Figure 4.21, the same graph, but with the normalized shared assets between each country pair is presented. One can notice that there is less saturation generally, and country pairs are more distributed. Moreover, some outliers appear: such as France-Lebanon.

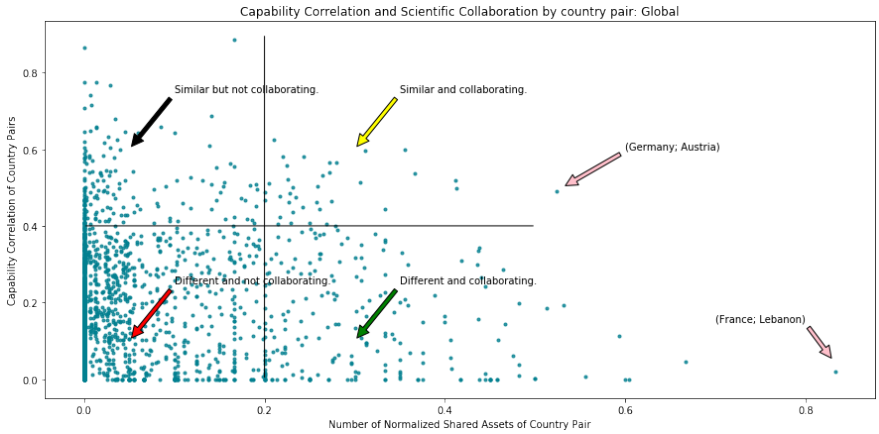


Figure 4.21: Capability correlation and normalized collaboration between country pairs. Interactive version: https://plot.ly/ duarteocarmo/34/.

4.2.4 Comparing Countries

Coming back to the more general analysis, in the same way as two years were compared in terms of capability, two countries will now be compared in terms of term pairs usage. It is worth noting that this approach is simply a deep dive into the capability matrices of two different countries and looking at the most common term pairs in each of them.

As an example, the countries Brazil and Denmark will be compared, their capability correlation is around 30%. The first result is the top term pairs for each of these countries presented in Tables 4.9 and 4.10. One can note that in the top term pairs of Brazil, there is a high number of term pairs related to sugar, sugarcane and cellulose. One the other hand, in the Denmark table, there is more stress on processing technologies (digestion, fermentation, hydrolysis), and outputs.

Similarly to what was done with the macro level analysis, the table of the most important term-pair usage differences was produced, and can be seen in Table 4.11. One can note a high number of term pairs that are not used at all by Denmark, and used in Brazil: “sugar-sugarcane”, “advanced biofuel-cellulosic ethanol”, “sugarcane-ethanol”. On the other hand, there is lower number of terms that are only used in Denmark (“straw-hydrolysis”). Moreover, feedstocks and related term pairs are common in this table, with terms such as sugar, sugarcane, or straw, being divisive between countries.

First Term	Second Term	Documents	Percentage
anaerobic digestion	biogas	31	0.088319
ethanol	fermentation	26	0.074074
ethanol	hydrolysis	23	0.065527
ethanol	straw	14	0.039886
bioethanol	fermentation	14	0.039886
yeast	fermentation	12	0.034188
ethanol	enzymatic hydrolysis	12	0.034188
ethanol	cellulosic ethanol	12	0.034188
hydrolysis	bioethanol	12	0.034188
fermentation	cellulosic ethanol	11	0.031339

Table 4.9: Top term pairs Denmark.

First Term	Second Term	Documents	Percentage
sugarcane	sugar	416	0.525917
ethanol	fermentation	401	0.506953
fermentation	sugar	210	0.265487
ethanol	cellulosic ethanol	208	0.262958
ethanol	sugar	207	0.261694
ethanol	advanced biofuel	200	0.252845
advanced biofuel	cellulosic ethanol	200	0.252845
fermentation	sugarcane	198	0.250316
ethanol	sugarcane	195	0.246523
ethanol	hydrolysis	42	0.053097

Table 4.10: Top term pairs Brazil.

First Term	Second Term	Denmark	Brazil	Difference
sugarcane	sugar	0.000000	0.525917	0.525917
ethanol	fermentation	0.074074	0.506953	0.432879
advanced biofuel	cellulosic ethanol	0.000000	0.252845	0.252845
ethanol	advanced biofuel	0.000000	0.252845	0.252845
fermentation	sugar	0.014245	0.265487	0.251242
fermentation	sugarcane	0.000000	0.250316	0.250316
ethanol	sugar	0.014245	0.261694	0.247449
ethanol	sugarcane	0.000000	0.246523	0.246523
ethanol	cellulosic ethanol	0.034188	0.262958	0.228770
anaerobic digestion	biogas	0.088319	0.015171	0.073148
ethanol	straw	0.039886	0.002528	0.037358
straw	hydrolysis	0.031339	0.000000	0.031339
fermentation	mixed biomass	0.025641	0.000000	0.025641
hydrolysis	bioethanol	0.034188	0.010114	0.024074
straw	wheat	0.022792	0.000000	0.022792
biogas	waste	0.025641	0.003793	0.021848
biodiesel	transesterification	0.000000	0.020228	0.020228
straw	fermentation	0.019943	0.000000	0.019943
vegetable oil	transesterification	0.000000	0.018963	0.018963
fermentation	cellulosic ethanol	0.031339	0.012642	0.018697

Table 4.11: Top term pairs usage differences in Denmark and Brazil. Percentages are 0-1.

4.2.5 Country Spectrums

Representing Country Spectrums

As a way of diving deeper into the country capability spectrums, understanding their composition, and making the analogy between term pairs and amino-acid pairs in DNA representations, in the following section the country spectrum concept was further developed.

Instead of focusing in the frequency of the appearance of a certain term pair, let us focus on whether a term-pair appears or not in the capability list of a country or not.

In Figure 4.22, for 7 countries, and the first 45 term pairs of the capability spectrum are represented. Even though this is a very small part of the spectrum (<1%), one can already see some term pairs that appear in several countries. "Natural Gas / Anaerobic Digestion" for instance, appears in Finland and Denmark. Moreover, there are a wide range of terms that only appear in one country. Such as terms related to "animal fats", in the case of Spain.

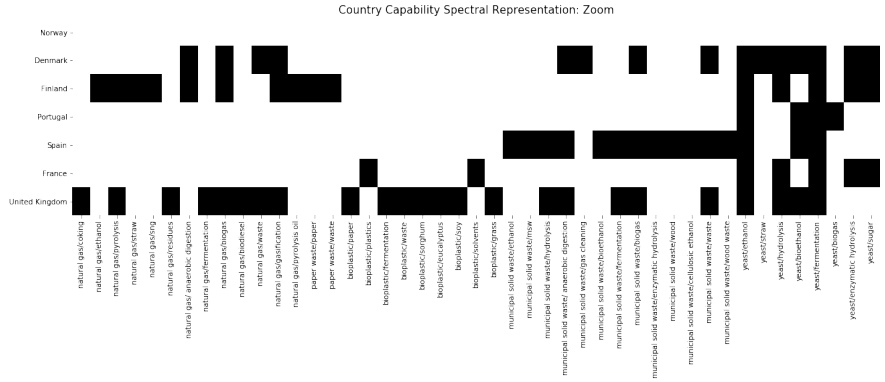


Figure 4.22: Country capability spectrum (first 45 term pairs) for a set of 7 countries.

Generalizing this capability spectrum concept to all of the countries all of the terms pairs, is a good way of visualizing the biofuels capability "DNA" of all of them. However, in order to improve the quality of this visualization, two adjustments were made:

- The order of the countries in the left hand side was adjusted to reflect the result of the clustering in the country correlation matrix.
- Only term pairs that were used by at least 2 countries were represented. This allowed the reduction of the original size of the capability spectrum from 58482 values, to 6236 values.

This representation can be consulted in Figure A.6.

The uniqueness of countries

Taking the capability spectrum of a country as a starting point, the next and final step of the analysis seeks to understand how unique each country is in terms of usage of terms pairs. Denmark, for example, in its capability spectrum, uses a total of 256 different term pairs. Of these 256 term pairs, there are a total of 21 that are only used by documents located in Denmark, these can be consulted in Table 4.12.

Taking this approach and applying it to all of the countries in the database, a uniqueness index was developed. The uniqueness index of a country is the ratio

Term pairs unique to Denmark	various grasses/straw
	various grasses/waste
	various grasses/garden waste
	various grasses/ethanol
	various grasses/enzymatic hydrolysis
	various grasses/hydrolysis
	industrial waste/gas cleaning
	sewage/gas cleaning
	mixed biomass/biogas
	mixed biomass/cellulosic ethanol
	mixed biomass/fermentation
	straw/garden waste
	grass/garden waste
	rapeseed oil/solvents
	garden waste/ethanol
	garden waste/enzymatic hydrolysis
	garden waste/hydrolysis
	rapeseed/solvents
	biogas/cellulosic ethanol

Table 4.12: Unique term pairs for Denmark.

between the total number of term pairs used by a country and the number of term pairs that are unique to that country. In the case of Denmark, this value would be equal to $21/256 = 0.082$.

With this approach, a table of the top 20 most unique countries was created. Table 4.13, presented below, one can see that the most unique country is the US, with an index of almost 0.50. This means that half the term pairs used by the US are only used by the US (!). The rest of the countries in the ranking have a relatively low number of term pairs, Lebanon, for example, with only 5 term pairs, of which 1 is unique. The top 20 most unique countries have either a very large number of term pairs or a very low number of term pairs.

4.3 Micro Level: Technological Capabilities and Organizations

On the third and final part of the main analysis, the results of the application of the methodology to the micro level will be presented. Here, the focus will rely on studying organizations as the units of technological capability. Moreover, the analysis will focus on the study of the Technical University of Denmark and of organizations based in Denmark.

4.3.1 Characterisation of organizations

The first part of the micro level analysis seeks to characterize an organization by its capability matrix, as previously done for years at the macro level, and countries at the meso level. By filtering all of the documents whose owner contained the "Technical University of Denmark", the technological capability matrix of the University was achieved. The normalized version of this matrix, has the properties described in Table 4.14. In its capability matrix, DTU has the maximum value in the position (200, 286) with the value of 103.8. This position refers to the terms biogas and anaerobic digestion

4.3.2 Organization correlation matrix and organization profiles

Following the methodology applied in sections 4.1.2 and 4.2.2, organizations become the focus. A comparison was made between the capability matrices of 112 organizations, by the means of the calculation of the Pearson correlation index between them. The database contains a total of 10638 organizations; therefore, in order to maintain the quality of visualizations, a triage of the organizations was made. By only retrieving organizations with 7 assets or more and for reference purposes added a number of Danish organizations. As a result the list was reduced to a total of 112 organizations. The Danish organizations added manually were: DTU, Novozymes, Aalborg University, Aarhus University, University of Copenhagen, University of Southern Denmark, DTU Riso, and Dong Energy.

By comparing every organization's capability matrix and calculating the correlation between each of them, the capability matrix of organizations was created (Figure

Country	Uniqueness	Unique Pairs	Total Pairs
United States of America	0.477474	2215	4639
Ukraine	0.263158	5	19
Indonesia	0.210526	12	57
Lebanon	0.200000	1	5
Bangladesh	0.200000	7	35
Cyprus	0.157895	3	19
Sweden	0.156250	70	448
Austria	0.150685	11	73
European Patent Office	0.140145	155	1106
El Salvador	0.132530	11	83
Australia	0.132296	34	257
Uganda	0.130435	3	23
South Africa	0.125749	21	167
Hong Kong	0.125000	3	24
Turkey	0.120773	25	207
People's Republic of China	0.113377	139	1226
Thailand	0.111111	9	81
Brazil	0.110672	56	506
United Kingdom	0.106870	70	655
Canada	0.106061	84	792

Table 4.13: Uniqueness ranking.

Property	**Value**
Shape	352x352
Max	0.82
Min	0.00
Mean	3.7e-04
Standard Deviation	7.8e-03
Symmetry	Yes

Table 4.14: Normalized capability matrix for DTU.

4.23). In the figure below is the clustered version of this matrix. In this clustering, Danish organizations do not appear as particularly related. DTU for example, appears in a cluster with a majority of Chinese organizations (Tsinghua University, Shanghai University and Beijing University). On another note, Dong, Riso, Aalborg and the University of Copenhagen are highly connected in this clustering. Novozymes appears rather isolated, while the University of Southern Denmark and Aarhus appear in the same cluster. The organizations in focus are highlighted in red in the graph below.

When cutting this correlation matrix in the rows corresponding to DTU (4.24) and Novozymes (4.25), we are able to access the profiles of both of these organizations. In the case of DTU, its top 3 correlations are with Shanghai Jiao Tong University (70%), Tsinghua University (70%) and the Portuguese University of Minho (69%). The organization that is not a university with the highest correlation with DTU is CSIR (African Research Organization). In the case of Novozymes, the correlations are on average lower. The Indian Institute of Chemical Technology (57%) and the University of Copenhagen (55%) make for the most related organizations.

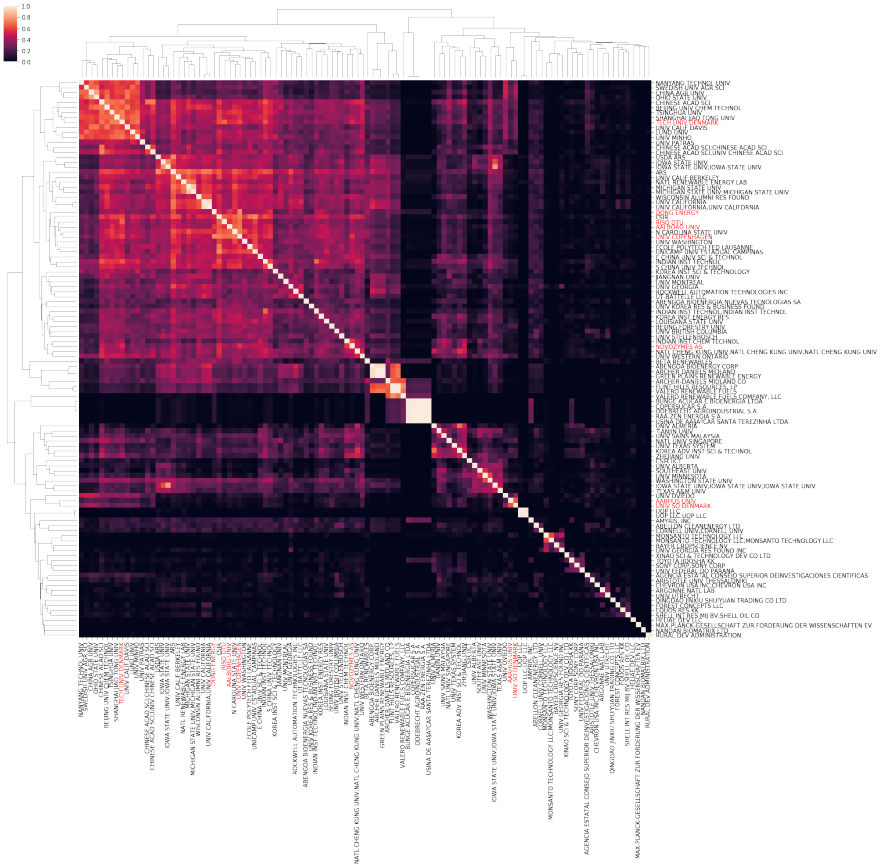


Figure 4.23: Organization correlation matrix: Clustered Version (Danish organizations in red).

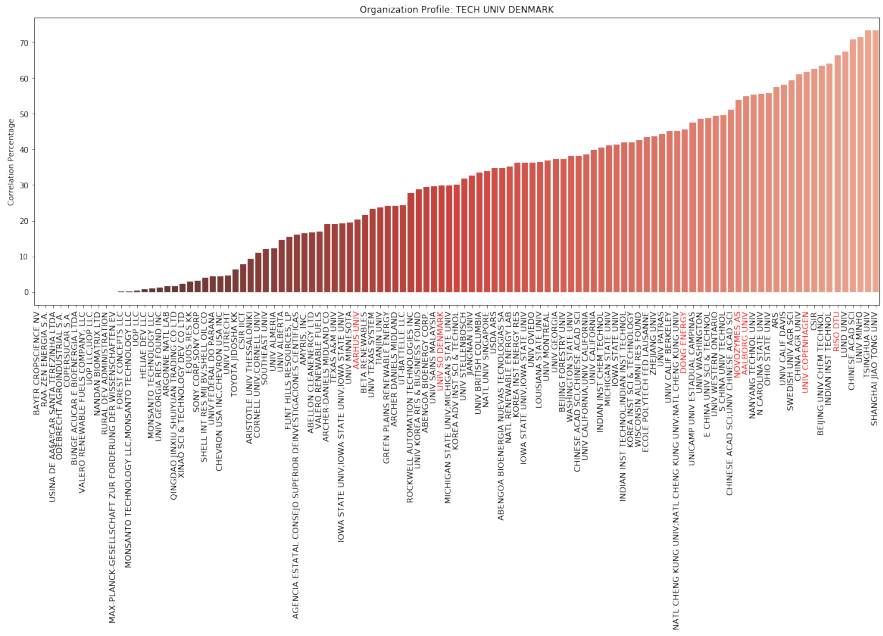


Figure 4.24: Country correlation profile: Technical University of Denmark.

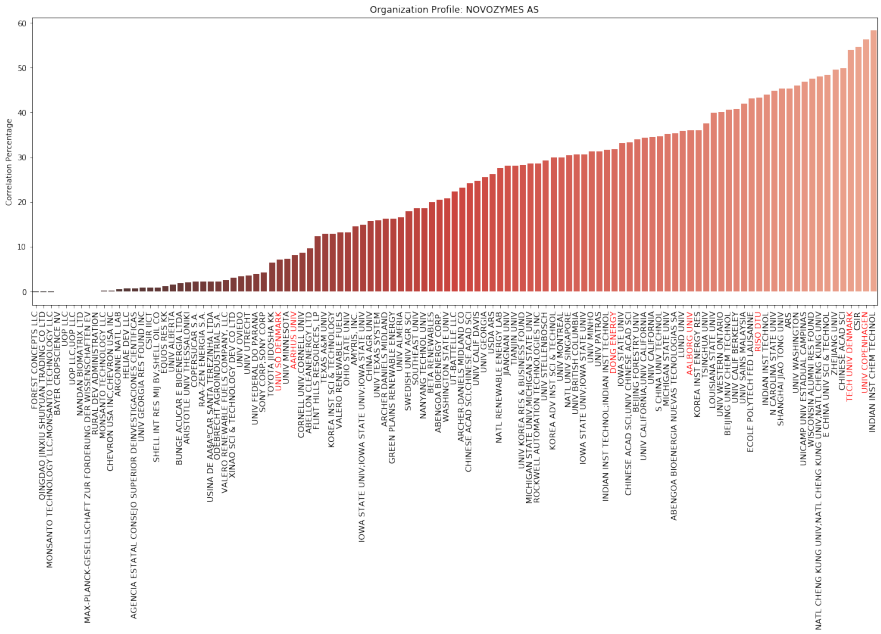


Figure 4.25: Country correlation profile: Novozymes AS.

4.3.3 Comparing organizations

In order to prove the concept of comparing the capabilities of two organizations, the Technical University of Denmark and the University of Tsinghua were selected. These organizations are similar by a factor of about 70%, and for this reason the differences in the usage of term pairs across documentation is not necessarily important.

Comparing the usage of all term pairs and then ordering the absolute relative difference, resulted in Table 4.15. Here, the top 20 term pairs that differed mostly in usage between these organizations are presented. The first observation to be made is the fact that several of these term pairs are simply not used by one organization. “Biodiesel-hydrolysis” and “sorghum-fermentation” for example, are simply not used by DTU; however, they both appear in 1.15% of all documents owned by Tsinghua University. Another interesting note is the difference in terms of feedstocks used. Tsinghua University for example, uses “sewage” and “waste” and “gasoline” in a more important way, while on the other side, “algae-anaerobic digestion” is not used

at all.

4.3.4 Collaborations

A particularity of the micro level analysis is the study of the types of collaboration made between organizations. By applying the same methodology as section 4.2.3.1 for each organization in the list, details about the nature of the collaborations that were put in place was obtained. By filtering these partners, and detecting the word "Univ" in their names, the program designed was able to detect the nature of the collaboration, and categorize them into "University Partnerships" or "Organizational Partnerships".

In the two bar charts below, the organizations in the database are ordered according to the percentage of partnerships of Universities.

Figure 4.26, presents all of the organizations in the original list. The Agricultural Research Service (ARS) appears in the first position with 70% of partnerships made with universities and 30% with other organizations. Approximately a third of the organizations queried, have no university partnerships, and only collaborate with businesses or organizations. The Danish organizations and universities (in red), all collaborate with universities, the one that does so the least is "Riso DTU", where university partnerships make up for approximately 48% of all collaborations. When looking at businesses, Novozymes' collaborations are 60% with universities, and Dong Energy 57%.

The second chart, Figure 4.27 is derived from the first, but on the x axis, only universities are presented. DTU, for example, is balanced in terms of university and organizational collaborations with ratios of 58% and 42% respectively. On the other side of the spectrum, two Brazilian universities (Universidade Federal do Parana and Universidade Estadual de Campinas) and one Canadian university (University of Alberta), only collaborate with businesses and research organizations.

There are no organizations that collaborate exclusively with universities, but organizations that collaborate exclusively with non-universities do exist. Moreover, the Danish organizations in focus (red) all collaborate with universities in at least 50% of their shared assets. Finally over half of the non-universities queried, collaborate almost exclusively with other non-universities.

First Term	Second Term	DTU	TSINGHUA	Difference %
bioethanol	fermentation	0.656250	0.188679	0.467571
anaerobic digestion	biogas	0.828125	0.452830	0.375295
ethanol	fermentation	0.734375	0.490566	0.243809
anaerobic digestion	algae	0.218750	0.000000	0.218750
anaerobic digestion	biodiesel	0.203125	0.000000	0.203125
hydrolysis	biogas	0.203125	0.000000	0.203125
biodiesel	transesterification	0.171875	0.000000	0.171875
ethanol	catalysis	0.171875	0.000000	0.171875
biogas	manure	0.203125	0.037736	0.165389
straw	fermentation	0.203125	0.037736	0.165389
ethanol	bioethanol	0.218750	0.075472	0.143278
methanol	catalysis	0.140625	0.000000	0.140625
methanol	transesterification	0.140625	0.000000	0.140625
ethanol	transesterification	0.140625	0.000000	0.140625
hydrolysis	bioethanol	0.250000	0.113208	0.136792
fermentation	sorghum	0.000000	0.132075	0.132075
hydrolysis	biodiesel	0.000000	0.132075	0.132075
hydrolysis	cellulose	0.187500	0.056604	0.130896
ethanol	straw	0.203125	0.075472	0.127653
algae	biogas	0.125000	0.000000	0.125000

Table 4.15: Term pair comparison of DTU and Tsinghua University.

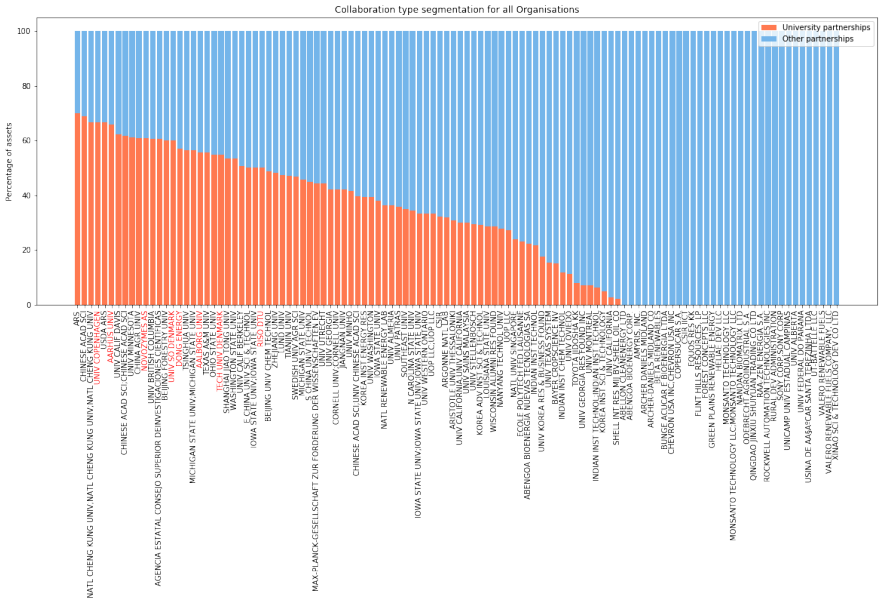


Figure 4.26: Collaboration by type of partnership: All organizations.

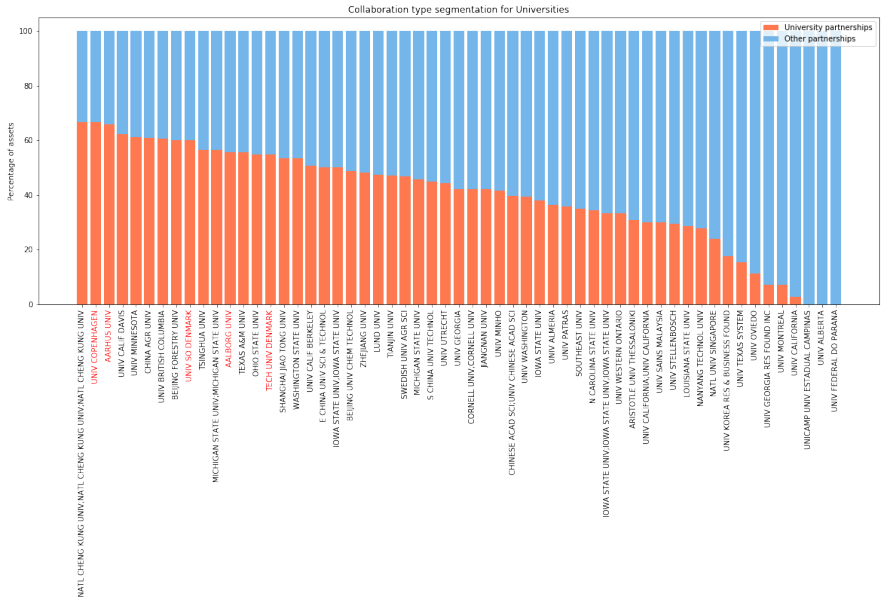


Figure 4.27: Collaboration by type of partnership: Universities.

4.3.5 Organizational spectrums

Similarly to the analysis and results carried out in section 4.2.5, the same approach was applied to organizations.

When observing the first representation of the organization spectrums (Figure 4.28) for DTU, Novozymes, the University of California Berkeley and Tsinghua University, one can notice immediately some differences and similarities between the organizations and the source of their differences. Some term pairs such as "sugar/-cellulose" appear in all of the 4 organizations. While others are exclusive to some organizations such as "willow/miscanthus" for DTU or "paper/grass" for the University of Berkeley. This representative visualization serves as a proof of concept to the following one, where all of the countries and term-pairs are presented.

In the full picture visualization (A.7), the countries were ordered according to the result of the hierarchical clustering of section 4.2.2. The first observation to be made is related to the long vertical lines along all of the organizational spectrums, these long bars are the representation of term pairs that are widely used by a large num-

ber of organizations, thus producing this effect. Moreover, some of the organizational spectrums seem more widely distributed than others. For instance, the organizational spectrum of Shell (5th to bottom) has a number of condensed black areas corresponding to many interrelated term pairs; on the other hand, the organizational spectrum of DTU appears to be more diverse and well distributed. Finally, one can note that the quality of clustering, this is because organizations and groups of organizations that use the same term pairs appear closer together.

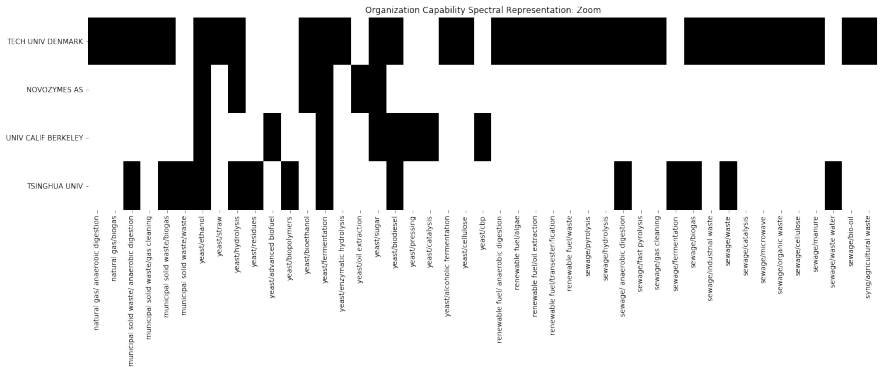


Figure 4.28: Organizational spectrums: Limited to 45 term pairs and 4 organizations.

4.4 Patents and Publications

4.4.1 Patent and Publication Matrix: Characterisation and differences

In this first part of the complementary analysis, the focus will not be in the macro, meso or micro level, but rather in comparison of two different types of assets present in the database, particularly patents and scientific publications. As a first result of the analysis, is the creation of two capability matrixes, one for patents and another for scientific publications. These matrices, represented as heat maps in Figure 4.29, are obtained through the same term-pair approach in conjunction with a filtering of the technological assets by type.

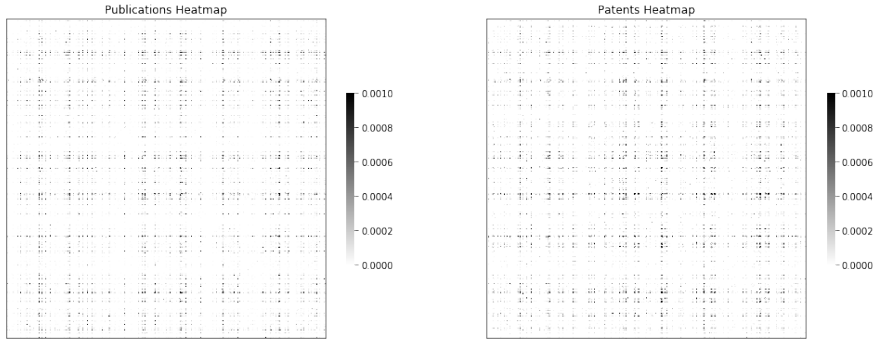


Figure 4.29: Heat maps of the capability matrices for patents and publications.

Following the same approach as before, and with the goal of understanding the differences in terms of utilization of different term pairs between patents and publications, these two normalized matrices were compared. As a result, a table of the most divergent term pairs in terms of usage was created (Table 4.16). Here, one can see the term pairs that are more published than patented and vice-versa, ordered by the absolute difference in percentage of utilization. For example, in the first position, the term pair "biogas-ethanol/anaerobic digestion" is used in 36% of publications, but only 0.44% in patents. Moreover, it should be noted that most of the term pairs in this table are either processing technologies (e.g. anaerobic digestion, fermentation, pyrolysis, hydrolysis) or output terms (e.g. biogas, ethanol), and that the presence

of feedstock term pairs is rather unimportant.

4.4.2 Evolution of asset types over time

In order to study the evolution of the systems of the two types of assets over time, the total number of records of each type from 1990 until 2017 was plotted in Figure 4.30. An explosion in the number of assets can be noted from the year of 2007, moreover, there seems to be a sharp downfall of the total number of assets from 2015. When comparing patents and publications, while the years of 2004-2010 have a far superior number of patent assets over published assets, this trend seems to be overcome after that same period, following a sharp rise in published assets.

First Term	Second Term	Patents	Publications	Difference
anaerobic digestion	biogas	0.000488	0.016448	0.015960
pyrolysis	bio-oil	0.000561	0.008202	0.007641
bioethanol	fermentation	0.002560	0.009282	0.006722
hydrolysis	bioethanol	0.001609	0.006748	0.005139
biodiesel	catalysis	0.000488	0.005337	0.004850
biogas	waste	0.001439	0.006182	0.004743
butanol	fermentation	0.006925	0.002377	0.004548
ethanol	catalysis	0.000317	0.004606	0.004289
ethanol	butanol	0.005364	0.001115	0.004250
ethanol	pressing	0.004828	0.000601	0.004227
biodiesel	transesterification	0.002975	0.007192	0.004217
anaerobic digestion	waste	0.000415	0.004441	0.004026
ethanol	enzymatic hydrolysis	0.002585	0.006574	0.003989
hydrolysis	biogas	0.001170	0.004937	0.003767
methanol	catalysis	0.000073	0.003474	0.003401

Table 4.16: Top term pairs usage differences in Patents and Publications. Percentages are 0-1.

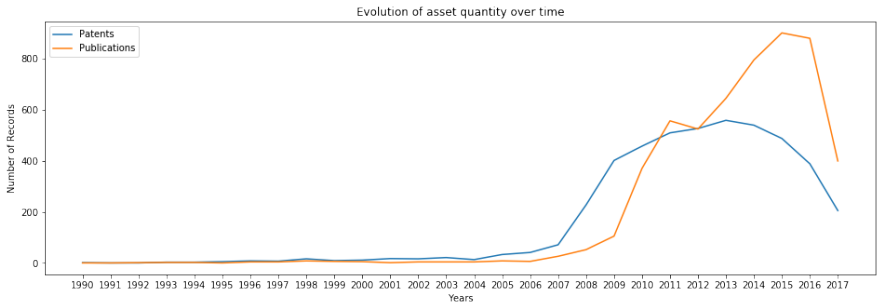


Figure 4.30: Evolution of biofuel patents and publications from 1990 until 2017.

Furthermore, another focus of this subsection is to understand how different terms are patented vs. published. As a proof of concept, 8 feedstock terms were chosen (waste, algae, cellulose, sugar, paper, wood, residues and corn) and their evolution over the period of time of 1990-2017 was studied. In Figure A.8, both the normalized and absolute versions are shown. Where one shows the total number of assets of each type with that term, the other shows the number of assets as a ratio of the total number of assets in that year. It can be noted that the behavior of the different terms follows the general pattern: they appear more in patents until a certain moment in time, and after that period, they appear more in scientific publications.

4.4.3 Term distribution by type of asset

The final part of this complementary analysis, seeks to understand the behavior of the different types of terms (feedstocks, processing technologies, and outputs) in terms of their balance in presence in patents or publications.

To understand this, the same approach was applied to 3 groups. In this approach all of the different terms are plotted in a graph where the x axis corresponds to percentage of patents, and the y axis corresponds to the publications. A perfect balance would align the terms perfectly following the x=y curve (a term is as patented as published). The further away from the x=y diagonal the data point is, the bigger the bias (towards publications or patents) towards one type of technological asset.

Though the three different groups were analyzed in this subsection we will focus on feedstocks as a proof of concept. As can be noted in Figure 4.31, there seems to be a relative balance between publications and patents in the case of feedstocks.

Moreover, in Table 4.17, the assets with the biggest distance to x=y were printed. The term “starch” for example, has a bias towards patents (3.04%) when comparing to publications (0.88%). However, when counting the global number of feedstocks, in 170 terms, 80 appear more in patents and 90 appear more in publications.

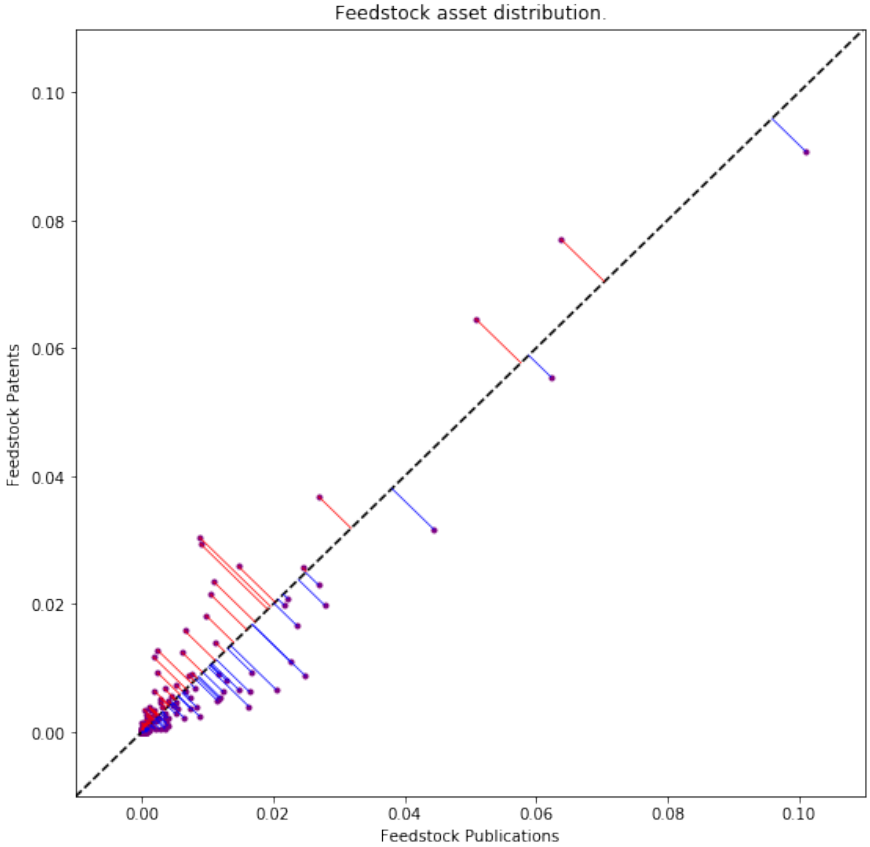


Figure 4.31: Distribution of Feedstock terms across asset type.

After analyzing the three different types of terms, another graph was produced, where all of the different biofuel related terms appear (Figure 4.32). It can be observed that, in a general manner, most terms have a rather low patenting and publication rate. Furthermore, a big discrepancy does not seem to exist between patents and publications when looking at all of the terms

Name	Patent	Publications	Distance to Mean	Bias
starch	0.030417	0.008816	0.015274	Patents
grain	0.029412	0.008946	0.014472	Patents
agriculture	0.008798	0.024893	0.011381	Publications
sugar	0.064605	0.050823	0.009745	Patents
waste water	0.006662	0.020420	0.009729	Publications
algae	0.077049	0.063788	0.009376	Patents
paper	0.031674	0.044470	0.009048	Publications
blend	0.023504	0.010956	0.008873	Patents
energy crops	0.004022	0.016271	0.008661	Publications
sewage	0.010935	0.022689	0.008311	Publications

Table 4.17: Most unbalanced feedstock terms.

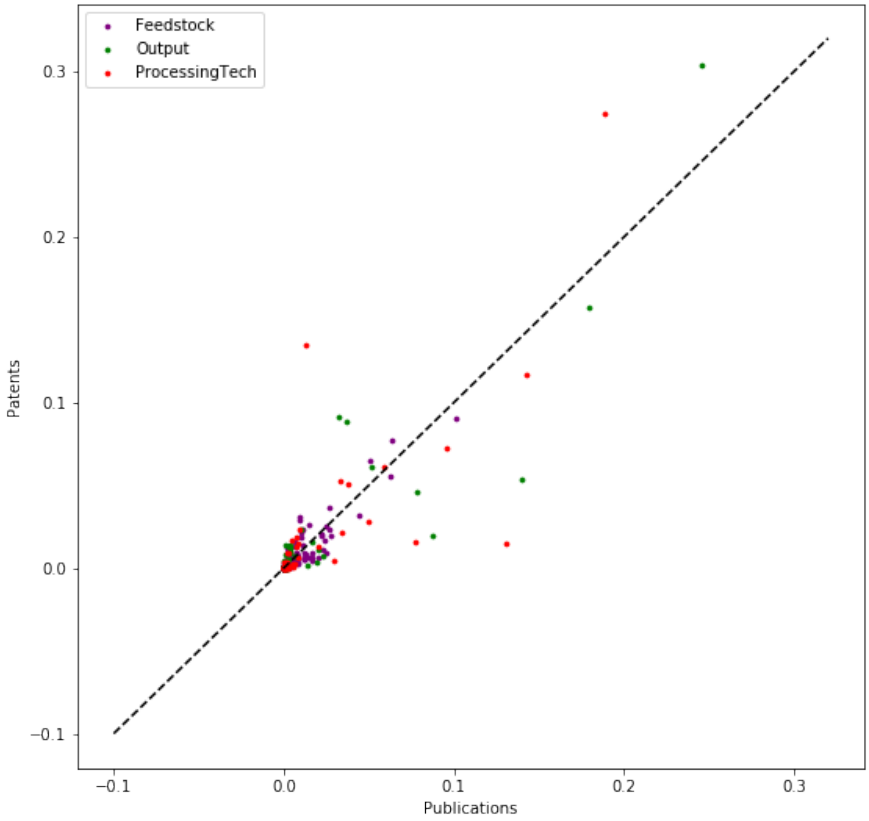


Figure 4.32: Distribution of all biofuel related terms by type.

4.5 Technological Capabilities: An Alternative Approach

In the final part of the complementary analysis, two unsupervised learning methods will be applied to the dataset at the meso level. In this approach, instead of focusing on term-pairs, the methodology will be based on simple term frequency. By simplifying the approach, one can possibly compare this approach with the one used throughout the analysis.

4.5.1 T-SNE algorithm implementation

The first step of the analysis is the creation of a matrix, X where each row represents a country, and each column a term related to biofuels. Each entry of that matrix corresponds to the number of times that particular term was used in documentation located in that. Finally, each row of that matrix was normalized in relation to the documents produced by that country. For example, the country of the US was divided by the total number of documents located in it. As a result, X has the shape (137, 352) which corresponds to 137 countries and 352 terms.

After creating this matrix, the t-sne algorithm was applied to it. The t-Distributed Stochastic Neighbor Embedding algorithm serves as a way of visualizing high dimensional data, in a two or three dimensional way((Maaten 2018)). For example, matrix X has a total of 352 dimensions, this number of elements is simply not possible to visualize. The algorithm applies a mathematical transformation that turns this matrix into a (137, 2) or (137,3) matrix, turning the visualization possible.

This algorithm can be fed with two arguments, the number of iterations and the perplexity. In general, the author of the algorithm advises the use of a perplexity parameter between 30-100. In Figure 4.33, the application of this algorithm with different values of perplexity can be observed. In them, each data point corresponds to one country, and the color-scale provides context as to the “richness” of that country (using the GDP per capita). The application does not result in any type of noticeable clustering, however, one can observe that from a perplexity of about 30, countries with a higher GDP per capita appear to be centered in the plot, while countries with lower GDP per capita, appear scattered around them.

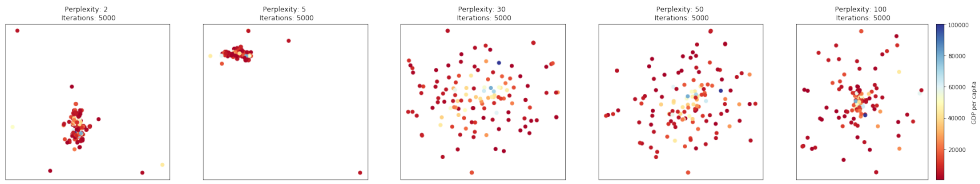


Figure 4.33: Application of the t-sne algorithm at the meso level.

4.5.2 Hierarchical clustering based on term frequency

The final result to be presented in this section is the result of application of hierarchical clustering in two different ways: Using term frequency first, and using term-pair frequency after. For both applications hierarchical clustering using the average distance was applied.

In the first figure (Figure 4.34), the dendogram that resulted from the application of hierarchical clustering to the term frequency matrix X is presented. When cutting the dendrogram where cluster have a size of 20 or less countries, the group that contains Denmark also contains the following countries: Austria, Belgium, Czech Republic, El Salvador, Finland, France, Germany, Greece, Hungary, Italy, Netherlands, Norway, Poland, Russia, Spain, Sweden, Switzerland, and the United Kingdom.

In the second figure (Figure 4.35), the same algorithm was applied to a matrix containing term-pairs - as in the main analysis presented in this section. When cutting the dendogram at the same level, Denmark's cluster also contains: Canada, Finland, France, Germany, India, Italy, Netherlands, China, Poland, Portugal, South Korea, Spain, Sweden, Taiwan, the United Kingdom and the US.

When observing both dendogram, one can notice that generally, the term-pair application seems to have more balanced clustering: Clusters are more equivalent in sizes, and the distance of the clustering (y-axis) is well distributed.

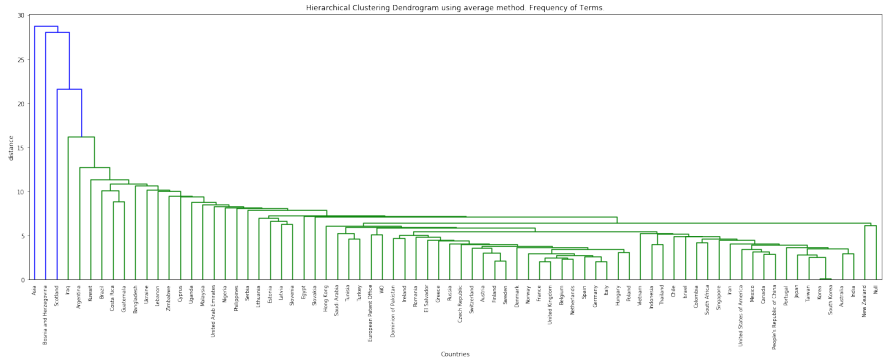


Figure 4.34: Hierarchical Clustering applied to the country level: Using term frequency approach.

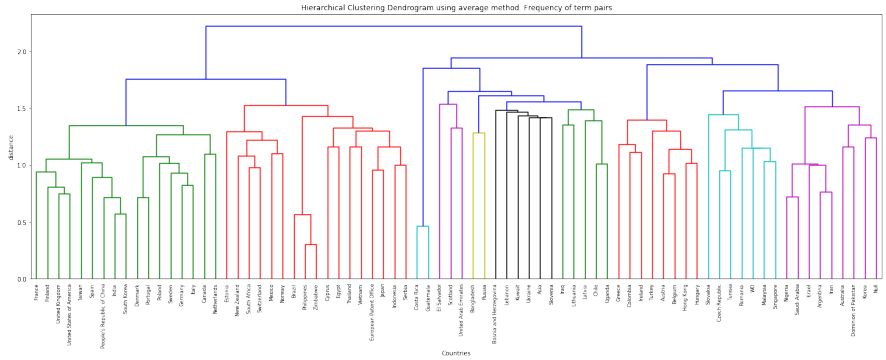


Figure 4.35: Hierarchical Clustering applied to the country level: Using term-pair frequency approach.

CHAPTER 5

Discussion and Perspectives

This chapter of the project will focus on creating a discussion and providing the author's perspectives on the results described in the previous chapter. Before diving into it, it should be noted that not all of the results will be discussed extensively, but only the ones that are related to the research questions. The chapter contains the following sections:

- Macro Level Analysis: A discussion and interpretation of the world-level results, including static and dynamic year-to-year analyses.
- Meso Level Analysis: A discussion and interpretation of the most relevant country-level results.
- Micro Level Analysis: A discussion and interpretation of the most relevant organisation-level results including important considerations on how organizations interact in relation to their technological capabilities related to biofuels.
- Complementary Analysis: A discussion of the outcomes of the patent and publications analysis as well as an interpretation of the alternative methodologies used.

Each section will finish with a discussion of limitations.

5.1 Macro Level: Time

A look at the whole dataset

The first result worth discussing is related to the application of the developed model to the whole database overall years.

When applying clustering to the capability matrix of the entire database, the natural relationships between every term can be observed. As described, the most related terms appear together in this clustering. For example, terms related to sugar, straw, waster or wood, show a high relationship in terms of usage. This serves as a validation and proof that in fact, the clustering makes sense. It is interesting to discuss two different phenomena that resulted from this analysis. First, most terms are clustered because of their scientific similarity, this happens particularly to feedstock terms. Secondly, some other terms such as outputs, are clustered in a not so linear form. For example, biogas, ethanol, and other types of fuels are also clustered. This second clustering can be due to a variety of factors, one possibility is the fact that they are researched using the same feedstocks or processing technologies. Another possibility is the fact that they are clustered for being intensely researched, and the "the goal" of most technological assets.

Correlation of Years

When observing the capability correlation matrix of the years, and particularly its clustered version, one cannot help but notice that as time goes on, years are increasingly related to each other. For example, the last 7 years (2010-2017) form a cluster where their correlation is of at least 70%. On the other hand, from 2004 to about 2009, years have on average a far inferior correlation between themselves. Finally, the years before 2004, make up the cluster of less related years. This might indicate a decrease in the experimentation in the field of biofuels research in the years of 2004-2009, and a consequent consolidation of the field in the modern years. This increase in consolidation can be related to factors such as: globalization, increase of knowledge sharing, the prevalence of information systems such as the internet, or even others.

Looking at the correlation of years with the year before, there appear to be years that somehow are almost unrelated to years before. From 1999 to 2001, the correla-

tion seems to increase until it drops considerably in 2002, and then regularly increases. This seems to indicate that there are "gaps" or "breaks" in scientific research particularly apparent in the years of 2002 and 2006. These gaps might be related to a change in scientific direction, the appearance of a breakthrough technology, or more generally, something that leads researchers to stop using a certain set of term combinations and start using a different one. Moreover, this correlation behaviour might be connected to the fact that in earlier years, less data is available, and therefore sparser, leading to a lower correlation (see limitations).

Another interesting concept worth noting and that can be easily validated is related to the concept of first, second, and third generation biofuels (Saladini et al. 2016). First generation biofuels are mainly related to crops and feedstocks related to human consumption, second generation biofuels are related to non-human consumable feedstocks and third generation biofuels which are related to aquatic feedstock. The normalized evolution of terms is compatible with this concept. When looking at the evolution of algae, sugar and wood, one can notice that sugar has a considerable usage in early years, wood follows with a high frequency after 1995. Finally, in recent years, the prevalence of the term algae is clear.

In summary, there are several models and reasons that can explain the relationship between the years and the evolution of the research landscape. One should note that the terms in the framework are based in all years; therefore, it is normal that the matrix is more and more saturated as the years go by.

Comparison of Years

When comparing two years in the database and the main term pairs that make up their capability matrices. Some results are worth discussing.

In the case of the comparison of 2010 and 2017, the years have the same term pairs generally speaking. However, the percentages of these term pairs can differ greatly. This leads to the belief that the research ecosystem is more "distributed" than older years. This somehow confirms the considerations in the previous section.

Furthermore, it is also interesting to note some areas of research that have highly increased or decreased in frequency. This behaviour supports the fact that some areas of research lose or gain interest over time. The reasons behind the behaviour can range from a general loss of interest, commercialization, or general trend towards a

particular research area. Anaerobic digestion for example sees an increase of about 1%, which can indicate a focus of the ecosystem on this particular processing technology. Although 1% might not sound like such an important increase, if the global panorama is considered, this has weight.

The comparison of two years and the calculation of the Pearson correlation index, also serves as a proxy for their similarity. This similarity (or lack of) can be due to a high number of small changes in term pair usage (likely), a number of large changes (less likely), or a combination of both (likely).

Contextualization

The contextualization part of this analysis is perhaps one of its most interesting aspects. This is because it tries to understand the relationship between certain external factors and the research ecosystem.

The adjusted price of barrel of oil is an indicator that was chosen mainly due to its reach as an external factor. The price of oil is known to influence a series of macro and micro economic indexes. When comparing it to the frequency of output terms over the years in the research ecosystem, some correlations are surprisingly high (butanol 85%, bioplastic 80%). A positive correlation means that the higher the price of oil, the bigger the number of produced records with that term. Several factors could explain this:

- Oil becomes more valuable, and therefore biofuels become more valuable as a way of addressing competition. This could lead to increased research.
- Oil becomes expensive, and companies look for alternate fuels, therefore increasing R&D focus on alternatives.

Butanol, biodiesel, and bioethanol are all used as alternative fuels, and all have a correlation of at least 70%. To understand more exactly the nature of this relationship, more research would need to be carried out, perhaps even a time series analysis.

The price of the kilo of sugar and its influence on the biofuel research ecosystem comes to evidence a more important series of facts, but also questions. One would naturally think that the price of sugar affects “sugar related” research. The analysis carried out confirms this: sugarcane and cellulosic sugars for instance, have an 80%

correlation with the price of sugar. However, when trying to understand exactly why if the price of a good increases, the research using that same good also increases, some questions surface. One would hope that if the price of goods increase, access to them becomes more economically challenging, and therefore research using them would decrease, causing a negative correlation. An opposing view is that as these goods become more valuable, the interest in it from an R&D perspective becomes greater. For now, this research can confirm that a relationship between the price of goods and their research exists, but the nature of it remains to be determined.

Limitations

When studying the system at a macro level, there are a few limitations that should be pointed out.

Firstly, the study is related to the entire system in a given year. This might not seem big, but in fact, the study is related to all of the patents and publications, from every organization, in every country in the world. One can therefore expect that a big part of the chronological behaviour is rather "smooth". This is because it is very possible that a certain phenomenon self-corrects when looking at the whole world. The system per se can be categorized as saturated.

Secondly, the volume of data along the years is not regular. In fact, there is an explosion of the number of technological assets in the more recent years. Although this was addressed by normalizing, one cannot forget the fact that the quality of documentation and general access to information has improved in the last years, causing recent documentation to be far richer than previously . This might make the analysis rather unbalanced, chronologically.

Finally, the use of the Pearson correlation index can also be seen as a limitation. Although this indicator is mathematically accurate in describing the relationship between two functions, it can be scarce in providing more information about the exact nature of a relationship. When analyzing the influence of external factors such as the price of sugar, carrying a time series analysis would allow a better understanding of the nature of the behavior observed.

5.2 Meso Level: Countries

Country Correlations

When observing the correlations in the correlation matrix produced in the results there is a wide range of possible interpretations. However, some hypothesis can be formulated as to the configuration of this clustering.

Language can play a big role as to the correlation of countries: the UK and USA have a 68% correlation, China and Taiwan a 70% correlation. Distance between countries also seems to be a source of correlation between them; Costa Rica and Guatemala have between themselves a correlation of 71% and a low correlation with every country with the exception of Brazil. However, there are some relationships that can be due to other factors such as economic partnerships (South Korea and USA - 65%), climate (Brazil, Zimbabwe and Philippines >80%), or simply the volume of research activity (the case of India). Moreover, some relationships would need a higher level of investigation to be understood, such as the high correlation (>58%) of Denmark with Portugal and Spain.

GDP per capita

In order to provide further context as to the reasons for the correlation among the capability of different countries, as suggested by literature, the GDP per capita of countries was extensively used.

One general trend that is worth commenting on is the fact that most country pairs are locked in an "area" characterized by a low GDP per capita difference (<40000$) accompanied by a low capability correlation among themselves (<40%).

The highest capability correlations are more frequent as the GDP per capita difference decreases. In fact, countries such as Brazil and Zimbabwe, which have a similar GDP per capita but a very high capability correlation (88%). On the other hand, some country pairs such as Denmark and India, or the US and China counter this trend, with high divergences in terms of GDP and high correlations, due perhaps to the unbalanced economic development of certain countries.

Generally speaking, the introduction of the average GDP per capita of a country pair comes to confirm the fact that lower GDP differences are related to a higher

capability correlation. This can be related perhaps to the fact that countries that are closer to each other tend to have a more similar economic capability, and more collaboration, but this cannot be said for certainty.

Collaborations

When studying the collaborations and similarity of countries a general trend surfaces: most countries are different in terms of capability and do not collaborate among themselves. This was expected.

All countries that are similar and collaborate are European Union countries (Portugal, Germany, Austria, Italy, and Belgium). This could indicate that the EU is a driving force of collaboration but that collaboration is not necessarily "innovative", since the countries are similar in terms of research. Moreover, in the global landscape, countries that are similar tend to not collaborate, which seems counterintuitive, but for purposes of innovation makes sense.

Furthermore, most collaborations are made of countries that are different in capability (Canada-Kuwait, El Salvador-Germany, France-Lebanon), rather than similar. This would indicate that most countries decide that collaboration is a way of accessing a research space that they have previously not seen, instead of a way of intensifying the research into their core areas of interest. This is the example of pairs such as France-Lebanon, Canada-Kuwait, Germany-Ukraine, Hungary-UK. The factors that lead to these relationships seem to be a mix of historic, economic and political factors.

Comparing Countries

When comparing countries, in the particular case of Brazil and Denmark there is one interpretation that could be made. The top term pairs in the Denmark capability matrix are highly related to outputs or processing technologies such as biogas, ethanol, or fermentation. The only feedstock term that appears in Denmark's top terms is "straw". On the other hand, looking at Brazil, there is a high prevalence of sugar, sugarcane and other feedstocks.

This can indicate that countries that have a higher prevalence of a certain industry/ raw material tend to focus their research in that particular term and what they can use it for. While on the other hand, countries that are less reliant on a particular

raw material tend to focus on the outputs and results of that research, without giving so much importance to the raw material that is used.

Country Spectrums

The country spectrum is enlightening as to the extreme heterogeneity and inequality in the world of research.

Most of the spectrum is empty which indicates that most countries have very little research when comparing to others. In contrast, the US is particularly impressive as its spectrum is almost entirely completed. This would mean that the US uses almost all term pairs that appear in the database. This can be due to their tradition in leading the research field, their economic development or even the intensive patenting culture. Other countries such as China or India only come close it.

The spectrum also shows if a country is adopting one of two strategies: focusing on certain areas, or distributing its capability across areas. Brazil for instance, seems to follow the first strategy distributing its interest in certain areas across the spectrum, one of them probably related to sugar. The second strategy is the one use by bigger players such as the US, India, or even China which have widely spread research interests.

Uniqueness

The uniqueness index comes to confirm the previous premise, the US leads with almost 50% uniqueness. Which means that most term pairs used by the country are only used by it. This shows not only the intensity of research but also the amount of innovation and accessing unresearched areas. When comparing to countries of similar size, such as China, one can say that the US is 40% more unique than China (10%), which sounds impressive. However, does more term usage necessarily indicate more innovation? Or more saturation?

Interestingly, there is a very large number of small countries in the uniqueness ranking such as Ukraine, Lebanon, Cyprus, or Bangladesh. This could mean that countries with no particular research intensity and low economic development tend to focus on areas that are special to them, or their location, and relatively “exotic”. Possibly this is made through collaborations (France-Lebanon example) where a more

established country accesses an untapped area through a small country with a particular capability, using it as a “research proxy”.

Limitations

The first limitation worth mentioning is the usage of the Pearson correlation index. Similarly to the macro analysis, the Pearson correlation index can be criticized by not giving the full picture when comparing countries and establishing a correlation matrix between them.

The second limitation is related to the use of the GDP as contextual information additional to the correlation index. Throughout the literature, several other economic, social and political indicators are used to explain the innovation in countries (Filippetti, A., Peyrache, A., 2011). Therefore it can be seen as naive to only use one indicator. It is interesting to think of other indicators that could further explain the relation between technological capabilities of countries: language, population, education, or even geography. Moreover, the relationships might also be a combination of different factors rather than a consequence of one.

The third limitation is related to collaboration. Throughout the analysis, a division was made between countries that collaborate and countries that do not. However, there is no quantitative index for “collaboration”, only the number of collaborations between two entities, For example, should countries that only worked once together, be seen as countries that collaborate, or should it be only considered as collaboration if they collaborate more than X times?

The fourth and final limitation worth mentioning is the spectral representation used to display the capabilities of countries. The sheer amount of term pairs in the database makes the visualizations rather difficult. Although an effort was made to reduce this number, some countries appear as almost empty spectrums while in fact, they used term pairs (as can be seen by the data).

5.3 Micro Level: Organizations

Organizational Correlations

When creating the correlation matrix of organizations and directly comparing it with the matrix of correlations at the meso level, several differences stand out.

Generally speaking, the correlations of organizations are more scattered and spread out. For instance, the matrix is made of a higher number of clusters of correlation that vary in the size of its organizations. The cluster that contains the Technical University of Denmark for example, is quite diverse: with presence of universities from China, the US, Singapore, Portugal and Sweden. This particular cluster is only made of universities and contains no businesses or associations. Other clusters follow this pattern, the fact of being made of only a particular type of institution, which means that universities and the industry still have a "gap" between them.

When focusing on organizations another pattern emerges, these are rarely highly related to other organizations, and instead only strongly connected to a smaller number of institutions. Finally, one can say that universities tend to aggregate themselves in clusters of innovation, while organizations tend to be more dynamic and targeted in their capabilities. This because the correlation matrix seems to separate universities and businesses. Thinking of the university-industry relationships, these relationships seem to be in the early stages of development: organizations are still "shy" in what regards the targeting of capabilities, while universities are more open and related to themselves.

Comparing Organizations

The organizational profiles of correlation are very good proxies for the innovation strategy that organizations follow. On one side, universities have a fuller spectrum with a higher average correlation, which indicates a more scattered field of capability. On the other side, organizations have a lower average which indicated that they focus on particular areas of interest in a more "closed" fashion.

The Technical University of Denmark is a good example of the above. The top 5 organizations that are more related to it, 3 are Chinese universities, with correlations upwards of 65%. The Chinese university of Tsinghua follows the same pattern. On

the other hand, American universities are characterized by higher collaboration with associations (ARS, CSIR) and other American universities.

Finally the comparison tables between DTU and Tsinghua Universities are a good example of how a high correlation expresses itself. The top 10 tables of term pairs have 7 term pairs in common. This indicates that the research strategies of these two universities are extremely aligned, at least in terms of area. This could also be an expression of research partnerships, which are known to exist between these institutions (Tsinghua 2018), and would be a good quantitative measure of the success of these partnerships.

Collaborations

The characterization of the nature of collaborations has several aspects worth discussing.

On the first hand, most universities (more than 90%) appear to have at least some sort of collaboration with other universities, which is an expression of the above hypotheses. Danish universities are particularly well classified in the spectrum with all having at least 60% of collaborations made with other universities. This same pattern occurs in Chinese and American universities. However, the bottom 4 universities that collaborates the least with other universities are the 2 Brazilian and the 2 Canadian institutions. This might be related to several factors: the fact that these countries are vast in size and their universities are "isolated", or the fact the organizations in these countries do not see value in academic collaboration. A report was found (Montreal Metropolitain 2011) where 37% of Canadian organizations did not find collaboration with universities to be relevant for example.

On the other hand, while observing non-universities, the landscape is quite the opposite. Most organizations simply do not collaborate with universities of any sort. This is more evidence of the fact that organizations are still closed to outside ventures. Of the organizations that collaborate with universities, most are associations such as ARS, ACAD, or USDA ARS. Which could mean that associations serve as a bridge between universities and the industry. Meanwhile, Danish organizations such as Novozymes or Dong Energy tend to go against this trend by collaborating more intensely with universities, this might be due to cultural factors (more openness to industry-university collaborations in Denmark), or strategic factors.

In summary, perfect university-industry collaboration seems to be still in its very early stages. Universities form clusters of academic partnerships and collaborations, but industry is still cautious, and reserved in opening up. Denmark is one of the countries "swimming against the current" and investing more seriously in these types of collaborations which is clearly shown by the data.

Limitations

The first limitation that was the most challenging in this subchapter of the analysis was related to the database itself. The AMICa pathfinder project successfully organized a database where every asset was connected to an owner. However, this characterization is not perfect, particularly in the names of organizations across the database, some names are not particularly well formatted, and other organizations are repeated under different names. This is a consequence of the fact that the AMICa project mixed a number of different databases. Although this does not seem to affect the "big picture" perspective, it could have some repercussions such as misinterpretation etc.

The collaboration part of this analysis was based on the simple detection of the string 'Univ' in an organization's name. Although this serves as a good proof of concept, not all universities and academic institutions use 'Univ' in their name, which could distort some of the collaboration numbers.

Finally, throughout the micro analysis, a distinction between Universities and Non-universities was made. However, the world is made of a wide range of different types of organizations: universities, businesses, associations, research centers, and more (Etzkowitz, 2001). Therefore, this simplification is limitative on the type of conclusions that could be drawn.

5.4 Complementary Analysis

Publications and Patents

When discussing the patents and publications analysis, the analysis tried to understand if they are unbalanced or if there is a presence of out of the ordinary behaviour.

Normally, one would expect scientific publications to lead the way in quantity: this is confirmed by the data. In fact, there is sharp rise in the number of scientific publications related to biofuels from 2006 to 2016. But contrary to what one would expect, patents are at the origin of this explosion, in fact, between 2006 and 2011, patents were more important. What appears to happen is that patents are guiding the direction of research, when their number rises; publications tend to rise, with some delay. When patents fall, publications fall sharply. This could be explained by the fact that biofuels is an "industry first" technology and that it is also becoming a more mature technology which is way past its exploration phase. This pattern repeats when looking at particular feedstock terms, where a small change in the rate of patenting, seems to disturb the rate of research on that particular term, with a certain delay.

When looking at the rate of patenting and the rate of publication of every term, one can notice the terms that are more mature and the ones that are still in the research/publication stage. Generally speaking, there seems to be a balance between patenting and publication of every term. One would expect some terms to be lost in the research phase and not adopted by patenting, but the imbalance of some terms (such anaerobic digestion) might be indicative that these are still in its research phase and will be adopted later.

Some of the limitations of this complementary study include the fact that an analysis on the global ecosystem is being made, which is quite saturated on that scale. Moreover, in order to really understand the relationship between patenting and publications, a time series analysis would be perhaps more telling. Finally, the fact that terms instead of term-pairs were used for this analysis can also be seen as a limitation, since the observation of term-pairs tends to add more granularity. This of course would have to be done carefully due to the very high number of term pairs.

Alternative Methods

One other part of the complementary analysis was the application of some unsupervised learning techniques to the data, in order to find some patterns.

However, both the application of the principal component analysis and the t-sne algorithm showed that these techniques are not as directly applicable and explanatory

as the system perspective. This is because the clustering of different countries was not directly evident from the application of the transformations.

An obvious limitation of this study is the study level for this part of the analysis. A very large number of other techniques for clustering, association mining, and density estimation exist for inferring a structure to data. However, the application of these techniques was deemed to be out of the scope of this thesis.

Clustering Quality

When comparing the quality of the clusters made with term frequency and term-pair frequency, the fact that term pair produces a much cleaner and reasonable clustering is quite obvious. In fact, just as predicted by the AMICa pathfinder briefing, a term per se is much less explanatory than the several associations to it. For example if only observing the term “anaerobic digestion”, all of the countries that use this processing technology would be deemed similar. But if we observe the term pairs “sugar/anaerobic digestion” and “wood/anaerobic digestion”, then the clustering will consider these as different features and cluster the countries that use each of them separately, therefore providing much more granularity to the analysis.

Of course, judging the clustering quality only conceptually and visually can be seen as a limitation. Techniques such as the Rand index, the Jaccard similarity, or the entropy of partitions exist to provide context on the clustering made. Although the application of these indicators is not in the range of this project , these could be applied with relative simplicity.

CHAPTER 6

Conclusion

In the final chapter of this project, a conclusion on the most important aspects of it will be made. The main focus of this part will be on the implications for academia and industry. The sections in this chapter are:

- Implications for theory: Here, the most relevant connections between the literature review and the project will be pointed out.
- Implications for practice: An overview of the main consequences this project will have for policy makers, industry stakeholders and researchers.
- Further Research: By taking the limitations and conclusions as a starting point, considerations on further academic research will be drawn.

6.1 Implications for theory

Macro and Meso level

When revisiting the academic study of national innovations systems of countries, most research focuses on the volume of technological assets, such as patents, produced by a certain country. Throughout this project, in order to characterize the innovation or capability of nation, the focus relied on the usage of field related term-pairs. This allowed us to obtain a more nuanced view on how different countries and regions conduct research on a certain industry topic. Although the volume of patents is a good indicator of the intensity of the research from a certain country (Furman, Porter, and Stern 2002), it fails to describe exactly how this country innovates. Furthermore, the analysis of industry terms allows for a quantitative comparison of countries, which the volume of patents does not. For these reasons, it is fair to deduce that in order

to fully understand the nuances of a country's research ecosystem, more focus should be put on industry specific terminology.

The literature review clearly showed that researchers have realized that innovation capability is related to a number of external indicators: GDP per capita, RD spending, openness to international trade etc. (Filippetti and Peyrache 2011). Although this project does not refute this in any way, it does allow for a simple verification of the relationship between external factors and the national innovation of countries. Just as was done with the GDP per capita and its relation to the capability correlation matrix, this project provides a framework which allows the usage of any external indicator and the consequential visualization of its influence in the research ecosystem. By doing this, the technological capability framework provides an additional tool that can help to find the perfect national innovation capacity indicator.

Most NIC studies focus on the evolution of the patenting volume over the course of a number of years. (Furman, Porter, and Stern 2002) It is worth noting that by combining the macro and meso levels of the analysis, it is possible to understand how the capabilities of countries change over time. For example, just as the capability correlation matrix of countries was produced, one could create one of these for each year, thus understanding in a deeper manner, the behaviour of the system.

Finally, the last area of the literature review related to the macro and meso level touched on Hidalgo's view on economic complexity(Hidalgo and Hausmann 2009). This project comes to validate the importance of the complexity of an ecosystem as a way of understanding its behaviour. Furthermore, just as the economic complexity view, this thesis focused on understanding the innovation systems as a network and the study of its characteristics. This could mean that the economic complexity methodology and way of thinking are applicable to other areas, such as the study of technological capability and innovation. The value of the system lies in its complexity, not in its volume or size.

Micro Level

In what concerns the micro or organizational level, the main focus of the literature review relied on the understanding of how open innovation is expressed in the ecosystem of research.

The first part of the analysis touched on open strategy as a crucial part of the open innovation process (OECD 2008). When looking at the correlations in the industry and the collaboration landscape between organizations, one cannot help but state that open strategy seems to be, at least at this moment, a utopian view. Although universities are intensely collaborating and seeking knowledge outside their premises, businesses are focused on targeted innovation topics and are very skeptical to the idea of open collaboration with universities. This would mean that at least in the biofuels ecosystem, businesses are betting on other types of open strategies, or simply not interested in pursuing that type of strategy.

In what regards patenting and appropriation of intellectual property (Pénin and Wack 2008), this project has shown that although patents influence scientific research, the number of patents has decreased over the last 4-5 years. However, the project does not seek to understand if this is due to a smaller rate of patenting or simply the expression of the maturity of a certain technology.

Finally, the study of the nature of the collaborations in the micro level presents a simple proof of concept that aims at quantifying exactly the triple helix relationships (Etzkowitz and Leydesdorff 2000). Researchers have stressed the importance of industry-academia-government relationships through case studies and examples. This project directly quantifies these relationships as a quota of the collaborations of every organization, thus quantifying the helix. However, in the biofuel ecosystem, this helix seems to be still in its embryonic phase, at least from the industry side.

Engineering Systems Perspective

This project serves as the direct application of the engineering systems perspective (Weck et al. 2011) to different levels of analysis of the biofuel research ecosystem.

The first implication that this project has towards theoretical studies on the topic is the fact that the engineering systems approach is a valid approach to describe, understand, and study the behaviour of a research ecosystem. Through the application of a network or graph based model, visualizations and quantitative studies (Albert-László Barabási 2016) allowed to deepen the knowledge of the system.

A second implication that this project might have is the demonstration of the scalability of the engineering systems perspective. By establishing a rigid framework

of term pairs, and filtering the input of documents, the same set of quantitative tools can be applicable to different levels of the system.

The third and final implication is the under explored relationship between big data analytics and the engineering systems view. The access to a large amount of information and the understanding of it, is a very serious modern challenge. By combining the raw information of big data (*EURITO, 2018*) and the structured thinking of an engineering systems perspective, several advances can be made in understanding how the world works.

6.2 Implications for practice

General outcomes

One of the objectives of this work and the AMICa project is to support data-driven decision making in industry and the public sector. As a result, this work is has implications for practitioners in both sectors. The following are some general implications that this project can have:

- The methodology applied to the biofuels research landscape serves as a **tool that allows the understanding and characterization of any research area**. It serves as a set of guidelines that will transform a database of technological assets into a characterization of the several units that make it, and how these are connected.
- The workflow applied during this thesis also allows, as demonstrated, to **study a particular research area at different levels**, macro, meso and micro.
- This thesis also provides an example of some of the many visualizations and characterizations that the AMICa pathfinder project can have, and bridges the gap between industrial mapping and the practical implications of such mapping.

Countries and International Organizations

For countries and international organizations, some of the main practical implications follow:

- This project provides **answers to relevant international level questions related to the research ecosystem of biofuels**. Through visualizations and quantifications, one can think of a multitude of other questions that could also be answered by the application of the same tools.
- Support of **transparent decision making** was mentioned as one of the main goals of the project. By developing this research, publishing it, and open sourcing the code that was developed, this framework can be used as a platform that not only helps countries in understanding the ecosystem but that also supports transparency in the decision making process. This because of the high quantitative degree of the analysis.
- By using the work developed in this thesis, countries can use the framework as a way of **evaluating their political strategies**. For example, the 'similar and collaborating' countries is proof that the European Union, to some degree, influences countries to collaborate: this research quantifies it.
- Finally, this project also allows for the **discovery of untapped research areas and the raising of new questions about the research ecosystem and its functioning**. This is critical for the sustainable development of research and the advance of the scientific field.

For organizations and industry

Some practical implications for organizations, the industry and researchers:

- The first practical implication is that organizations and businesses are now capable of **quantifying and evaluating the success of their innovation strategies**. Similarly to what can be done with financial statistics, organizations and universities will now be able to see if collaboration or investment in a particular area actually materialized or not. Moreover, this project will allow institutions to really understand their capability DNA.
- This analysis will also allow organizations to develop **market analyses that will allow them to elevate the knowledge on a particular research subject**, such as biofuels. This contextualization is crucial to the development of new innovations in the field.

- Furthermore, as for countries, organizations and researchers will be able to find **untapped areas where innovation is possible or new partners that suit a certain strategy**. This work thus creates a sandbox for innovation exploration.

- Finally, this framework allows for the **quantification of university-industry-government collaboration**, and consequently gives a method to quantify it. This is one way in which this project allows a push for open innovation and industry wide collaboration.

6.3 Further research

The final part of this chapter will establish some recommendations and ideas for further research on this topic. These recommendations are based on some of the limitations previously established, or were noted during the development of the analysis. A list of the most important ones follows:

- The most important further research topic is the expansion of the framework and the experimentation with it in other research areas. As a proof of concept, this project was based on the biofuels ecosystem; however, it's certainly interesting to think of the possibilities of its application to other fields. For instance, after the design of a database one could easily use exactly the same approach.

- The second area of research is related to the relation between external factors (price of sugar and oil) and the production of technological assets. To study this, this project used the Pearson correlation, which produced very interesting results. However, a relationship of this nature should be studied using a time-series analysis, in order to understand exactly how the behaviour changes (or not). The application of a time series analysis and a wide range of external factors is a very interesting research topic. From its application, one could understand exactly what moves the research field.

- Although the GDP per capita and the GDP difference between countries was widely used, academia has shown that these indicators might be just the tip of the iceberg in explaining the innovation systems of countries. One possibility is

to dive into this topic is the application of a very large number of indexes (GDP, HDI, alphabetisation, number of universities etc.) and to study their influence when comparing to the capability correlation matrix. This is a direct way of understanding the impact of every index. Moreover, one could mix different indexes and see which combination better explains the capability correlation matrix.

- The uniqueness of countries and of their usage of term pairs served as proof of concept to a notion of "specialization" of countries in a certain area. This could obviously be expanded to periods of time, or organizations. What made this year, organization, special? The notion of uniqueness is very valuable in the context of innovation, and further work should be developed in this area.

- In the study of patents and publications, the volume of terms of every type was used. A more interesting and granular analysis would be to conduct the same study but with the usage of term-pairs. As previously demonstrated, term pairs provide a richness to the analysis that term frequency simply does not deliver. Also regarding the complementary analysis, the application of more unsupervised learning techniques in order to discover other clusters is definitely an interesting area of research.

- In what regards the study of the nature of collaborations, a very rudimentary analysis was carried out by the detection of the string 'Univ' in the name of organizations. One can envision a deeper study into the types of collaborations in the industry, which would serve even better the triple helix quantification. Here, a distinction between university-industry, industry-government, and government-university could be made. To do this, several natural language processing techniques (such as regex) would have to be used.

APPENDIX A

Figure Appendix

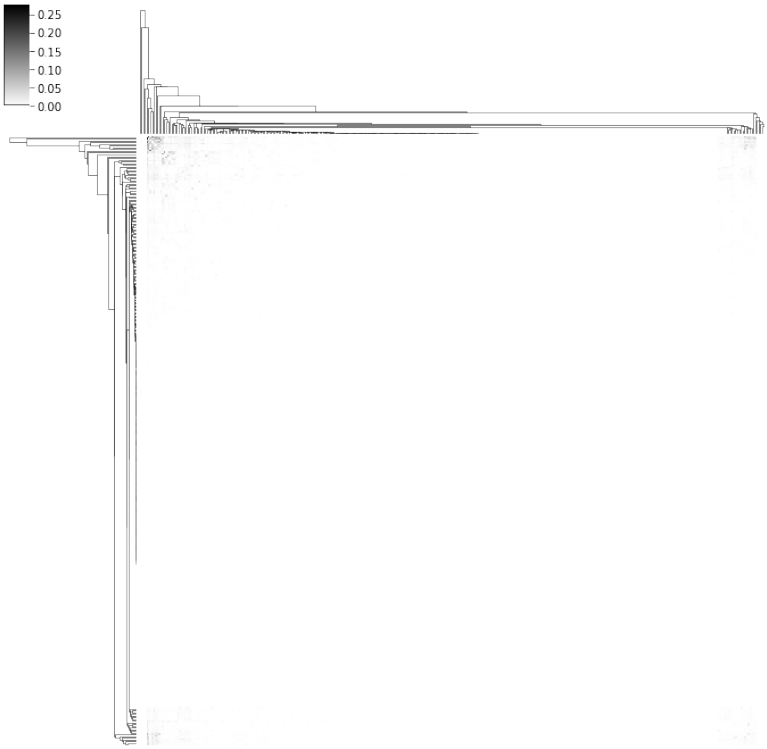


Figure A.5: Clustered version of the AMICa database capability matrix.

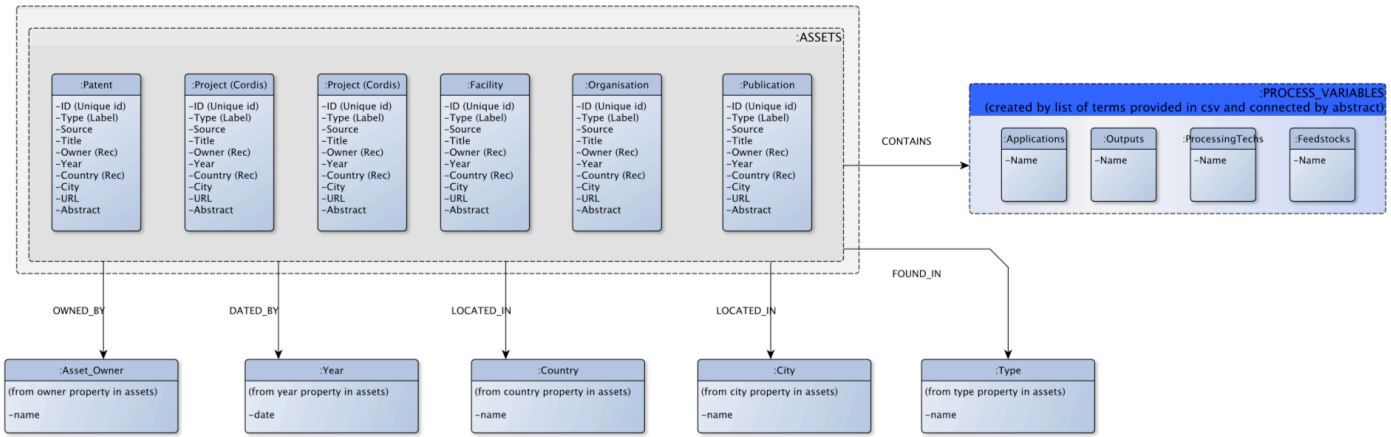


Figure A.1: Simplified data model retrieved from the AMICa pathfinder repository (*AMICa, 2018*).

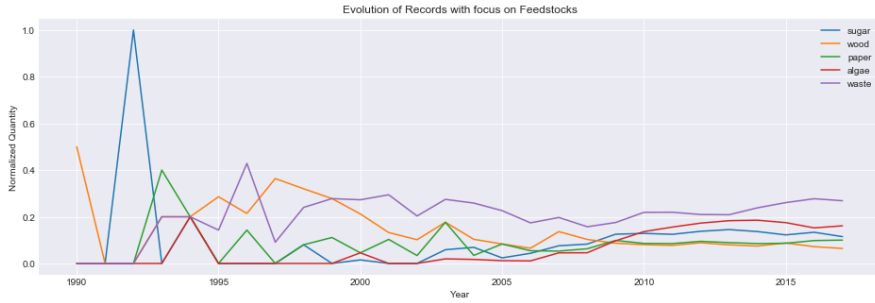


Figure A.2: Normalized quantity of terms over time: Feedstocks.

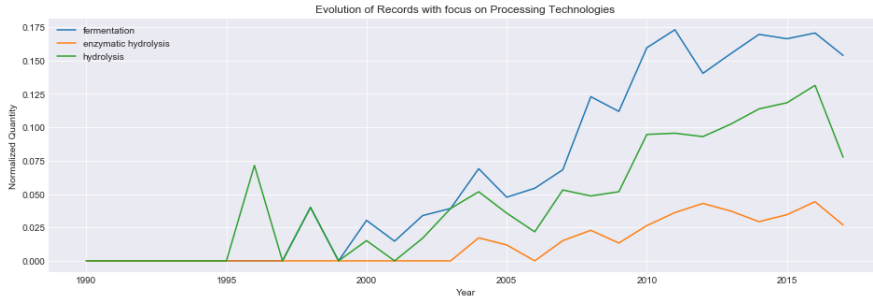


Figure A.3: Normalized quantity of terms over time: Processing Technologies.

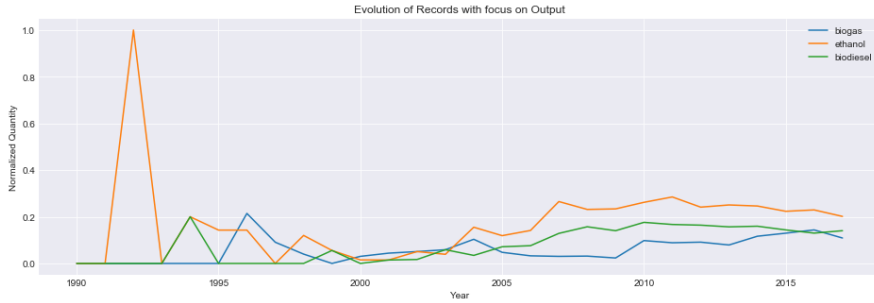


Figure A.4: Normalized quantity of terms over time: Outputs.

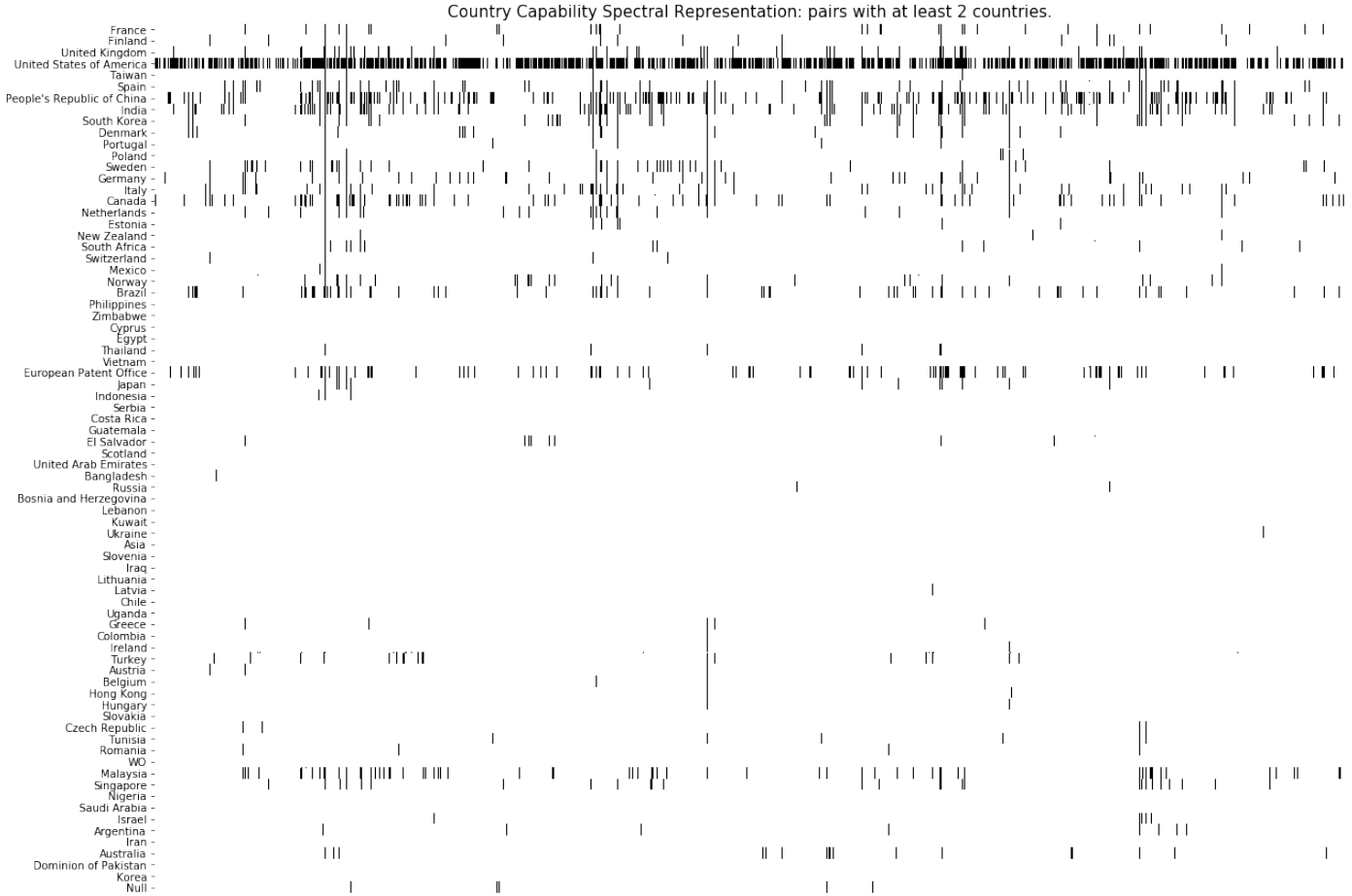


Figure A.6: Country capability spectrum for all countries (only term pairs that appear in at least 2 countries).

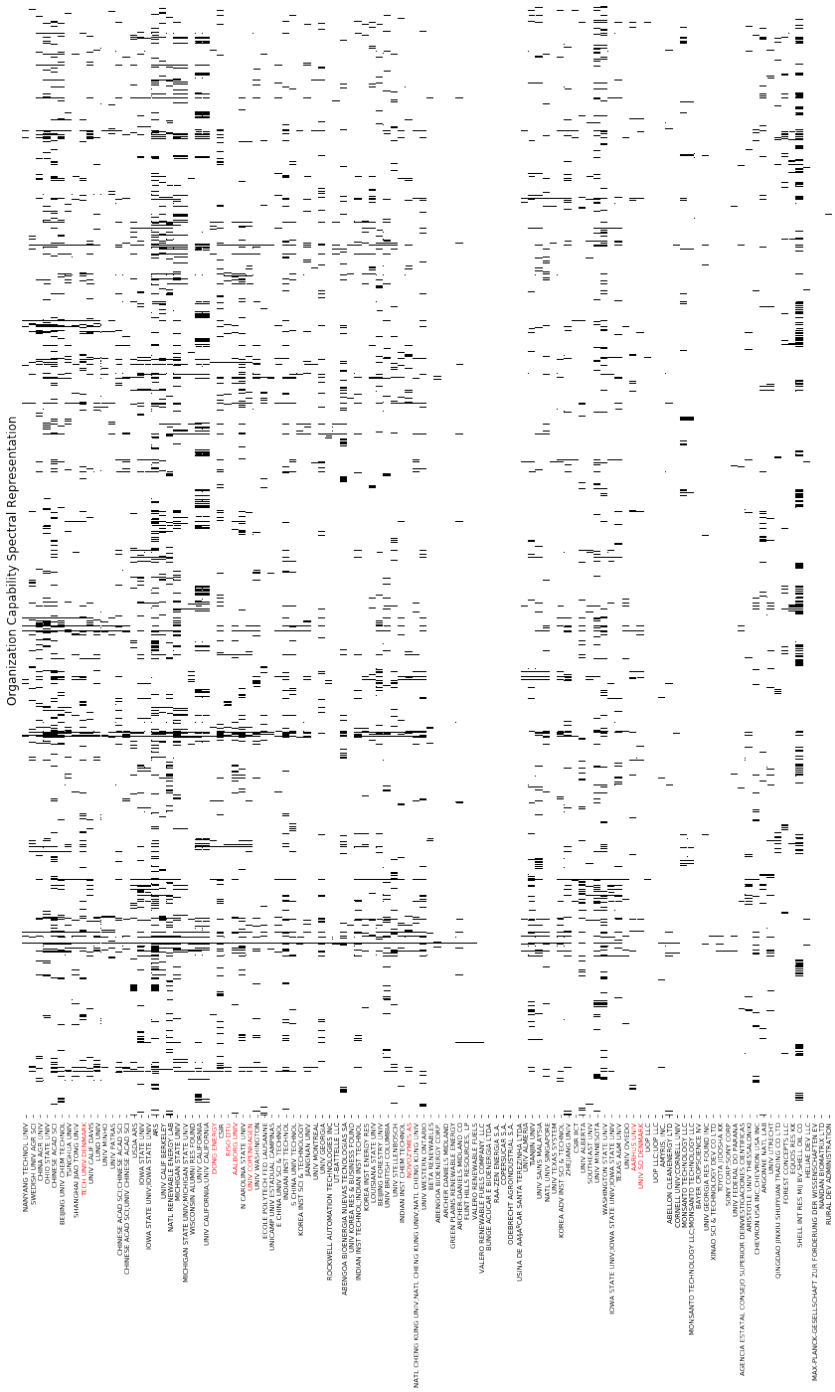


Figure A.7: Organizational spectrums: Complete spectrum visualization.

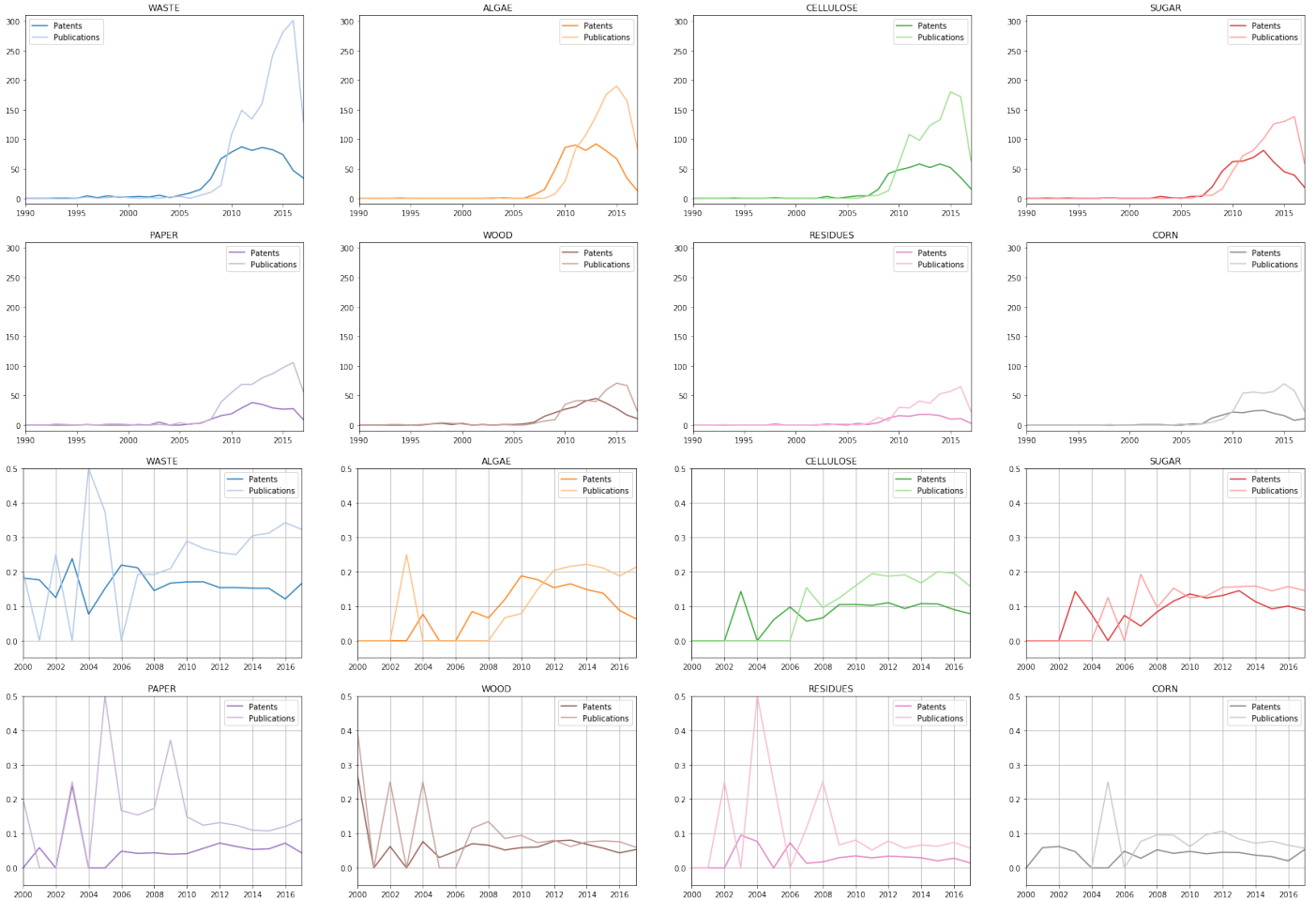


Figure A.8: Absolute and relative quantity of assets with a certain feedstock term over time: Patents and publications.

Bibliography

Albert-László Barabási (2016). *Network Science:* February, p. 456. ISBN: 978-1107076266. URL: http://barabasi.com/book/network-science{\%}5Cnhttp://www.amazon.com/Network-Science-Albert-L-225-Barab/dp/1107076269.

AMICa, 2018. URL: https://amica-pathfinder.net/ (visited on 05/24/2018).

Assis Mota, Alexandre, Lia Toledo Moreira Mota, and Anfre Morelato (2007). "Visualization of power system restoration plans using CPM/PERT graphs". In: *IEEE Transactions on Power Systems* 22.3, pp. 1322–1329. ISSN: 08858950. DOI: 10.1109/TPWRS.2007.901118. URL: http://ieeexplore.ieee.org/document/4282007/.

Bahar, Dany, Ricardo Hausmann, and Cesar A. Hidalgo (2014). "Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?" In: *Journal of International Economics* 92.1, pp. 111–123. ISSN: 00221996. DOI: 10.1016/j.jinteco.2013.11.001. URL: https://scholar.harvard.edu/files/dbaharc/files/bhh-jie.pdf.

Bessant, John (2008). "Dealing with discontinuous innovation: the European experience". In: *International Journal of Technology Management* 42.1/2, p. 36. ISSN: 0267-5730. DOI: 10.1504/IJTM.2008.018059. URL: http://www.inderscience.com/link.php?id=18059.

Blessing, Lucienne T.M. and Amaresh Chakrabarti (2009). *DRM, a design research methodology*. Vol. 1, pp. 1–397. ISBN: 9781848825864. DOI: 10.1007/978-1-84882-587-1. arXiv: arXiv:1011.1669v3. URL: http://books.google.com/books?hl=en{\&}lr={\&}id=KdR4OmWtQdIC{\&}oi=fnd{\&}pg=PA1{\&}dq=DRM,+a+Design+Research+Methodology{\&}ots=Ocn1JOFcN1{\&}sig=pc7NahqVKKe8U0{_}ugmMUAbtYm98.

Brill, James H. (1998). "Systems engineering? A retrospective view". In: *Systems Engineering* 1.4, pp. 258–266. ISSN: 1098-1241. DOI: 10.1002/(SICI)1520-6858(1998)1:4<258::AID-SYS2>3.0.CO;2-E. URL: http://doi.wiley.com/10.1002/{\%}28SICI{\%}291520-6858{\%}281998{\%}291{\%}3A4{\%}3C258{\%}3A{\%}3AAID-SYS2{\%}3E3.0.CO{\%}3B2-E.

Browning, T. R. (2001). "Applying the design structure matrix to system decomposition and integration problems: A review and new directions". In: *IEEE Transactions on Engineering Management* 48.3, pp. 292–306. ISSN: 00189391. DOI: 10.1109/17.946528. arXiv: arXiv:1011.1669v3.

Castellacci, Fulvio and Jose Miguel Natera (2011). "A new panel dataset for cross-country analyses of national systems, growth and development (CANA)". In: *Innovation and Development* 1.2, pp. 205–226. ISSN: 2157-930X. DOI: 10.1080/2157930X.2011.605871. URL: http://www.tandfonline.com/doi/abs/10.1080/2157930X.2011.605871.

— (2013). "The dynamics of national innovation systems: A panel cointegration analysis of the coevolution between innovative capability and absorptive capacity". In: *Research Policy* 42.3, pp. 579–594. ISSN: 00487333. DOI: 10.1016/j.respol.2012.10.006. URL: http://dx.doi.org/10.1016/j.respol.2012.10.006.

CORDIS (2018). *European Commission's community research and development information service*. URL: https://cordis.europa.eu/.

Crossref (2018). URL: https://www.crossref.org/.

Danilovic, Mike and Tyson R. Browning (2007). "Managing complex product development projects with design structure matrices and domain mapping matrices". In: *International Journal of Project Management* 25.3, pp. 300–314. ISSN: 02637863. DOI: 10.1016/j.ijproman.2006.11.003. URL: https://www.sciencedirect.com/science/article/abs/pii/S0263786306001645.

Data, Inflation (2018). *Inflation Data*. URL: https://inflationdata.com/Inflation/Inflation_Rate/Historical_Oil_Prices_Table.asp.

Dodder, Rs, Jm Sussman, and Jb McConnell (2004). "The Concept of the 'CLIOS Process': Integrating the Study of Physical and Policy Systems Using Mexico City as an Example". In: *... Engineering Systems ...*, pp. 1–48. ISSN: 0011-3131. URL: http://esd.mit.edu/symposium/pdfs/papers/dodder.pdf.

DTU (2018). *Engineering Management DTU*. URL: http://www.dtu.dk/english/Education/msc/Programmes/industrial_engineering_management.

DTU's strategy 2014-2019 - DTU. URL: http://www.dtu.dk/english/About/ORGANIZATION/Strategy (visited on 05/24/2018).

Eppinger, Steven (2001). *Innovation with the Speed of Information.pdf*. URL: https://hbr.org/2001/01/innovation-at-the-speed-of-information.

Etzkowitz, H. (2003). "Innovation in innovation: the Triple Helix of university-industry-government relations". In: *Social Science Information Sur Les Sciences Sociales* 42, pp. 293–337. ISSN: 0539-0184. DOI: 10.1177/05390184030423002. arXiv: 0911.2730. URL: //000185234300002.

Etzkowitz, Henry and Loet Leydesdorff (2000). "The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations". In: *Research policy* 29.2, pp. 109–123.

EURITO, 2018. URL: https://cordis.europa.eu/project/rcn/211945{_}en.html (visited on 05/24/2018).

Faber, Jan and Anneloes Barbara Hesen (2004). "Innovation capabilities of European nations: Cross-national analyses of patents and sales of product innovations". In: *Research Policy* 33.2, pp. 193–207. ISSN: 00487333. DOI: 10.1016/S0048-7333(03)00122-7. URL: https://ac-els-cdn-com.proxy.findit.dtu.dk/S0048733303001227/1-s2.0-S0048733303001227-main.pdf?{_}tid=ad647b70-188c-11e8-817f-00000aacb35f{\&}acdnat=1519385500{_}ad3ad89a5a4a416f9a0b08cd5a27fc43.

Filippetti, Andrea and Antonio Peyrache (2011). "The Patterns of Technological Capabilities of Countries: A Dual Approach using Composite Indicators and Data Envelopment Analysis". In: *World Development* 39.7, pp. 1108–1121. ISSN: 0305750X. DOI: 10.1016/j.worlddev.2010.12.009. URL: http://dx.doi.org/10.1016/j.worlddev.2010.12.009.

Freeman, C. and Soete, L. (2000). *The Economics of Industrial Innovation, Cambridge, MA, MIT Press*. 2nd. ISBN: 1844800938. URL: https://mitpress.mit.edu/books/economics-industrial-innovation.

Furman, Jeffrey L., Michael E. Porter, and Scott Stern (2002). "The determinants of national innovative capacity". In: *Research Policy* 31.6, pp. 899–933. ISSN: 00487333. DOI: 10.1016/S0048-7333(01)00152-4. arXiv: arXiv:1011.1669v3. URL: https://ac-els-cdn-com.proxy.findit.dtu.dk/S0048733301001524/1-s2.0-S0048733301001524-main.pdf?{_}tid=spdf-af419319-4cfb-4c5f-af1f-66fe11d57d41{\&}acdnat=1519380813{_}c003162172b639276aea6702fe9935ba.

Gassmann, Oliver (2006). "Opening up the innovation process: towards an agenda". In: *R&d Management* 36.3, pp. 223–228.

Giannopoulou, Eleni et al. (2010). "Implications of openness: A study into (all) the growing literature on open innovation". In: *Journal of Technology Management and Innovation* 5.3, pp. 162–180. ISSN: 07182724. DOI: 10.4067/S0718-27242010000300012. URL: http://www.jotmi.org.

Hall, Arthur David (1962). *A methodology for systems engineering*. Van Nostrand, 478p. ISBN: 0442030460. URL: https://findit.dtu.dk/en/catalog/2351747888.

Hidalgo, C. A. and R. Hausmann (2009). "The building blocks of economic complexity". In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10570–10575. ISSN: 0027-8424. DOI: 10.1073/pnas.0900943106. arXiv: 0909.3890. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0900943106.

IBM Analytics. URL: https://www.ibm.com/analytics/hadoop/big-data-analytics (visited on 05/24/2018).

Léon, Gonzalo (2007). "Cooperative Models for Information Technology Transfer in the Context of Open Innovation". In: *Organizational Dynamics of Technology-Based Innovation: Diversifying the Research Agenda*. Vol. 235. Boston, MA: Springer US, pp. 43–61. ISBN: 9780387728032. DOI: 10.1007/978-0-387-72804-9_5. URL: http://link.springer.com/10.1007/978-0-387-72804-9{_}5.

Lewontin, Richard (2000). *Triple helix*. URL: http://triplehelix.stanford.edu/triplehelix (visited on 05/24/2018).

Maaten, Laurens van der (2018). *t-SNE*. URL: https://lvdmaaten.github.io/tsne/.

Merriam-Webster Dictionary (1828). *Shift | Definition of Shift by Merriam-Webster*. URL: https://www.merriam-webster.com/dictionary/shift (visited on 05/24/2018).

Miles, R. F. (1973). "Systems concepts: Lectures on contemporary approaches to systems." In: ISSN: 0018-9472. DOI: 10.1109/TSMC.1974.4309372. URL: https://ntrs.nasa.gov/search.jsp?R=19730050770.

Montreal Metropolitain, Chambre du commerce du (2011). "A look at Canadian university-industry Collaboration". In: http://www.ccmm.ca/documents/activities$_pdf/autres/2010savoir_2011_en.pdf$.

Moran, J J (1994). *The qfd book - the team approach to solving problems and satisfying customers through quality function deployment - guinta,lr, praizler,nc*. Vol. 11. 3. Amacom, pp. 275–276. ISBN: 081445139X.

Morgan, Lorraine and Patrick Finnegan (2008). "Deciding on open innovation: An exploration of how firms create and capture value with open source software". In: *IFIP International Federation for Information Processing*. Vol. 287, pp. 229–246. ISBN: 9780387875026. DOI: 10.1007/978-0-387-87503-3_13. URL: http://link.springer.com/10.1007/978-0-387-87503-3{_}13.

Mulder, Karl (2006). *Sustainable Developement for Engineers*.

Nelson, Richard R. (1993). *National Innovation Systems: A comparative Analysis*, pp. 3–21. ISBN: 0195076168. DOI: 10.1016/0048-7333(96)00880-3.

OECD (2008). "Open Innovation in Global Networks". In: *OECD Policy Brief* November, pp. 9–13. DOI: 10.1787/9789264047693-en. URL: http://www.oecd-ilibrary.org/science-and-technology/open-innovation-in-global-networks{_}9789264047693-en.

— (2018). *OECD Database*. URL: https://www.oecd-ilibrary.org/science-and-technology/the-oecd-regpat-database_241437144144.

OpenRefine (2018). *OpenRefine*. URL: http://openrefine.org/.

Pénin, Julien and Jean Pierre Wack (2008). "Research tool patents and free-libre biotechnology: A suggested unified framework". In: *Research Policy* 37.10, pp. 1909–1921. ISSN: 00487333. DOI: 10.1016/j.respol.2008.07.012. URL: http://linkinghub.elsevier.com/retrieve/pii/S0048733308001601.

Perkmann, Markus and Kathryn Walsh (2007). "University-industry relationships and open innovation: Towards a research agenda". In: *International Journal of Management Reviews* 9.4, pp. 259–280. ISSN: 14608545. DOI: 10.1111/j.1468-2370.2007.00225.x. URL: http://doi.wiley.com/10.1111/j.1468-2370.2007.00225.x.

Piaszczyk, C (2011). "Model Based Systems Engineering with Department of Defense Architectural Framework". In: *Systems Engineering* 14.3, pp. 305–326. ISSN: 1098-1241, 1098-1241. DOI: 10.1002/sys. URL: http://ezproxy.lib.ucf.edu/login?url=http://search.proquest.com/docview/926281441?accountid=10003{\%}5Cnhttp://sfx.fcla.edu/ucf?url{_}ver=Z39.88-2004{\&}rft{_}val{_}fmt=info:ofi/fmt:kev:mtx:journal{\&}genre=article{\&}sid=ProQ:ProQ:civilengineering{\&}atitle=Model+Based+Syste.

Porter, M. E. (1998). "Clusters and the new economics of competition." In: *Harvard business review* 76.6, pp. 77–90. ISSN: 00178012. DOI: 10.1042/BJ20111451. arXiv: /dx.doi.org/10.1016/0048-7333(84)90018-0 [http:]. URL: http://www.ncbi.nlm.nih.gov/pubmed/10187248.

Romer, Paul M. (1990). "Endogenous Technological Change". In: *Journal of Political Economy* 98.5, Part 2, S71–S102. ISSN: 0022-3808. DOI: 10.1086/261725. URL: http://www.journals.uchicago.edu/doi/10.1086/261725.

Saladini, Fabrizio et al. (2016). "Guidelines for emergy evaluation of first, second and third generation biofuels". In: *Renewable and Sustainable Energy Reviews* 66, pp. 221–227.

Sheard, Sarah A and Ali Mostashari (2009). "Principles of complex systems for systems engineering". In: *Systems Engineering* 12.4, pp. 295–311.

Tsinghua (2018). *Collaborative Programs*. URL: http://www.tsinghua.edu.cn/publish/enven/6296/index.html.

Weck, Olivier de et al. (2011). *Engineering Systems: Meeting Human Needs in a Complex Technological World*, p. 224. ISBN: 978-0262016704.

World Bank (2017). *DataBank | The World Bank*. URL: http://databank.worldbank.org/data/home.aspx (visited on 05/24/2018).

List of Figures

List of Tables

Listings

{2.7182818284}ϑφερτυθιοπσδφγηξκλ