UNIVERSITY OF THESSALY

MASTER OF SCIENCE

Modeling and Analysis of Innovation with Artificial Intelligence

Author:
Athanasios ZOUMPEKAS

Supervisor:
Emmanouil VAVALIS

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Electrical & Computer Engineering

October 4, 2019

Declaration of Authorship

I, Athanasios ZOUMPEKAS, declare that this thesis titled, "Modeling and Analysis of Innovation
with Artificial Intelligence" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

Περίληψη

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Μεταπτυχιακό Δίπλωμα

Μοντελοποίηση και Ανάλυση Καινοτομίας με τη χρήση Τεχνητής Νοημοσύνης

Αθανάσιος Ζουμπέκας

Στους σημερινούς καιρούς ταχείας τεχνολογικής και οικονομικής αλλαγής, είναι ζωτικής σημασίας για μια χώρα να αξιολογήσει τα πλεονεκτήματα και τις αδυναμίες της όσον αφορά τις επιδόσεις της στην καινοτομία. Σκοπός αυτής της ερευνητικής μελέτης είναι να αξιολογήσει και να συγκρίνει την καινοτομία της Ελλάδας σε σχέση με την Ευρωπαϊκή Ένωση χρησιμοποιώντας τους δείκτες του Ευρωπαϊκού Πίνακα Αποτελεσμάτων Καινοτομίας. Συγκρίναμε τα αποτελέσματα της Ελλάδας με τις μέσες βαθμολογίες της ΕΕ κατά την περίοδο 2010-2017. Αναλύουμε τη συστηματική υπεραπόδοση και τη χαμηλή απόδοση της Ελλάδας και τις τάσεις των δεικτών αυτών με την πάροδο των ετών χρησιμοποιώντας στατιστικές τεχνικές και μεθόδους. Επιπλέον, χρησιμοποιούμε τεχνικές μηχανικής μάθησης για να καθορίσουμε και να παρουσιάσουμε τα πιο σημαντικά χαρακτηριστικά που οδηγούν τη διακύμανση της συνολικής βαθμολογίας καινοτομίας σε επίπεδο ΕΕ και Ελλάδας. Πιστεύουμε ότι αυτή η εργασία παρέχει εξηγήσεις και στοιχεία που βοηθούν τη χώρα να εκτιμήσει τα πλεονεκτήματά της και να αντιμετωπίσει τα μειονεκτήματα.

UNIVERSITY OF THESSALY

Abstract

Department of Electrical & Computer Engineering

Master of Science

Modeling and Analysis of Innovation with Artificial Intelligence

Athanasios ZOUMPEKAS

In the current times of rapid technological and economic change, it is crucial for a country to assess its strengths and weaknesses regarding its innovation performance. The purpose of this research study is to evaluate and compare the innovativeness of Greece relative to the European Union using the indicators from the European Innovation Scoreboard. We compare the scores of Greece with the EU average scores over the period 2010-2017. We analyze systematic overperformance and underperformance of Greece and the trends of these indicators over the years utilizing statistical techniques and methods. Furthermore, we use machine learning techniques to determine and display the most important features that drive the fluctuation of summary innovation score of EU and Greece level. It is our belief that this thesis provides explanations and evidence to help the country value its advantages and deal with the disadvantages.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors Prof. Elias Houstis and Prof. Emmanouil Vavalis for the continuous support of my MSc study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Michael Vassilakopoulos and Prof. Yeoryios Stamboulis for their encouragement, insightful comments, and hard questions.

Finally, I must express my very profound gratitude to my parents, to my brother and to my friends for providing me with unfailing support and continuous encouragement through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you all !

Author
Athanasios ZOUMPEKAS

Contents

List of Figures

List of Tables

List of Abbreviations

EU	**E**uropean **U**nion
GR	**G**reece
R&D	**R**esearch and **D**evelopment
SMEs	**S**mall and **M**edium-sized **E**nterprises
KETs	**K**ey **E**nabling **T**echnologies
EIS	**E**uropean **I**nnovation **S**coreboard
UK	**U**nited **K**ingdom
ICT	**I**nformation **C**ommunication **T**echnology
IPR	**I**ntellectual **P**roperty **R**ights
PCT	**P**atent **C**ooperation **T**reaty
GEM	**G**lobal **E**ntrepreneurship **M**onitor
TEA	**T**otal **E**ntrepreneurial **A**ctivity)
GDP	**G**ross **D**omestic **P**roduct
PPS	**P**urchasing **P**ower **S**tandard
EBOPS	**E**xtended **B**alance **O**f **P**ayments **S**ervices classification
GLS	**G**eneralized **L**east **S**quares
OLS	**O**rdinary **L**east **S**quares
AR	**A**uto**R**egressive
GLSAR	**G**eneralized **L**east-**S**quares regression with **A**uto**R**egressive errors
CD	**C**oordinate **D**escent
SVM	**S**upport **V**ector **M**achine
MMH	**M**aximal **M**argin **H**yperplane
MMC	**M**aximal **M**argin **C**lassifier

Dedicated to my beloved family & friends . . .

Chapter 1

Introduction

1.1 Innovation

Innovation is one of the two fundamental functions of an organization [1]. It is the procedure of translating an idea or invention into a good or service that creates value or for which customers will pay. The idea must be able to be replicated at an economical cost and must satisfy a specific need. Innovation implicates intentional application of information, creativity and lead in deriving greater or different values from resources [2]. It involves all processes by which new ideas are generated and converted into useful products. It also includes the developing of new sources of supply with raw materials [3]. In business, innovation is the outcome of applied ideas by the company in order to further satisfy the requirements and expectations of the customers.

Innovation can be also defined as a process that provides added value and a degree of novelty to the organization, suppliers and customers, developing new procedures, solutions, products and services and new ways of marketing. It is the adoption of new or significantly improved elements to create added value to the organization directly or indirectly for its customers [4]. In a social context, innovation aids in the development of new methods for alliance creation, joint venturing, flexible work hours, and creation of buyers' purchasing power. It is synonymous with risk-taking. Organizations and companies that develop new revolutionary products and services take on the great risk because they create new markets.

1.1.1 The Importance of Innovation

For the rest of this these we concentrate firstly on the European Union and secondly on Greece. Innovation is one of the most important concerns of each organization. Its role in the development and coordination of the market is intrinsic. The importance of innovative applications is crucial in all human areas from product development, methods of management, ways of doing works and beyond [5].

On the other hand, industry is crucial for competitiveness and innovation is a key factor in this regard [6]. Industry commonly accounts for around 80% of a country's exports. Some 65% of private sector research and development (R&D) investment comes from manufacturing. Thus, industrial modernization in every country must be broad-reaching and include:

- the successful commercialization of product and service innovations,
- the industrial exploitation of innovative manufacturing technologies and
- innovative business models.

Organizations who prioritize innovation are likewise the individuals who experience the most astounding increment in turnover. Some 79% of companies that introduced at least one innovation since 2011 experienced an increase of their turnover by more than 25% by 2014 [7].

Small and medium-sized enterprises (SMEs) are specific targets for innovation policy. The smaller the company is, the more it faces constraints to innovation or to the commercialization of its innovations. Some 63% of companies with between 1 and 9 employees declared having introduced at least one innovation since 2011, compared to 85% of companies with 500 employees or more. Some 71% of companies with between 1 and 9 employees encountered difficulties commercializing their innovations due to a lack of financial resources, compared to 48% of companies with 500 employees or more [8].

1.1.2 Monitoring Innovation

The European Commission provides various tools that map, monitor and assess the EU's performance in different innovation areas. The information provided helps policy makers and practitioners at EU, national and regional levels to benchmark their performance and policies and to learn about new trends and emerging business opportunities that can inform evidence-based policy making [9]. The list of the Commission' s tools include the following [9].

- European Innovation Scoreboard,
- Regional Innovation Scoreboard,
- European Public Sector Innovation Scoreboard,
- Innobarometer,
- Regional Innovation Monitor Plus,
- Business Innovation Observatory,
- Digital Entrepreneurship Monitor,
- European Cluster Observatory,
- Key Enabling Technologies (KETs) Observatory and
- KETs Technology Infrastructure Mapping

In this thesis, we are analyzing the country level innovation performance. Thus, we are using the data from the European Innovation Scoreboard, version 2018. Following we briefly describe the tools listed above, highlighting the one we utilize.

European Innovation Scoreboard

The European Innovation Scoreboard (EIS) provides a comparative assessment of research and innovation performance in Europe. It assesses the relative strengths and weaknesses of national research and innovation systems, and helps countries and regions identify the areas they ought to address [9] for their further development.

The 2018 edition of the scoreboard highlights that the EU's innovation performance keeps on improving. The progress of innovation performance is accelerating,

and that the outlook is positive. Since 2010, the EU's average innovation performance has increased by 5.8 percentage, and it is expected to improve by an extra 6 percentage over the next 2 years.

According to EIS, Sweden remains the EU innovation leader, followed by Denmark, Finland, the Netherlands, the UK, and the Luxembourg. Moreover the fastest growing innovators are Lithuania, the Netherlands, Malta, the UK, Latvia, and France [9].

Other European Commission Tools

Besides EIS, the EU provides various tools to monitor innovation performance as listed above. In regional level there are two tools available the Regional Innovation Scoreboard, which is a regional extension of the EIS, assessing the innovation performance of European regions on a limited number of indicators and the Regional Innovation Monitor Plus, which provides a platform for sharing knowledge and know-how on major innovation and industrial policy trends in the EU regions.

In addition, there are more specific tools to sectors of a country such as the European Public Sector Innovation Scoreboard, Innobarometer, the Business Innovation Observatory, the Digital Entrepreneurship Monitor, the European Cluster Observatory, the Key Enabling Technologies (KETs) Observatory, and the KETs Technology Infrastructure Mapping. European Public Sector Innovation Scoreboard is a tool developed from the EU to improve the ability to benchmark the innovation performance of the public sector in whole Europe. Industry, businesses and entrepreneurship surely need more targeted tools such as the Innobarometer, which is a survey on activities and attitudes from the general public and European businesses related to innovation, the Business Innovation Observatory, which provides evidence and evaluation on the latest innovative trends in business and industry and the Digital Entrepreneurship Monitor, which displays a comparative assessment of the enabling factors that create a fertile ground for digital entrepreneurs to flourish and operate successfully.

Last but not least, the European Cluster Observatory is a tool providing statistical information, analysis and mapping of clusters in Europe. The Key Enabling Technologies (KETs) Observatory provides the EU, national policymakers and business stakeholders with information on the performance of the EU Member States and competing economies regarding the deployment of KETs. KETs Technology Infrastructure Mapping allows SMEs and other stakeholders to identify technological service centers active in the field of KETs.

1.2 Purpose of this Thesis

The specific purpose of this thesis is to compare innovativeness of Greece versus EU using the indicators provided by the European Innovation Scoreboard. We compare the scores of Greece with the EU average scores over the period 2010-2017. We briefly analyze systematic over-performance and under-performance of Greece and the trends of these indicators over the years. Furthermore, we use machine learning and statistical techniques to determine the most important features that drive the fluctuation of the summary innovation score at EU and Greek level.

1.3 Thesis Organization

This research thesis is organized as follows. In chapter Background Material we present the background information on the topic of modeling and analysis of innovation. We explain in depth the composite indicators and each indicator respectively. Moreover we present and provide further information on statistical and machine learning techniques and algorithms utilized in this study and other related studies. In section 2.4 we exhibit selected research publications and other related projects on this topic of interest.

Chapter Methodology depicts our methodology and processes of extracting conclusions of data. We present our data collection, the data pre-processing methods and the analytics work-flow. In addition, we comment the software tools we utilize for the analysis of the data. In chapter Data Analysis we analyze the aforementioned data, i.e. the composite indicators and the simple indicators of innovation. The numerous charts show the fluctuation and the percentage change of the indicators related to Greece and EU from year 2010 to 2017.

By using statistics and machine learning methods, we show and evaluate the trends and the importance of each indicator in chapter Predictive Analytics. We analyze and compare the trends of indicators to show systematic over- or underperformance of Greece versus EU. Furthermore, we utilize a modeling technique on data to be able to evaluate the importance of each indicator related to Greece and EU respectively. Chapter Conclusion summarizes our observations, draws conclusions. and suggest further related research directions.

Chapter 2

Background Material

2.1 European Innovation Scoreboard

In this section, we present and comment on the indicators used in our analysis. They consist the basis of the annual European Innovation Scoreboard (EIS) which provides a comparative evaluation of the research and innovation performance of the EU Member States. It specifically offers comparisons on relative strengths and weaknesses regarding the research and innovation systems in the country level. It is essentially provides assistance to Member States in order to assess areas in which they need to focus their efforts to boost their innovation performance.

The main indicator is the **Summary Innovation Index** that summarizes the range of different indicators of innovation and measures the total innovation performance.

We next report on the definitions, the explanations and the methods of calculation for composite indicators and indicators, as these are provided by the EIS 2018 Methodology report, in the following subsections. For more detailed information the reader is referred to [10].

2.1.1 Composite Indicators

The EIS 2018 discriminates between four main types of indicators and ten innovation dimensions, capturing in total 27 different indicators [10]. These four main categories and their composite indicators are:

Framework conditions: main drivers of innovation performance external to the firm. Composite indicators:

- **Human resources** includes three indicators and calculates the availability of a high-skilled and educated workforce. Human resources captures New doctorate graduates, Population aged 25-34 with completed tertiary education, and Population aged 25-64 involved in education and training.
- **Research systems** includes three indicators and gauges the international competitiveness of the science base by concentrating on International scientific co-publications, Most cited publications, and Foreign doctorate students.
- **Innovation-friendly environment** captures the environment in which enterprises operate and includes two indicators. The Broadband penetration among enterprises and Opportunity-driven entrepreneurship indicators measure the degree to which individuals seek entrepreneurial activities as they look at new opportunities, for instance resulting from innovation.

Investments: investments made in both the public and business sector. Composite indicators:

Finance and support includes two indicators and measures the availability of finance for innovation projects by Venture capital expenditures, and the support of governments for research and innovation activities by R&D expenditures in universities and government research organizations.

Firm investments includes three indicators of both R&D and non-R&D investments that firms make to generate innovations, and the efforts enterprises make to upgrade the ICT skills of their personnel.

Innovation activities: different aspects of innovation in the business sector. Composite indicators:

Innovators includes three indicators measuring the share of firms that have introduced innovations onto the market or within their organizations, covering both product and process innovators, marketing and organizational innovators, and SMEs that innovate in-house.

Linkages includes three indicators measuring innovation capabilities by looking at collaboration efforts between innovating firms, research collaboration between the private and public sector, and the extent to which the private sector finances public R&D activities.

Intellectual assets captures different forms of Intellectual Property Rights (IPR) generated in the innovation process, including PCT patent applications, Trademark applications, and Design applications.

Impacts: the effects of firms' innovation activities. Composite indicators:

Employment impacts measures the impact of innovation on employment and includes two indicators measuring Employment in knowledge-intensive activities and Employment in fast-growing firms in innovative sectors.

Sales impacts measures the economic impact of innovation and includes three indicators measuring exports of medium and high-tech products, Exports of knowledge-intensive services, and Sales due to innovation activities.

2.1.2 Indicators

We next provide brief descriptions of the basic indicators.

New doctorate graduates per 1000 population aged 25-34 This indicator is a fraction which has as numerator the number of doctorate graduates and denominator the population between and including 25 and 34 years. It is a measure of the supply of new second-stage tertiary graduates in all fields of training. For most countries, it captures PhD graduates.

Percentage population aged 25-34 having completed tertiary education This indicator is a fraction which has as numerator the number of persons in age class with some form of post-secondary education and denominator the population between and including 25 and 34 years. This is a general indicator of the supply of advanced skills. It is not limited to science and technical fields, because the adoption of innovations in many areas, in particular in the service sectors, depends on a wide range of skills. The indicator focuses on a relatively young age cohort of the population, aged 25 to 34, and will therefore easily and quickly reflect changes in educational policies leading to more tertiary graduates.

Percentage population aged 25-64 involved in lifelong learning This indicator is a fraction which has as numerator the target population for lifelong learning statistics referring to all persons in private households aged between 25 and 64 years and denominator the total population of the same age group, excluding those who did not answer the question concerning participation in (formal and non-formal) education and training. The information collected relates to all education or training, whether or not relevant to the respondent's current or possible future job. Data are collected through the EU Labour Force Survey. The reference period for the participation in education and training is the four weeks preceding the interview, as is usual in the Labour Force Survey. Lifelong learning encompasses all purposeful learning activity, whether formal, non-formal or informal, undertaken on an ongoing basis with the aim of improving knowledge, skills and competence. The intention or aim to learn is the critical point that distinguishes these activities from non-learning activities, such as cultural or sporting activities.

International scientific co-publications per million population This indicator is a fraction which has as numerator the number of scientific publications with at least one co-author based abroad (where abroad is non-EU for the EU28) and denominator the total population. International scientific co-publications are a proxy for the quality of scientific research as collaboration increases scientific productivity.

Scientific publications among the top 10% most cited publications worldwide as % of

total scientific publications of the country This indicator has the number of scientific publications among the top-10% most cited publications worldwide and total number of scientific publications as numerator and denominator respectively. The indicator is a measure for the efficiency of the research system, as highly cited publications are assumed to be of higher quality. There could be a bias towards small or English-speaking countries given the coverage of Scopus' publication data.

Foreign doctorate students as a % of all doctorate students This indicator has the number of doctorate students from foreign countries and total number of doctorate students as numerator and denominator respectively. The share of foreign doctorate students reflects the mobility of students as an effective way of diffusing knowledge. Attracting high-skilled foreign doctorate students will secure a continuous supply of researchers.

Broadband penetration This indicator has the number of enterprises with a maximum contracted download speed of the fastest fixed internet connection of at least 100 Mb/s and total number of enterprises as numerator and denominator respectively. Realizing Europe's full e-potential depends on creating the conditions for electronic commerce and the Internet to flourish. This indicator captures the relative use of this e-potential by the share of enterprises that have access to fast broadband.

Opportunity-driven entrepreneurship (Motivational index) This index is calculated as the ratio between the share of persons involved in improvement-driven entrepreneurship and the share of persons involved in necessity-driven entrepreneurship. Data from Global Entrepreneurship Monitor (GEM) distinguish between two types of entrepreneurship:

1. opportunity-driven entrepreneurship and
2. necessity-driven entrepreneurship.

The first includes persons involved in TEA (Total Early-Stage Entrepreneurial Activity) who claim to be driven by opportunity as opposed to finding no other option for work and who indicate the main driver for being involved in this opportunity is being independent or increasing their income, rather than just maintaining their income. The second includes persons involved in TEA who are involved in entrepreneurship because they had no other option for work. GEM has constructed the Motivational index to measure the relative degree of improvement-driven entrepreneurship.

R&D expenditure in the public sector (% of GDP) This indicator is a fraction which has as numerator all R&D expenditures in the government sector and the higher education sector and denominator the Gross Domestic Product (GDP). R&D expenditure represents one of the major drivers of economic growth in a knowledge-based economy. As such, trends in the R&D expenditure indicator provide key indications of the future competitiveness and wealth of the EU. Research and development spending is essential for making the transition to a knowledge-based economy as well as for improving production technologies and stimulating growth.

Venture capital (% of GDP) This indicator is a fraction which has as numerator the venture capital expenditures and denominator the GDP. Venture capital expenditures is defined as private equity being raised for investment in companies. Management buyouts, management buy-ins, and venture purchase of quoted shares are excluded. Venture capital includes early-stage (seed plus start-up) and expansion and replacement capital. The amount of venture capital is a proxy for the relative dynamism of new business creation. In particular for enterprises using or developing new (risky) technologies, venture capital is often the only available means of financing their (expanding) business.

R&D expenditure in the business sector (% of GDP) This indicator has all R&D expenditures in the business sector and GDP as nominator and denominator respectively. It captures the formal creation of new knowledge within firms. It is particularly important in the science-based sectors (pharmaceuticals, chemicals and some areas of electronics) where most new knowledge is created in or near R&D laboratories.

Non-R&D innovation expenditures (% of turnover) This indicator is a fraction of the sum of total innovation expenditure for enterprises, excluding intramural and extramural R&D expenditures divided by total turnover for all enterprises. It measures non-R&D innovation expenditure as a percentage of total turnover. Several of the components of innovation expenditure, such as investment in equipment and machinery and the acquisition of patents and licenses, measure the diffusion of new production technology and ideas.

Enterprises providing training to develop or upgrade ICT skills of their personnel This indicator is a fraction of the number of enterprises that provided any type of training to develop Information Communication Technology (ICT) related skills of their personnel divided by the total number of enterprises. ICT skills are particularly important for innovation in an increasingly digital economy.

The share of enterprises providing training in that respect is a proxy for the overall skills development of employees.

SMEs introducing product or process innovations as % of SMEs This indicator is a fraction of the number of Small and medium-sized enterprises (SMEs) who introduced at least one product innovation or process innovation either new to the enterprise or new to their market and the total number of SMEs. A product innovation is the market introduction of a new or significantly improved good or service with respect to its capabilities, user friendliness, components or sub-systems. A process innovation is the implementation of a new or significantly improved production process, distribution method, or supporting activity. Technological innovation, as measured by the introduction of new products (goods or services) and processes, is a key ingredient to innovation in manufacturing activities. Higher shares of technological innovators should reflect a higher level of innovation activities.

SMEs introducing marketing or organisational innovations as % of SMEs This is the number of SMEs who introduced at least one new organizational innovation or marketing innovation divided by the total number of SMEs. An organizational innovation is a new organizational method in an enterprise's business practices (including knowledge management), workplace organization or external relations that has not been previously used by the enterprise. A marketing innovation is the implementation of a new marketing concept or strategy that differs significantly from an enterprise's existing marketing methods and which has not been used before. Many firms, in particular in the services sectors, innovate through other non-technological forms of innovation. Examples of these are marketing and organizational innovations. This indicator captures the extent to which SMEs innovate through non-technological innovation.

SMEs innovating in-house as % of SMEs This is the number of SMEs with in-house innovation activities divided by the total number of SMEs. In-house innovating enterprises are defined as enterprises which have introduced product or process innovations either themselves or in co-operation with other enterprises or organisations. This indicator measures the degree to which SMEs, that have introduced any new or significantly improved products or production processes, have innovated in-house. It is limited to SMEs, because almost all large firms innovate and because countries with an industrial structure weighted towards larger firms tend to do better.

Innovative SMEs collaborating with others (% of SMEs) This indicator shows the number of SMEs with innovation co-operation activities divided by the total number of SMEs. The aforementioned SMEs had any co-operation agreements on innovation activities with other enterprises or institutions in the three years of the survey period. This indicator measures the degree to which SMEs are involved in innovation co-operation. Complex innovations, in particular in ICT, often depend on the ability to draw on diverse sources of information and knowledge, or to collaborate in the development of an innovation. This indicator measures the flow of knowledge between public research institutions and firms, and between firms and other firms. The indicator is limited to SMEs, because almost all large firms are involved in innovation co-operation.

Public-private co-publications per million population This is the number of public-private co-authored research publications divided by total population. The

definition of the "private sector" excludes the private medical and health sector. Publications are assigned to the country/countries in which the business companies or other private sector organizations are located. This indicator captures public-private research linkages and active collaboration activities between business sector researchers and public sector researchers resulting in academic publications.

Private co-funding of public R&D expenditures (percentage of GDP) This is all R&D expenditures in the government sector and the higher education sector financed by the business sector divided by Gross Domestic Product (GDP). This indicator measures public-private co-operation. University and government R&D financed by the business sector are expected to explicitly serve the more short-term research needs of the business sector.

PCT patent applications per billion GDP (in PPS) This indicator is a fraction which has as numerator the number of patent applications filed under the Patent Cooperation Treaty (PCT), at international phase, designating the European Patent Office (EPO) and denominator the GDP in Purchasing Power Standard. Patent counts are based on the priority date, the inventor's country of residence and fractional counts. The capacity of firms to develop new products will determine their competitive advantage. One measure of the rate of new product innovation is the number of patents. This indicator measures the number of PCT patent applications.

Trademark applications per billion GDP (in PPS) This is the number of trademark applications applied for at European Union Intellectual Property Office plus number of trademark applications applied for at World Intellectual Property Office ("yearly Madrid applications by origin") divided by GDP in in Purchasing Power Standard. Trademarks are an important innovation indicator, especially for the service sector. The Community trademark gives its proprietor a uniform right applicable in all Member States of the European Union through a single procedure which simplifies trademark policies at European level. It fulfills the three essential functions of a trademark: it identifies the origin of goods and services, guarantees consistent quality through evidence of the company's commitment vis-à-vis the consumer, and it is a form of communication, a basis for publicity and advertising.

Design applications per billion GDP (in PPS) This is an indicator which shows the number of individual designs applied for at European Union Intellectual Property Office divided by GDP in Product in Purchasing Power Standard. A design is the outward appearance of a product or part of it resulting from the lines, contours, colours, shape, texture, materials and/or its ornamentation. A product can be any industrial or handicraft item including packaging, graphic symbols and typographic typefaces but excluding computer programs. It also includes products that are composed of multiple components, which may be disassembled and reassembled. Community design protection is directly enforceable in each Member State and it provides both the option of an unregistered and a registered Community design right for one area encompassing all Member States.

Employment in knowledge-intensive activities (% of total employment) This is the number of employed persons in knowledge-intensive activities in business industries divided by the total employment. Knowledge-intensive activities are

defined, based on EU Labour Force Survey data, as all NACE Rev.2 industries at 2-digit level where at least 33% of employment has a tertiary education degree. These activities provide services directly to consumers, such as telecommunications, and provide inputs to the innovative activities of other firms in all sectors of the economy.

Employment in fast-growing enterprises (% of total employment) This is a fraction which has as numerator the number of employees in high-growth enterprises in 50% 'most innovative' industries, including numerous NACE industries, and denominator the total employment for enterprises with 10 or more employees. This indicator provides an indication of the dynamism of fast-growing firms in innovative sectors as compared to all fast-growing business activities. It captures the capacity of a country to rapidly transform its economy to respond to new needs and to take advantage of emerging demand.

Exports of medium and high technology products as a share of total product exports This is the value of medium and high tech exports, in national currency and current prices, including exports, divided by the value of total product exports. This indicator measures the technological competitiveness of the EU, i.e. the ability to commercialize the results of research and development (R&D) and innovation in international markets. It also reflects product specialization by country. Creating, exploiting and commercializing new technologies are vital for the competitiveness of a country in the modern economy. Medium and high technology products are key drivers for economic growth, productivity and welfare, and are generally a source of high value added and well-paid employment.

Knowledge-intensive services exports as % of total services exports This indicator is a fraction which has as numerator the exports of knowledge-intensive services and denominator the total value of services exports. The aforementioned exports are defined as the sum of credits in Extended Balance of Payments Services Classification (EBOPS) 2010 items. The indicator measures the competitiveness of the knowledge-intensive services sector. Competitiveness-enhancing measures and innovation strategies can be mutually reinforcing for the growth of employment, export shares, and turnover at the firm level. It reflects the ability of an economy, notably resulting from innovation, to export services with high levels of value added, and successfully take part in knowledge-intensive global value chains.

Sales of new-to-market and new-to-firm innovations as % of turnover This is the sum of total turnover of new or significantly improved products, either new-to-the-firm or new-to-the-market, for all enterprises divided by total turnover for all enterprises. This indicator measures the turnover of new or significantly improved products and includes both products which are only new to the firm and products which are also new to the market. The indicator thus captures both the creation of state-of-the-art technologies (new to-market products) and the diffusion of these technologies (new-to-firm products).

2.2 Statistics

Statistics is a branch of mathematics operating with information collection. It is used to inform scientific decision-making in the absence of complete information about

phenomena of interest. Data is collected and then analyzed to draw conclusions in a statistical way in order to interpret and present them [11], [12]. In general in order to apply statistics to a problem, one should start with a statistical population or a model to be analyzed. Populations can be different groups of people or objects. Statistics deals with every aspect of data, involving the organization of data acquisition in terms of the design of surveys and experiments [11]. Representative sampling guarantees that inferences and conclusions can be reasonably extended from the sample to the population.

In data analysis there are two principal statistical methods, descriptive and inferential statistics. Descriptive statistics summarize data from a sample utilizing criteria such as the mean or standard deviation. Inferential statistics draw conclusions from data that are subject to random variation (e.g., observational errors, sampling variation).

2.2.1 Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. More formally, they are used to present quantitative descriptions in a manageable form. In a research project we may have lots of measures and a large number of people or objects on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of every quantitative analysis of data. With descriptive statistics we are simply describing what is or what the data shows [13].

2.2.2 Inferential Statistics

Inferential statistics are distinguished from descriptive statistics. With inferential statistics, the goal is to reach conclusions that extend beyond the immediate data alone. For example, we use inferential statistics to try to infer from the sample data what the population might think. In addition, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study [13]. Thus, we use this type of statistics to make inferences from our data to more general conditions.

The Student's t-test is one type of inferential statistics. It is used to determine whether there is a significant difference between the means of two groups. In this thesis we utilize a variation of the normal t-test, namely two sample t-test assuming unequal variances. The mathematical formulation of this t-test is given next.

Let $\bar{x}$ and $\bar{y}$ be the sample means and s_x and s_y be the sample standard deviations of two sets of data of size n_x and n_y respectively. Also, μ_x and μ_y denote the population means. If x and y are normal, or n_x and n_y are sufficiently large for the Central Limit Theorem to hold, then the random variable

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{2.1}$$

has distribution $T(m)$ where

$$m = \frac{(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y})^2}{\frac{(\frac{s_x^2}{n_x})^2}{n_x - 1} + \frac{(\frac{s_y^2}{n_y})^2}{n_y - 1}}. \tag{2.2}$$

This t-test can be used to test the difference between sample means even when the population variances are unknown and unequal. The resulting test, called Welch's t-test [14], will have a lower number of degrees of freedom than $(n_x - 1) + (n_y - 1)$, which was sufficient for the case where the variances were equal.

2.2.3 Hypothesis Testing

Hypothesis testing is a use of inferential statistics to determine the probability that a given hypothesis is true. The common process of hypothesis testing contains four steps [15], [16].

1. State the null hypothesis H_0 and the alternative hypothesis H_a. Regularly the H_0 means that the observations are the result of pure chance. The H_a states that the observations show a real effect combined with a component of chance variation.

2. Identify a test statistic that can be used to assess the truth of the null hypothesis.

3. Compute the p_{value}, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true. The smaller the p_{value}, the stronger the evidence against the null hypothesis.

4. Compare the p_{value} to an acceptable significance value alpha (α). If $p_{value} \leq \alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

Commonly an alpha value of $\alpha = 0.05$ is used, which mean 95% statistical significance level [17].

2.2.4 Correlation Analysis

Correlation analysis is a technique of statistical evaluation. It is used to consider the strength of a relationship between two, numerically measured, continuous variables [18]. This specific kind of analysis is helpful when a researcher wants to establish if there are conceivable associations between variables. Correlation analysis does not decides circumstances and end results. However, this is not the situation on the grounds that different factors that are absent in the exploration may have affected on the outcomes.

If correlation is found between two variables it implies that when there is an efficient change in one variable, there is additionally an orderly change in the other. Thus, the variables alter together over a certain period of time [19]. If there is correlation found, depending upon the numerical values measured, this can be either positive or negative. Positive correlation exists if one variable increases simultaneously with the other, i.e. the high numerical values of one variable relate to the high numerical values of the other. Negative correlation exists if one variable decreases

when the other increases, i.e. the high numerical values of one variable relate to the low numerical values of the other.

Pearson's product-moment coefficient is the measurement of correlation [20]. It ranges, (depending on the correlation), between $+1$ and -1.

- $+1$ indicates the strongest positive correlation possible.
- -1 indicates the strongest negative correlation possible.

In this manner the nearer the coefficient to both of these numbers the stronger the correlation of the data it represents. On this scale 0 indicates no correlation, hence values closer to zero highlight weaker/poorer correlation than those closer to $+1/-1$ [19]. Coefficients close to 1 or -1 mean that the series' are strongly correlated or inversely correlated respectively, and coefficients close to zero mean that the values are not correlated, and fluctuate independently of each other.

In machine learning the correlation analysis is fundamental in order to avoid multicollinearity issues. Multicollinearity is the occurrence of high correlations among independent variables in a multiple regression model. Multicollinearity can lead to skewed or misleading results when a researcher or analyst attempts to determine how well each independent variable can be used most effectively to predict or understand the dependent variable in a model. In statistical modelling, multicollinearity can lead to wider confidence intervals and less reliable probability values (p_{value}) for the independent variables.

2.2.5 Trend Analysis

Trend analysis is a well-known method of collecting information and attempting to detect patterns. In stock trading a trend analysis is a method of analysis that allows traders to predict what will happen with a stock in the future. It is based on historical data about the stock's performance given the overall trends of the market and particular indicators within the market. In general it is often used to predict future events.

In statistics, trend analysis often refers to methods and techniques for extracting an underlying pattern of behavior in a time series which would otherwise be partly or nearly completely hidden by noise. If the trend can be assumed to be linear, trend analysis can be undertaken within a formal regression analysis. If we do not assume the linear trend, then estimation can be done by non-parametric methods, e.g. Mann-Kendall test, which is a version of Kendall rank correlation coefficient [21].

Linear trend estimation is a statistical technique to support the interpretation of data. In time series, trend estimation can be used to make and justify statements about tendencies in the data, by relating the measurements to the times at which they occurred. Linear trend analysis expresses data as a linear function of time, and can be utilized to decide the significance of differences in dataset. Especially, it may be possible to determine if observations exhibit an increasing or decreasing trend which is statistically distinguished from random behaviour.

Fitting a trend line can be commonly done by least-squares method. We use generalized least-squares method, which is a variation of the least squares. The generalized least squares (GLS) estimator of the coefficients of a linear regression is a generalization of the ordinary least squares (OLS) estimator [22]. It is used to deal with situations in which the OLS estimator is not the best linear unbiased estimator because one of the main assumptions of the Gauss-Markov theorem, namely that of homoskedasticity and absence of serial correlation, is violated [23]. In such

situations, provided that the other assumptions of the Gauss-Markov theorem are satisfied, the GLS estimator is the best linear unbiased estimator.

In standard linear regression models we observe data $\{y_i, x_{ij}\}_{i=1,\dots,n,j=2,\dots,k}$. The response values form a vector $y = (y_1, \dots, y_n)^T$ and the predictor or feature values form the design matrix $X = (x_1^T, \dots, x_n^T)^T$, where $x_i = (1, x_{2i}, \dots, x_{ki})$ denotes a vector of the k predictor variables including a constant for the i_{th} unit. Below there is the regression equation:

$$y = X\beta + \epsilon \tag{2.3}$$

The assumptions of the model are commonly three:

1. X has full rank
2. $E[\epsilon|X] = 0$
3. $Cov[\epsilon|X] = \Omega$

where Ω is a known non-singular covariance matrix Ω.

Here $\beta \in \mathbb{R}^k$ denotes the vector of regression coefficients that must be estimated from the data. Suppose b is a candidate estimate for β. Then the residual vector for b will be $y - Xb$. The generalized least squares method estimates β by minimizing the squared distance of the residual vector $\hat{\beta} = argmin_b (y - Xb)^T \Omega^{-1} (y - Xb)$. Since the objective is a quadratic form in b, the estimator has an explicit formula

$$\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y. \tag{2.4}$$

A time series regression model can be written as $y_t = X_t\beta + \epsilon_t$, where t denotes simply the time step. The errors (ϵ_t) may be correlated with each other. In other words, we have auto-correlation or a dependency between the errors. We may consider situations in which the error at one specific time is linearly related to the error at the previous time. That is, the errors themselves follow a simple linear regression model that can be written as $\epsilon_t = \rho\epsilon_{t-1} + \omega_t$ [24]. Here, $|\rho| < 1$ is called the autocorrelation parameter and the ω_t term is a new error term that follows the usual assumptions that we make about regression errors. Thus, this model says that the error at time t is predictable from a fraction of the error at time $t - 1$ plus some new perturbation ω_t. The model for the ϵ_t errors of the original Y versus X regression is an autoregressive model for the errors, specifically $AR(1)$ in this case.

One reason why the errors might have an autoregressive structure is that the Y and X variables at time t may be related to the Y and X measurements at time $t - 1$. These relationships are being absorbed into the error term of our multiple linear regression model that only relates Y and X measurements made at concurrent times. Notice that the autoregressive model for the errors is a violation of the assumption that we have independent errors and this creates theoretical difficulties for ordinary least squares estimates of the beta coefficients. There are several different methods for estimating the regression parameters of the Y versus X relationship when we have errors with an autoregressive structure.

In this study, we utilize generalized least-squares (GLS) regression method with autoregressive errors (AR), namely GLSAR.

2.3 Machine Learning Algorithms & Techniques

This section provides background information on machine learning methods and techniques to determine the most important features-indicators.

2.3.1 Feature Importance

The feature selection is a fundamental technique in machine learning. We often need to decide which of the features provided to us we may drop and which we ought to keep [25]. There are several feature selection methods including dimensionality reduction [26]. In general there are two main categories of feature importance methods, the model agnostic and the model based. Model agnostic feature selection techniques, such as forward feature selection, basically extract the most important features required for the optimal value of chosen key performance indicator. However, this approach have generally one drawback, its large time complexity. In order to circumvent that issue feature importance can directly be obtained from the model being trained. The aforementioned method is the model based approach. We utilize model based importance of features in our study.

2.3.2 Cross-validation

Cross-validation is a statistical method used to estimate the effectiveness of machine learning models. It is widely used in applied machine learning to compare and select a model for a given predictive modeling problem. It is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

We utilize cross-validation in occasions where there are limited data samples. It is a resampling based technique used to evaluate machine learning models. The procedure has a single parameter, namely k, that corresponds to the number of groups that a given data sample is to be split into. Thus, this method is called k-fold cross-validation [27].

It has low complexity and in general results in a less biased or less optimistic estimate of the model skill than other methods, such as the train-test split of the dataset. The method can be described by the following steps:

1. Perform random shuffling of the dataset.
2. Split the dataset into k groups.
3. For each individual group:
   (a) Take the group as a hold out or test dataset.
   (b) Take the remaining groups as a training dataset.
   (c) Fit a model on the training set and evaluate it on the test set.
   (d) Preserve the evaluation score.
4. Summarize the skill of the model using the average measurements of evaluation stage.

To sum up, each sample in the dataset is utilized in the hold out set one time and in the training set $k - 1$ times [27]. The results of a k-fold cross-validation method are often summarized with the mean of the model skill scores [28].

2.3.3 Logistic Regression

Logistic Regression is a binary classification algorithm. It aims to find the best hyperplane in k-dimensional space that separates the two classes, minimizing logistic loss [29]. The mathematical formulation of the logistic loss is the following:

$$L(y_i, \hat{y}_i) = -\frac{1}{n}\sum_{i=1}^{n} log(1 + e^{-y_i(w^T x_i + b)}) \tag{2.5}$$

where y_i and $\hat{y}_i$ denote the label of point i and the model prediction $w^T x_i + b$ respectively. The w_i is the weight vector, x_i the input vector and b the bias term of the equation.

In order to estimate the importance of each feature to the model output the k dimensional weight vector is used. Large absolute values of w_j signify higher importance of the jth feature in the prediction of class. The optimization algorithm minimizes loss by setting learning large weights for features more important in predicting a data point to belong to the positive class and similarly for negative class. We utilize the *liblinear* solver for the optimization problem. The solver uses a coordinate descent (CD) algorithm that solves the optimization problem by successively performing approximate minimization along coordinate directions or coordinate hyperplanes [30].

2.3.4 Decision Trees

Decision trees method is one of the most popular machine learning algorithms. This method is easily interpretable because decision trees can be easily visualized. In general, it breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. The leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

The traditional and core algorithm for building decision trees is called $ID3$. It uses a top-down, greedy search through the space of possible branches with no backtracking [31]. In order to construct the decision tree the entropy and the information gain are used. Entropy $H(S)$ is a measure of the amount of uncertainty in the dataset S. Information gain $IG(A)$ measures the difference in entropy from before to after the dataset S is split on an attribute A. In other words, it accounts to how much uncertainty in S was reduced after splitting set S on attribute A. The attribute with the smallest entropy is used to split the set S on each iteration. Meanwhile, the attribute with the largest information gain is used to split the set S on each iteration. Entropy and information gain have the following formal mathematical definitions:

$$\mathrm{H}(S) = \sum_{x \in X} -p(x)\log_2 p(x) \tag{2.6}$$

$$IG(S, A) = \mathrm{H}(S) - \sum_{t \in T} p(t)\mathrm{H}(t) = \mathrm{H}(S) - \mathrm{H}(S|A) \tag{2.7}$$

where S denotes the current dataset for which entropy is being calculated, X denotes the set of classes in S and $p(x)$ the proportion of the number of elements in class x to the number of elements in set S. The T denotes the subsets created from splitting set S by attribute A such that $S = \bigcup_{t \in T} t$.

Random Forest

Random forest is a variation of the decision tree model. Actually, it is an ensemble model using multiple decision trees as base learners. The base learners are high variance, low bias models. The variance of the overall model is reduced by aggregating the decisions taken by all base learners to predict the response variable. The idea is to ensure that each base learner learns a different aspect of data. This is achieved via both row and column sampling [32]. In a classification setting the aggregation is done by taking a majority vote.

At each node of a decision tree, the feature to be used for splitting the dataset is decided based on information gain criterion or the more computationally inexpensive Gini impurity reduction. The feature that maximizes information gain (or reduction in Gini impurity) is selected as the splitting feature. Data is then divided to its children according to the value of splitting feature. Data belonging to each category of splitting feature goes to a separate child. The mathematical formulation of Gini impurity is given below.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{2.8}$$

where D denotes the dataset, k the number of classes and p_i the probability of a point belonging to class i.

In this study, we use Gini importance, thus feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. Therefore, the higher the value is the more important is the feature. For each decision tree, a node importance is calculated using Gini Importance, assuming only two child nodes (binary tree) as follows

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \tag{2.9}$$

where ni_j is the importance of j^{th} node, w_j is the weighted number of samples reaching node j, C_j is the impurity value of node j, $left(j)$ is the child node from left split on node j and $right(j)$ the right one. The importance of each feature on a decision tree is then calculated as:

$$fi_j = \frac{\sum_{j:l} ni_j}{\sum_{k \in z} ni_k} \tag{2.10}$$

where fi_j is the importance of feature i, ni_j is the importance of node j, l is the number of splits of node j on feature i and z is the set of all nodes. The multiple fi then are normalized in the range [0,1]. This is done by dividing by the sum of all feature importance values. The final feature importance of the Random Forest classifier is the average value over all the trees.

Extra-Trees

Extra-Trees stands for Extremely Randomized Trees. This classifier is an ensemble learning method fundamentally based on decision trees. Extra-Trees classifier randomizes certain decisions and subsets of data to minimize over-learning from the data and overfitting. Extra-Trees is similar to Random Forest. This algorithm builds and fits multiple trees and splits nodes using random subsets of features. However,

the two key differences between Random Forest and Extra-Trees Classifiers are the following :

- Extra-Trees method does not bootstrap observations. It uses sampling without replacement.
- Extra-Trees nodes are split on random splits of a random subset of the features selected at every node, not the best splits.

From a statistical point of view, dropping the bootstrapping idea leads to an advantage in terms of bias, whereas the split-point randomization has often an excellent variance reduction effect. This method has yielded state-of-the-art results in several high-dimensional complex problems. From a functional point of view, the Extra-Tree method produces piece-wise multi-linear approximations, rather than the piece-wise constant ones of Random Forests [33].

The feature importance is computed as described above in Random Forest using Gini impurity.

2.3.5 Support Vector Machines

Support vector machine (SVM) is a powerful supervised learning model for prediction and classification. The fundamental idea of SVM is to map the training data into higher dimensional space using a nonlinear mapping function and then perform "linear" regression in higher dimensional space in order to separate the data [34]. A predetermined kernel function is used for data mapping. Data separation is done by finding the optimal hyperplane. This optimal hyperplane is called the Support Vector with the maximum margin from the separated classes [35].

SVMs construct linear separating hyperplanes in high-dimensional vector spaces. Data points are viewed as $(\overrightarrow{x}, y)$ tuples, $\overrightarrow{x} = (x_1, \ldots, x_p)$ where the x_j are the feature values and y is the classification. Optimal classification occurs when such hyperplanes provide maximal distance to the nearest training data points.

If we consider a real-valued p-dimensional feature space, known mathematically as $\mathbb{R}^p$, then our linear separating hyperplane is an affine $p-1$ dimensional space embedded within it. If we consider an element of our p-dimensional feature space, i.e. $\overrightarrow{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$, then we can mathematically define the related hyperplane by the following equation:

$$b_0 + \sum_{j=1}^{p} b_j x_j = 0 \tag{2.11}$$

this is nothing more than a multi-dimensional dot product, and as such can be written even more succinctly as:

$$\overrightarrow{b} \cdot \overrightarrow{x} + b_0 = 0 \tag{2.12}$$

A formulation of a mathematical separating property described below:

- $\overrightarrow{b} \cdot \overrightarrow{x} + b_0 > 0, \quad if \quad y_i = 1$
- $\overrightarrow{b} \cdot \overrightarrow{x} + b_0 < 0, \quad if \quad y_i = -1$

This basically states that if each training observation is above or below the separating hyperplane, according to the geometric equation which defines the plane, then its associated class label will be $+1$ or -1.

The concept of the maximal margin hyperplane (MMH) is straight forward. MMH is the separating hyperplane that is farthest from any training observations, and is thus optimal. We compute the perpendicular distance from each training observation $\overrightarrow{x}_i$ for a given separating hyperplane. The smallest perpendicular distance to a training observation from the hyperplane is known as the margin. The MMH is the separating hyperplane where the margin is the largest. This guarantees that it is the farthest minimum distance to a training observation.

The procedure for determining a maximal margin hyperplane for a maximal margin classifier (MMC) is as follows. Given n training observations $\overrightarrow{x}_1, \ldots, \overrightarrow{x}_n \in \mathbb{R}^p$ and n class labels $y_1, \ldots, y_n \in \{-1, 1\}$, the MMH is the solution to the following optimization procedure:

Maximize $M \in \mathbb{R}$, by varying $b_1, \ldots, b_p$ such that:

- $\sum_{j=1}^{p} b_j^2 = 1$
- $y_i(\overrightarrow{b} \cdot \overrightarrow{x} + b_0) \geq M, \forall i = 1, \ldots, n$

To sum up this is the maximal margin classifier (MMC).

The optimization procedure in support vector classifier differs from that described above for the MMC. We need to introduce new parameters, namely $n \in i$ values (known as the slack values) and a parameter C. We wish to maximize M, across $b_1, \ldots, b_p, \epsilon_1, \ldots, \epsilon_n$ such that:

- $\sum_{j=1}^{p} b_j^2 = 1$
- $y_i(\overrightarrow{b} \cdot \overrightarrow{x} + b_0) \geq M(1 - \epsilon_i), \forall i = 1, \ldots, n$
- $\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C$

where C denotes a non-negative tuning parameter. M still represents the margin and the slack variables ϵ_i allow the individual observations to be on the wrong side of the margin or hyperplane. Basically, the ϵ_i inform us of the location of i_th observation relative to the margin and hyperplane. If $\epsilon_i = 0$ then the x_i training observation is on the correct side of the margin. If $\epsilon_i > 0$ then the x_i is on the wrong side of the margin. Finally, if $\epsilon_i > 1$ then the x_i is on the wrong side of the hyperplane. On the other side, C is a parameter which "controls" how much the individual ϵ_i can be modified to violate the margin. The values of the aforementioned parameter control the trade-off between bias and variance of the support vector linear classifier model.

Support vector machine (SVM) classifier model is somehow an extension of support vector linear classifier that results from expansion of the feature space through the use of functions known as kernels. Calculating the solution to the optimization problem, the algorithm only needs to make use of inner products between the observations and not the observations themselves. Recall that an inner product is defined for two p-dimensional vectors u, v as $[\overrightarrow{u}, \overrightarrow{v}] = \sum_{j=1}^{p} u_j v_j$. Thus, the inner product for two observations is $[\overrightarrow{x_i}, \overrightarrow{x_k}] = \sum_{j=1}^{p} x_{ij} x_{kj}$. A linear support vector classifier for a particular observation $\overrightarrow{x_i}$ can be represented as a linear combination of inner products: $f(\overrightarrow{x}) = b_0 + \sum_{i=1}^{n} \alpha_i [\overrightarrow{x}, \overrightarrow{x_i}]$, where α_i denotes the coefficient for each training point.

Slack variables can be utilized, ϵ_i and ϵ_i^*, in order to accomplish an acceptable degree of miss classification error. So now, there seems to be a constrained minimum

optimization problem, as this addition has occurred.

$$minR(w, \epsilon_i^*) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} (\epsilon_i + \epsilon_i^*) \tag{2.13}$$

where w is the vector of weights of the model.

The objective of SVM is to minimize ϵ_i , ϵ_i^* and $\|w\|^2$. The above optimization with constraint can be changed over by methods for Lagrangian multipliers to a quadratic programming problem. Therefore, the form of the solution can be given by the following equation:

$$f(x) = \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{2.14}$$

In equation 2.11 the K is the kernel function and its values is an inner product of two vectors x_i and x_j in the feature space $\phi(x_i)$ and $\phi(x_j)$ and satisfies the Mercer's condition. A real valued function $K(x, y)$ satisfies Mercer's condition if $\int \int K(x, y) g(x) g(y) dxdy \geq 0$ for all square-integrable functions $g(x)$. A function $f(x)$ is square-integrable if $\int_{-\infty}^{+\infty} |f(x)|^2 dx$. Therefore, $K(x_i, x_j) = \phi(x_i)\phi(x_j)$.

In Table 2.1 there are some common kernels used with SVM.

Kernel	**Formula**
Polynomial	$K(x_i, x_j) = (x_i x_j + 1)^d$
Gaussian	$K(x, y) = exp(-\frac{\|x-y\|^2}{2\sigma^2})$
Gaussian radial basis function	$K(x_i, x_j) = exp(-\gamma \left\|x_i - x_j\right\|^2)$
Laplace radial basis function	$K(x, y) = exp(-\frac{\|x-y\|}{\sigma})$
Hyperbolic tangent	$K(x_i, x_j) = tanh(kx_i x_j + c)$
Anova radial basis	$K(x, y) = \sum_{k=1}^{n} exp(-\sigma(x^k - y^k)^2)^d$

TABLE 2.1: Common SVM kernels

SVMs have numerous advantages over classical classification approaches like artificial neural networks, decision trees and others. Good performance in high dimensional spaces can be considered as an advantage. Moreover, the support vectors depend on a little subset of the training data which gives SVM an awesome computational advantage.

We utilize support vector machines with linear kernel for our research. Thus in order to estimate the importance of each feature to the model output the k dimensional weight vector is used. Large absolute values of w_j signify higher importance of the jth feature in the prediction of class. The absolute size of the weight coefficients in relation to each other can then be used to determine feature importance from data.

2.4 Related Work

Innovation analysis is considered fundamental and one of the most important keys for economic growth of each country. Many researchers from different fields focus their efforts in studying various aspects of innovation. Since 1990, a lot of research work has been conducted using advanced tools and indicators examining innovation.

Small and medium sized enterprises (SMEs) are a principal part of the industry and economy in all modern countries. In 1990, despite the fact that only little was known about SME innovation activities, Hyvarinen Lisa reviewed the definitions of innovation technology and factors on the background of innovation activities of SMEs. In the aforementioned research work, various concepts approaching innovativeness of SMEs and their contribution to total innovation are explained [36].

Furthermore, there is a limited amount of scientific work regarding the Greek innovativeness and economic performance. Researchers used empirical analysis and showed the unfriendliness of the Greek private sector to invest in R&D and the low productivity of innovation [37]. Others investigate the impact of the indicator R&D activity on operational performance of SMEs extending the objective on the operational performance of SMEs in the small open Greek economy [38].

Additional studies focus on the significance and awareness of a set of established strategic influences of technological innovation in the context of European newly-industrialized countries. Research studies such as [39], provided evidence from interviews conducted on Greek manufacturing firms (mainly SMEs) measuring their innovation rate as well as key performance indicators. Using statistical analysis tools summarized and highlighted the most important indicators having the major importance influence of innovation. This study also indicates that the Greek institutional context had insufficient important influences of innovation and the highly innovative companies were the ones to overcome barriers such as the low supply of technology and other innovation obstacles.

Moreover, such studies developed also in regional level and provide evaluation of the numerous policy instruments used by regional governments in Europe to promote innovation activity in SMEs [40]. Scientists try to find patterns of innovation in regional innovation structures which are becoming increasingly diverse, complex and nonlinear. To address these issues, they use multi-output models [41].

There is a variation in methods utilized by researchers trying to forecast or analyze in depth innovation or specifically indicators of innovation. A wide variety of machine learning and deep learning algorithms is commonly used. Advanced machine learning methods such as ensemble decision trees are utilized in study [42]. They demonstrated the use of ensembles of decision trees to model the intrinsic nonlinear characteristics of the innovation process and apply their method for predicting innovation activity to chemical companies. In addition, other studies use nonlinear methods based on Artificial Intelligence, namely neural networks [41], [43]–[46]. In the aforementioned study [45], they model and forecast innovation performance using a neural network model with fuzzy rules and provide evidence from Taiwanese manufacturing industry. They also implement an adaptive neuro-fuzzy inference system to measure the innovation performance through technical information resources and innovation objectives. In [46], they develop an Artificial Neural Network classification method and prediction model that can assist companies especially SMEs in evaluating Advanced Manufacturing Technology implementation contributing to innovation.

In addition, the fact that decision makers have to group the object of their analysis into homogeneous classes is very common and that's why they use clustering algorithms, as presented in numerous papers [43], [46]–[48]. Others prefer more traditional techniques based on statistical analysis and equation modeling [49], [50].

We should clearly note that there is a variety of data sources utilized, including combined data sources or a single data source. Data sources are ranging from traditional methods like interviews [39] to well-formed databases obtained from Eurostat's official website [48], [49], [51], World Bank Database, SCImago Journal

[43]. Researchers also use data from companies providing business services, such as ICAP [37].

We want to highlight the research from Rotterdam School of Management regarding the innovativeness of the Netherlands compared to European Union (EU) countries [50]. They statistically compared the Netherlands versus EU in indicators of innovation using data from European Innovation Scoreboard database. Briefly, the methods they utilized are generalized least squares regression for trend estimation and statistics.

In this thesis we review several of those methods described above, implement the associated statistical techniques and machine learning models and apply them on indicators time series data. We assess systematic over-performance and underperformance of Greece relative to EU countries and compare the trends of Greece and EU regarding the composite and simple indicators of innovation using data from European Innovation Scoreboard database. Moreover, we apply machine learning models to determine in a model-based sense the most important features-indicators affecting the fluctuation of summary innovation index.

We would like to close this section by pointing out that [50] is the study most closely related to ours. They tackle the same problem of analyzing innovativeness in the country-level through a very interesting, innovative and effective methodology. We inspired from their analysis of the Netherlands relative to the EU regarding innovation performance. Besides, we consider a similar approach to analyze the case of Greece. Furthermore, we add scientific value by involving machine learning to estimate the most important indicators, combining more information on this study.

Chapter 3

Methodology

3.1 Data Collection and Pre-processing

We use data from the European Innovation Scoreboard (EIS) database version 2018, comprising Greece and the EU average for the period 2010-2017. The data is collected from European Innovation Scoreboard website free of charge. The data is of high quality with minor missing observations for Greece.

The EIS 2018 database comprises of many dimensions. We utilize composite indicators and each individual indicator for the years 2010-2017 in our study. These all indicators are explained in detail in Indicators above.

We filter the EIS 2018 database to select only indicators and composite indicators regarding the European Union and Greece. We clean the data of missing values and we drop the two indicators, whose values are missing for Greece. These indicators are "Foreign doctorate students as a % of all doctorate students" and "Employment in fast-growing enterprises (% of total employment)". For each indicator and composite indicator we construct time series data from 2010 to 2017 using the normalized scores provided by the database. In total we have time series data comprising of 25 indicators and 11 composite indicators (including the summary innovation index).

3.2 Software Tools

For our study we utilize Python programming language [52] and use the data science and machine learning platform, called Anaconda [53]. Python is a very powerful programming language used for many different applications. Anaconda is a free open source distribution of the Python and R programming languages for large-scale data processing, predictive analytics, and scientific computing, that aims to simplify package management and deployment. The main Python libraries used are Pandas, NumPy and SciPy, which are present in the Anaconda installation package. Pandas is an open source, library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language [54]. SciPy is an open source Python library used for scientific computing and technical computing [55]. NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays [56].

For visualization purposes, we use the state of the art visualization libraries, Matplotlib and Plotly for Python [57], [58]. In addition, Jupyter Notebook is used for our implementation, because it provides high productivity features [59].

3.3 Analytics Work-flow & Methodology

We separate the analytics process in three main stages.

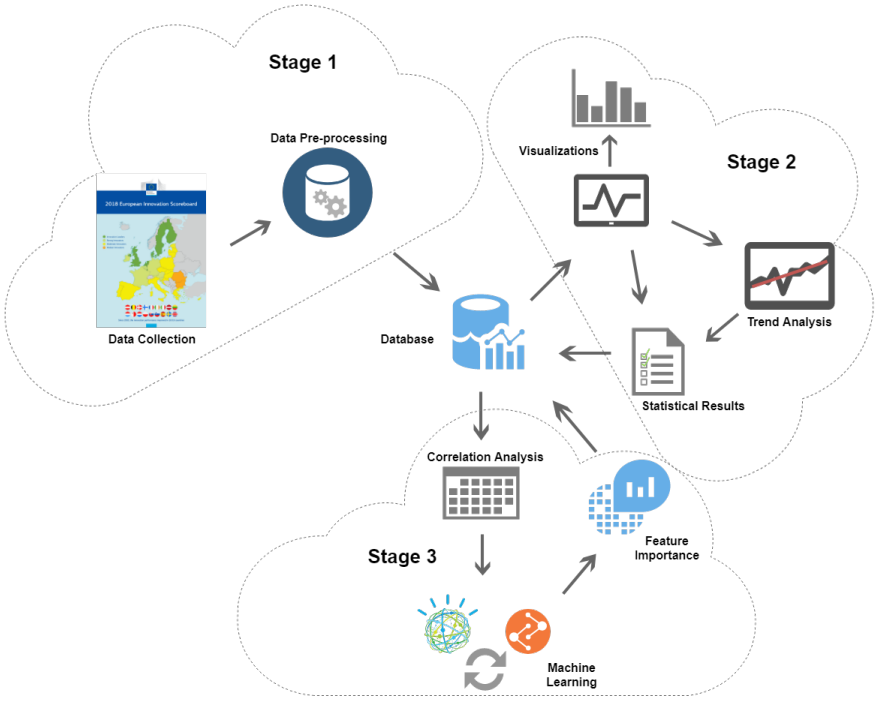

FIGURE 3.1: Analytics Work-flow & Methodology

First Stage - Pre-processing

The first stage deals with the data acquisition and pre-processing stage. We collect the data for years 2010 to 2017 and process them in order to build our dataset for our study, as described above in Data Collection and Pre-processing and **??** sections.

Second Stage - Statistical Analysis

The second stage consists of statistical analysis of the aforementioned time series data. We visualize each indicator and composite indicator of Greece versus the EU average over the time frame (2010-2017). The visualization process includes two graphs per indicator (or composite indicator), the actual values of time series and the percentage change each year. Then we explain the graphs and provide statistical test to measure the systematic out-performance and under-performance of Greece compared to EU. These comparisons are concluded from a two samples t-test assuming unequal variances between the (non-missing) data observations over the studied period. We state a hypothesis H_0 to test whether Greece outperforms on average the EU. For the aforementioned test, we accept or reject the H_0 according to the t-statistic (t), and p_{value}. We report the t-statistic (t), p_{value} and our decision based on an alpha level $\alpha = 0.05$, i.e. 95% statistical significance. The H_0 is rejected at $\frac{p_{value}}{2} \leq 0.05$, because it is a one-tailed test.

Below there is a simple if-then-else statement, we use for our hypothesis testing:

if (t-statistic < 0) and ($\frac{p_{value}}{2} \leq 0.05$) **then**
  Reject the H_0
else
  Accept the H_0
end if

In addition, we statistically analyze the trendlines (assumed to be linear) for each indicator (or composite indicator) for Greece and for the EU using generalized least squares regression with a correction for autocorrelation on the years at Greece and EU level. We report the slope coefficient of the trendline (b), its standard error (se) and the level of significance of the trend (p_{value}) respectively. Then, we compare the trends for Greece and the EU in a statistical manner using a $z-test$ (see equation 3.1). The motivation and the details of our way of comparing two trendlines (linear regression lines), are given in [60]. It is again a hypothesis testing, where H_0: $b_1 = b_2$, i.e. $b_1 - b_2 = 0$ and the alternative hypothesis H_1: $b_1 \neq b_2$, i.e. $b_1 - b_2 \neq 0$. We use an alpha level of statistical significance equal to $\alpha = 0.05$, i.e. 95% statistical significance. Thus, we conclude insignificance for $p_{value} > 0.05$ (two-tailed test). Again, we report the slope coefficient of the trendline (b), its standard error (se), the z_{score} and the level of significance of the difference in trend (p_{value}) respectively.

$$Z = \frac{b_1 - b_2}{\sqrt{se_1^2 + se_2^2}}. \tag{3.1}$$

Third Stage - Machine Learning

The third stage includes the machine learning part of this research study. In order to estimate the importance of each indicator to the final output of innovation of Greece and EU, i.e. to what extend it affects the summary innovation index of Greece or EU, we utilize the following methodology. Firstly, we do a correlation analysis of the indicators providing the correlation heat-maps for interpretation purposes. We decide to drop highly correlated features from our analysis, i.e correlation coefficient above 0.90 ($r > 0.90$). This is a common tactic to avoid multicollinearity issues and get better results from model based feature importance and generally in machine learning algorithms.

Then, we construct a vector (v) which models the fluctuation of the summary innovation index over the time frame (2010-2017) with the following technique. We slide a window with step equal to one year over the time frame starting from the beginning and if the value of the summary innovation index in present year is higher than the value from previous year then $v_i = 1$, else $v_i = 0$. Thus, it is a binary classification problem with $X_{features}$ the indicators and label y the vector v. We use 3-fold cross validation on our data to train 4 machine learning models for classification. The models are Logistic Regression, Random Forest, Extra-Trees and Support Vector Machines. For each model, we average the estimates of feature importance across all 3 folds of cross validation to get a better estimate of model based feature importance. In order to get a final summary of the most important features for Greece and EU, we turn each of the aforementioned model based feature importance in percentage feature importance for each model and then we average on percentages across on 4 models. We choose to express feature importance in percentage values, because the procedure and values for calculating the most important features are different in each model. Thus, we need to average on percentage values for comparability purposes.

Finally, we discuss the most important indicators which drive and affect the most the fluctuation of summary innovation index at Greece and EU level.

Chapter 4

Data Analysis

4.1 Composite Indicators

In section A.1 of the Appendix we provide the visualizations of time series data and percentage change regarding the years 2010-2017 for each composite indicator used in this study. Table 4.1 displays inferential statistics on the aforementioned composite indicators time series data. All values in this table are rounded to 3 decimal values.

The summary innovation index of EU is higher than Greece over the whole period of our study. EU outperforms Greece in total innovation. Observing the percentage change graph of this composite indicator A.1, there is a 12% decrease of innovation in Greece level on year 2014 compared to previous year. However both EU and Greece show upward trend in summary innovation index from 2014 till now.

It is worth to highlight the composite indicator "Innovators" A.7, where Greece outperforms the EU average in all years from 2010 to 2017. We point out that this composite indicator is rather important as it is comprised of three simple indicators. In a period of economic crisis of Greece, there is a systematic over-performance in average of Greece versus EU in the share of firms that have introduced innovations onto the market or within their organizations, covering both product and process innovators, marketing and organizational innovators, and SMEs that innovate in-house.

Following, there is a statistical hypothesis testing of the systematic over-performance or under-performance of Greece compared to EU average. We state a hypothesis for testing. We denote μ_{eu} and μ_{gr} the mean of EU and Greece (GR) values. We state the null hypothesis $H_0 \rightarrow \mu_{eu} \geq \mu_{gr}$ and the alternative hypothesis $H_1 \rightarrow \mu_{eu} < \mu_{gr}$. With a two samples statistical t-test assuming unequal variances, we test whether the EU indicator has greater value in average than GR indicator. In this way, we test for over-performance or under-performance of GR compared to EU average. Table 4.1 contains the composite indicators, the $t - statistic$ of the t-test, the p_{value} and our decision on H_0 based on 95% statistical significance ($\alpha = 0.05$).

Composite Indicator	t-statistic	p_{value}	Decision on H_0
Summary_Innovation_Index	25.489	0.000	accept
Human_Resources	10.143	0.000	accept
Research_Systems	4.335	0.001	accept
Innovation-friendly_environment	14.443	0.000	accept
Finance_and_support	15.453	0.000	accept
Firm_investments	14.918	0.000	accept
Innovators	-2.977	0.013	reject
Linkages	7.203	0.000	accept
Intellectual_assets	23.950	0.000	accept
Employment_impacts	14.611	0.000	accept
Sales_impacts	5.337	0.001	accept

TABLE 4.1: Composite Indicators: T-test on H_0

As we can clearly see, we have strong evidence to reject the null hypothesis H_0 ($\frac{p_{value}}{2} \leq 0.05$), that the EU average value is higher than Greece value. This means, that we statistically confirm the systematic over-performance of Greece versus EU in the composite indicator "Innovators".

4.2 Indicators

In section A.2 of the Appendix we provide the visualizations of time series data and percentage increase versus years 2010-2017 for each indicator used in this study. Table 4.2 displays inferential statistics on the aforementioned indicators time series data. All values in this table are rounded to 3 decimal values.

We focus on indicators where Greece outperforms EU with strong statistical evidence, according to table 4.2. As clearly seen, Greece exceed EU in average in six indicators, namely "Innovative SMEs collaborating with others (% of SMEs)", "International scientific co-publications per million population", "Non-R&D innovation expenditures (% of turnover)", "Percentage population aged 25-34 having completed tertiary education", "Sales of new-to-market and new-to-firm innovations as % of turnover" and "SMEs introducing marketing or organizational innovations as % of SMEs".

We observe systematic over-performance of Greece in all years from 2010 to 2017 in all of the above mentioned indicators except "Non-R&D innovation expenditures (% of turnover)", where we have equal score in years 2016 and 2017 of Greece and EU. According to our methodology, we report the $t - statistic = -2.85$ and $p_{value} = 0.013$.

Regarding the indicator "Innovative SMEs collaborating with others (% of SMEs)", Greece has the same score from 2010 to 2013, while EU average suffers from a terrible decrease in 2012 compared to 2011 about 25%. In 2016 Greece scores show more than 20% increase compared to only 10% of EU scores compared to previous year. We note that $t - statistic = -6.438$ and $p_{value} = 0$.

Greece and EU exhibit a serious increase in indicator of International scientific co- publications per million population year by year. As we observe from graphs in figure A.18, in 2017 the indicator's score increased by nearly 60% and 65% since 2010 for Greece and EU respectively. We report that $t - statistic = -2.36$ and $p_{value} = 0.034$.

It is thought-provoking that SMEs introducing marketing or organizational innovations as % of SMEs of both Greece and EU average show a decrement through the time frame of study. Since 2010 Greece has lost nearly 28% and EU 18% of their SMEs respectively introducing marketing or organizational innovations as % of SMEs. This can be justified by figure A.33. We report that $t - statistic = -4.72$ and $p_{value} = 0.001$.

While Greece has been outperforming EU average from year 2010 to 2014 regarding the sales of new-to-market and new-to-firm innovation as% of turnover, then a huge decrease (above 50%) throws them below EU since 2014. Later in 2016, both of them show an increment of around 10%. We report that $t - statistic = -2.116$ and $p_{value} = 0.071$.

Regarding the tertiary education field, and specifically the percentage of population aged 25-34, Greece outperforms EU average over the time frame of study. The difference seems to get larger as the growth rate of Greece is getting bigger since 2014. This can be clearly observed in figure A.24, where the percentage increment of Greece is approximately double compared to EU year by year since 2014. We report that $t - statistic = -3.576$ and $p_{value} = 0.005$.

The stastical hypothesis testing is following below. We denote μ_{eu} and μ_{gr} the mean of EU and Greece (GR) values. We state the null hypothesis $H_0 \rightarrow \mu_{eu} \geq \mu_{gr}$ and the alternative hypothesis $H_1 \rightarrow \mu_{eu} < \mu_{gr}$. With a two samples statistical t-test assuming unequal variances, we test whether the EU indicator is greater value in average than GR indicator. In this way, we test for overperformance or underperformance of GR compared to EU average. The table 4.2 shows the indicators, the $t - statistic$ of the t-test, the p_{value} and our decision on H_0 based on 95% statistical significance ($\alpha = 0.05$).

In table 4.2 the names of indicators are truncated into long acronyms for readability purposes. The full names of indicators are listed below:

Broadband_penetration Broadband penetration.

Venture_capital Venture capital (% of GDP).

Design_applications Design applications per billion GDP (in PPS).

Trademark_apps Trademark applications per billion GDP (in PPS).

Employment_activities Employment in knowledge-intensive activities (% of total employment).

Enterprises_training Enterprises providing training to develop or upgrade ICT skills of their personnel.

Innovative_Smes Innovative SMEs collaborating with others (% of SMEs).

International_publications International scientific co-publications per million population.

Knowledge_exports Knowledge-intensive services exports as % of total services exports.

New_doctorate_grads New doctorate graduates per 1000 population aged 25-34.

Non_rd Non-R&D innovation expenditures (% of turnover).

Opportunity_enterpre Opportunity-driven entrepreneurship (Motivational index).

Pct_patent PCT patent applications per billion GDP (in PPS).

Percentage_tertiary_edu Percentage population aged 25-34 having completed tertiary education.

Percentage_lifelong_learning Percentage population aged 25-64 involved in lifelong learning.

Private_co_funding Private co-funding of public R&D expenditures (percentage of GDP).

Public_private_pubs Public-private co-publications per million population.

Rd_business R&D expenditure in the business sector (% of GDP).

Rd_public R&D expenditure in the public sector (% of GDP).

Sales Sales of new-to-market and new-to-firm innovations as % of turnover.

Scientific_pubs Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country.

Smes_in_house SMEs innovating in-house as % of SMEs.

Smes_marketing SMEs introducing marketing or organizational innovations as % of SMEs.

Smes_product SMEs introducing product or process innovations as % of SMEs.

Exports_technology Exports of medium and high technology products as a share of total product exports.

Indicator	t-statistic	p_{value}	Decision on H_0
Broadband_penetration	8.252	0.000	accept
Venture_capital	12.576	0.000	accept
Design_applications	29.715	0.000	accept
Trademark_apps	7.914	0.000	accept
Employment_activities	8.434	0.000	accept
Enterprises_training	8.047	0.000	accept
Innovative_Smes	-6.438	0.000	reject
International_publications	-2.360	0.034	reject
Knowledge_exports	5.828	0.001	accept
New_doctorate_grads	8.001	0.000	accept
Non_rd	-2.850	0.013	reject
Opportunity_enterpre	27.904	0.000	accept
Pct_patent	65.418	0.000	accept
Percentage_tertiary_edu	-3.576	0.005	reject
Percentage_lifelong_learning	37.953	0.000	accept
Private_co_funding	8.679	0.000	accept
Public_private_pubs	33.776	0.000	accept
Rd_business	32.700	0.000	accept
Rd_public	6.434	0.000	accept
Sales	-2.116	0.071	reject
Scientific_pubs	14.342	0.000	accept
Smes_in_house	-0.970	0.355	accept
Smes_marketing	-4.720	0.001	reject
Smes_product	-1.570	0.143	accept
Exports_technology	41.010	0.000	accept

TABLE 4.2: Indicators: T-test on H_0

Chapter 5

Predictive Analytics

5.1 Trend Analysis

5.1.1 Trend Analysis for Composite Indicators

This section provides trend analysis for composite indicators of Greece and EU average. In addition, we statistically compare the two trendlines. In table 5.1 we observe trendline statistics of composite indicators for Greece and the EU, indicating the upward or downward trends. All values in this table are rounded to 3 decimal values.

Firstly, we highlight the summary of innovation score. The observable trends in Summary Innovation Index in aforementioned tables are positive for EU, but negative for Greece. The score of Greece decreases by factor of 0.2% per year, while the score of the EU increases 0.5% per year. However, the slope of Greece's trendline is not significant at 95% statistical significance ($p_{value} = 0.575$), while EU is statistical significant ($p_{value} = 0.011$). Please note that the difference between the trendlines of Greece and EU is calculated as described in 3.3 using a statistical $z - test$, according to equation 3.1. This is a hypothesis testing where we aim to test if the two slopes of the trendlines are significantly different. Thus, H_0: $b_1 = b_2$, i.e. $b_1 - b_2 = 0$ and the alternative hypothesis H_1: $b_1 \neq b_2$, i.e. $b_1 - b_2 \neq 0$. The difference of Greece and EU trendlines in Summary Innovation Index is statistically significant (b = -0.007, se = 0.003, $p_{value} < 0.05$).

We focus on statistically significant trends ($p_{value} < 0.05$). So, for the rest of the document we state only the trends that have $p_{value} < 0.05$. For Greece, there is upward trend in Human Resources, Research Systems, Finance and Support, Intellectual Assets, and Sale Impacts. Human Resources and Research Systems exhibit an increment by a factor of 1.1% every year, while Finance and Support and Intellectual Assets 1.9% and 1.6% respectively. Sale Impacts of Greece decrease by a factor of 4.5% every year.

EU shows upward trend in Summary Innovation Index, Human Resources, Research Systems, Innovation-friendly Environment, Firm Investments and Innovators. EU Human Resources and Firm investments show an increment of 1.4% and 1.3% every year respectively. Research systems and Innovation-friend Environment display increase in score of 0.7% and 1.7% every year respectively. However Innovators in EU average present a decrease by 1.7%.

We compare the trends of Greek composite indicators versus EU in a statistical manner in table 5.1, as described above. Once again, we highlight only the statistical significant results, which means that $H_0 : \quad b_{gr} = b_{eu}$ can be rejected at 95% significance ($p_{value} < 0.05$). Thus, we observe significant difference in trendlines between EU and Greece in Summary Innovation Index, Research Systems, Firm Investments, Intellectual Assets, Employment Impacts and Sales Impacts. We report the slope

difference b_{diff}, the standard error se, the z_{score} of the statistical test and the p_{value}, in table 5.1.

5.1.2 Trend Analysis for Indicators

This section provides trend analysis for simple indicators of Greece and EU average. In addition, we statistically compare the two trendlines. In table 5.2 we observe trendline statistics of Greece and EU indicators respectively. All values in this table are rounded to 3 decimal values.

We indicate the upward or downward trends in both trendlines of Greece and EU and then we statistically compare the two trendlines. We concentrate again on indicators with $p_{value} < 0.05$, these are the gray shaded values in tables. We distinguish positive and negative trends and note the factor of growth or decrease each year in parenthesis.

For Greece level, we witness upward trend in the following indicators:

- Design applications per billion GDP (in PPS) (1.5%),
- Trademark applications per billion GDP (in PPS) (3.0%),
- International scientific co-publications per million population (1.5%),
- Percentage population aged 25-34 having completed tertiary education (1.9%),
- Percentage population aged 25-64 involved in lifelong learning (0.9%),
- R&D expenditure in the business sector (% of GDP) (1.1%),
- R&D expenditure in the public sector (% of GDP) (4.6%).

Furthermore, there is a decreasing trend in:

- Venture capital (% of GDP) (0.7%),
- Knowledge-intensive services exports as % of total services exports (4.2%),
- Sales of new-to-market and new-to-firm innovations as % of turnover (10%),
- SMEs introducing marketing or organizational innovations as % of SMEs (4.4%).

Greece's R&D expenditure in the public sector and Trademark applications are growing in a fast pace. However Sales of new-to-market and new-to-firm innovations of Greece are going downwards rapidly.

For EU level, we witness upward trend in the following indicators:

- Broadband penetration (3.5%),
- Venture capital (% of GDP) (3.5%),
- Trademark applications per billion GDP (in PPS) (0.5%),
- Employment in knowledge-intensive activities (% of total employment) (0.6%),
- Enterprises providing training to develop or upgrade ICT skills of their personnel (1.6%),
- International scientific co-publications per million population (1.3%),

	GR			EU			Difference			
Trendline	b	se	p_{value}	b	se	p_{value}	b_{diff}	se	z_{score}	p_{value}
Summary_Innovation_Index	-0.002	0.003	0.575	0.005	0.001	0.011	-0.007	0.003	-2.227	0.013
Human_Resources	0.011	0.002	0.008	0.014	0.002	0.000	-0.003	0.003	-1.086	0.139
Research_Systems	0.011	0.001	0.001	0.007	0.001	0.000	0.004	0.002	2.251	0.012
Innovation-friendly_environment	0.005	0.005	0.393	0.017	0.006	0.032	-0.013	0.008	-1.621	0.053
Finance_and_support	0.019	0.004	0.004	0.016	0.006	0.054	0.004	0.007	0.520	0.302
Firm_investments	0.002	0.003	0.446	0.013	0.003	0.010	-0.010	0.004	-2.489	0.006
Innovators	-0.025	0.010	0.062	-0.017	0.004	0.008	-0.008	0.011	-0.717	0.237
Linkages	0.001	0.004	0.785	0.003	0.004	0.419	-0.002	0.005	-0.386	0.350
Intellectual_assets	0.016	0.002	0.000	-0.002	0.001	0.121	0.018	0.002	9.425	0.000
Employment_impacts	0.009	0.005	0.149	-0.001	0.002	0.727	0.010	0.006	1.726	0.042
Sales_impacts	-0.045	0.010	0.005	0.006	0.002	0.060	-0.051	0.010	-5.182	0.000

TABLE 5.1: Composite Indicators Trendline Statistics - Greece (GR), EU and Difference between EU and Greece (GR). With gray color, we denote statistical significance at 95% level ($\alpha = 0.05$).

- Knowledge-intensive services exports as % of total services exports (0.6%),
- New doctorate graduates per 1000 population aged 25-34 (3.1%),
- Percentage population aged 25-34 having completed tertiary education (0.9%),
- Percentage population aged 25-64 involved in lifelong learning (0.1%),
- R&D expenditure in the business sector (% of GDP) (0.8%),
- Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country (0.3%),
- Exports of medium and high technology products as a share of total product exports (1.4%).

In addition, we observe a downward trend in:

- PCT patent applications per billion GDP (in PPS) (0.6%),
- R&D expenditure in the public sector (% of GDP) (0.4%),
- SMEs introducing marketing or organizational innovations as % of SMEs (1.2%),
- SMEs introducing product or process innovations as % of SMEs (2.0%).

We characterize the rate of change in indicators of EU more stable than of Greece. Actually, this is because EU indicator scores are average values of EU countries. However we want to highlight the fast pace of increment in broadband penetration, venture capitals and new doctorate graduates. EU SMEs introducing product or process innovations is decreasing by a factor of 2.0% year by year.

We present the statistical significant differences of Greece and EU trendlines in table 5.2, regarding the simple indicators. We report the slope difference b_{diff}, the standard error se, the z_{score} of the statistical test and the p_{value}. Following, there is a distinction of significant differences in trendlines:

- Design applications per billion GDP (in PPS),
- Trademark applications per billion GDP (in PPS),
- Sales of new-to-market and new-to-firm innovations as % of turnover,
- Broadband penetration,
- Venture capital (% of GDP),
- Enterprises providing training to develop or upgrade ICT skills of their personnel,
- International scientific co-publications per million population,
- Knowledge-intensive services exports as % of total services exports,
- New doctorate graduates per 1000 population aged 25-34,
- Percentage population aged 25-34 having completed tertiary education,
- Percentage population aged 25-64 involved in lifelong learning,

- PCT patent applications per billion GDP (in PPS),
- R&D expenditure in the public sector (% of GDP),
- SMEs introducing marketing or organizational innovations as % of SMEs.

	GR			EU			Difference		
Trendline	b	se	p_{value}	b	se	p_{value}	b_{diff}	se	z_{score}	p_{value}
Broadband_penetration	0.010	0.006	0.144	0.035	0.009	0.010	-0.025	0.010	-2.391	0.008
Venture_capital	-0.007	0.002	0.011	0.035	0.013	0.048	-0.042	0.013	-3.139	0.001
Design_applications	0.015	0.002	0.000	-0.003	0.001	0.068	0.019	0.002	8.425	0.000
Trademark_apps	0.030	0.003	0.000	0.005	0.001	0.006	0.025	0.003	7.088	0.000
Employment_activities	0.009	0.005	0.151	0.006	0.001	0.000	0.003	0.005	0.506	0.307
Enterprises_training	-0.010	0.010	0.372	0.016	0.005	0.021	-0.026	0.011	-2.348	0.009
Innovative_Smes	0.010	0.008	0.244	0.010	0.009	0.298	0.000	0.012	0.030	0.488
International_publications	0.015	0.001	0.000	0.013	0.000	0.000	0.002	0.001	2.041	0.021
Knowledge_exports	-0.042	0.004	0.000	0.006	0.001	0.003	-0.048	0.004	-11.882	0.000
New_doctorate_grads	0.004	0.003	0.235	0.031	0.006	0.004	-0.027	0.007	-4.013	0.000
Non_rd	0.005	0.007	0.498	0.014	0.007	0.114	-0.008	0.010	-0.807	0.210
Opportunity_enterpre	-0.001	0.005	0.896	-0.000	0.003	0.937	-0.000	0.006	-0.062	0.475
Pct_patent	0.003	0.001	0.069	-0.006	0.001	0.007	0.009	0.002	4.745	0.000
Percentage_tertiary_edu	0.019	0.004	0.007	0.009	0.002	0.009	0.010	0.005	2.149	0.016
Percentage_lifelong_learning	0.009	0.002	0.008	0.001	0.000	0.039	0.008	0.002	3.696	0.000
Private_co_funding	-0.003	0.008	0.683	-0.001	0.001	0.240	-0.002	0.008	-0.247	0.403
Public_private_pubs	-0.004	0.003	0.314	0.001	0.002	0.698	-0.004	0.004	-1.183	0.118
Rd_business	0.011	0.003	0.011	0.008	0.001	0.001	0.003	0.003	1.063	0.144
Rd_public	0.046	0.008	0.003	-0.004	0.001	0.042	0.050	0.008	5.900	0.000
Sales	-0.100	0.030	0.021	-0.002	0.006	0.809	-0.098	0.031	-3.197	0.001
Scientific_pubs	0.006	0.003	0.085	0.003	0.001	0.009	0.003	0.003	1.085	0.139
Smes_in_house	-0.012	0.014	0.445	-0.012	0.005	0.063	0.001	0.015	0.038	0.485
Smes_marketing	-0.044	0.007	0.001	-0.020	0.004	0.002	-0.024	0.008	-3.102	0.001
Smes_product	-0.018	0.015	0.279	-0.018	0.005	0.012	-0.000	0.015	-0.016	0.494
Exports_technology	0.006	0.005	0.312	0.014	0.003	0.005	-0.008	0.006	-1.366	0.086

TABLE 5.2: Indicators Trendline Statistics - Greece (GR), EU and Difference between EU and Greece (GR). With gray color, we denote statistical significance at 95% level ($\alpha = 0.05$).

5.2 Machine Learning

Next we present our results obtained through machine learning techniques. We use classification methods to estimate the importance of each indicator to the summary innovation index fluctuation over the time frame of study. Specifically we utilize Logistic Regression, SVM (linear kernel), Random Forest Classifier and Extra Trees Classifier. We cross validate the training of models (3-fold cross validation) and by keeping the model estimate in each fold, we average the feature importance on 3 estimators of each classifier. We thus summarize, the four model-based feature importance values and we report the indicator-feature importance in percentage. For more information on our methodology, please backtrack in Analytics Work-flow & Methodology, also in background information in Statistics, Machine Learning Algorithms & Techniques.

5.2.1 Indicator Importance

This section analyzes the results of our methodology in order to indicate the most important indicators. We use correlation analysis in order to avoid including the highly correlated features in our analysis. We then briefly describe our modeling technique of the time series classification process. Tables 5.3 and 5.4 refer to the summary feature importance in percentage of all models (average values).

Correlation Analysis

Figures 5.1 and 5.2 show the correlation heat-maps for EU and Greek indicators respectively. We use heat-maps for pairwise correlation analysis for visualization and interpretation purposes. In order to avoid multicollinearity issues the threshold for dropping the highly correlated features is set 0.90. Thus, we drop the features having 0.90 correlation from our analysis.

EU

According to our technique, we leave out the following indicators in EU level (figure 5.1):

- Employment in knowledge-intensive activities (% of total employment),
- International scientific co-publications per million population,
- Knowledge-intensive services exports as % of total services exports,
- New doctorate graduates per 1000 population aged 25-34,
- Non-R&D innovation expenditures (% of turnover),
- Percentage population aged 25-34 having completed tertiary education,
- R&D expenditure in the business sector (% of GDP),
- Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country,
- SMEs introducing marketing or organizational innovations as % of SMEs,
- SMEs introducing product or process innovations as % of SMEs.

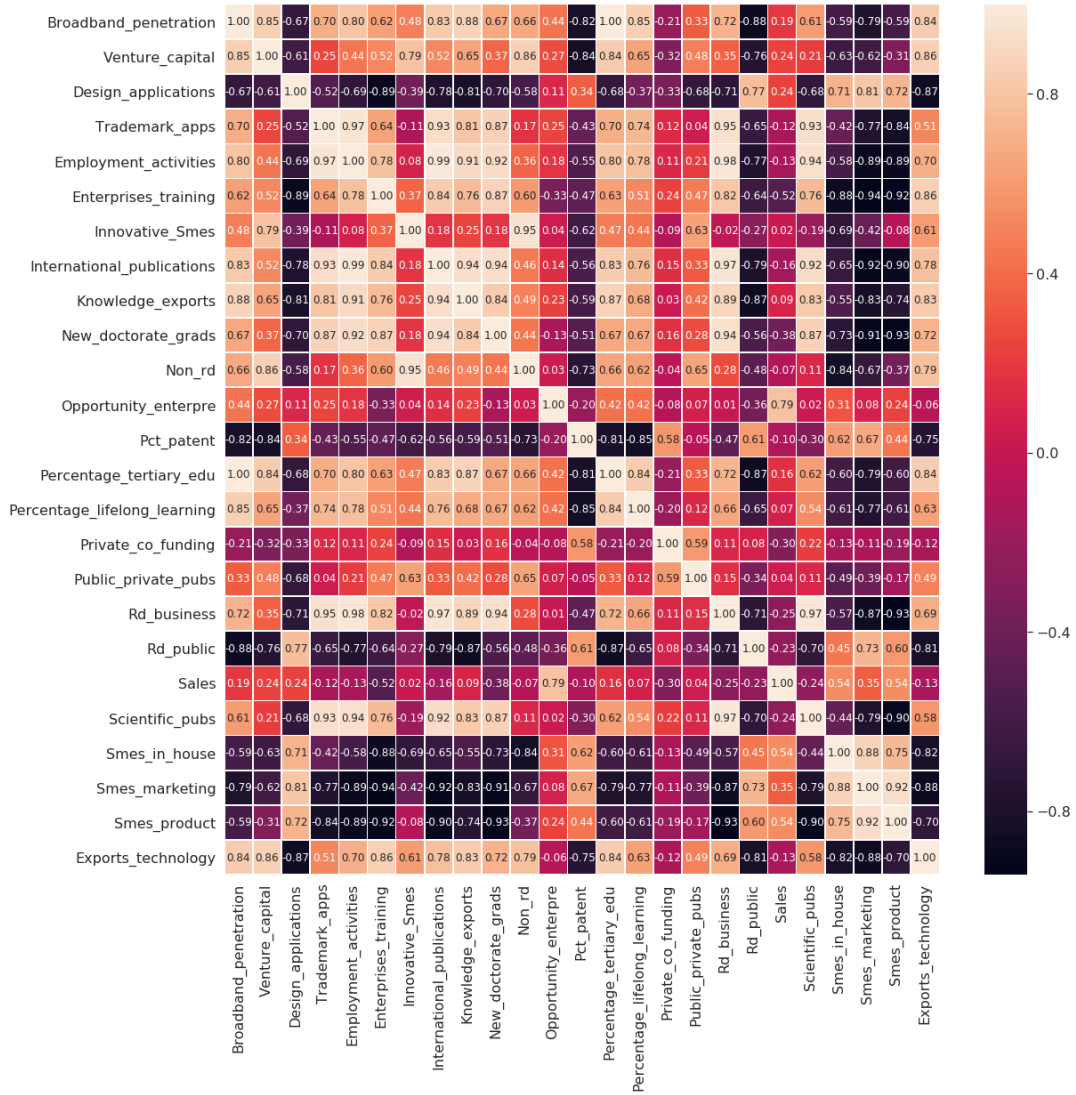

FIGURE 5.1: Correlation Heat-map: Indicators-EU

Greece

According to our technique, we leave out the following indicators in Greece level (figure 5.2):

- Trademark applications per billion GDP (in PPS),
- Employment in knowledge-intensive activities (% of total employment),
- International scientific co-publications per million population,
- Knowledge-intensive services exports as % of total services exports,
- Percentage population aged 25-64 involved in lifelong learning,
- R&D expenditure in the business sector (% of GDP),
- R&D expenditure in the public sector (% of GDP),
- Sales of new-to-market and new-to-firm innovations as % of turnover,

- Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country,
- SMEs innovating in-house as % of SMEs,
- SMEs introducing marketing or organizational innovations as % of SMEs,
- SMEs introducing product or process innovations as % of SMEs.

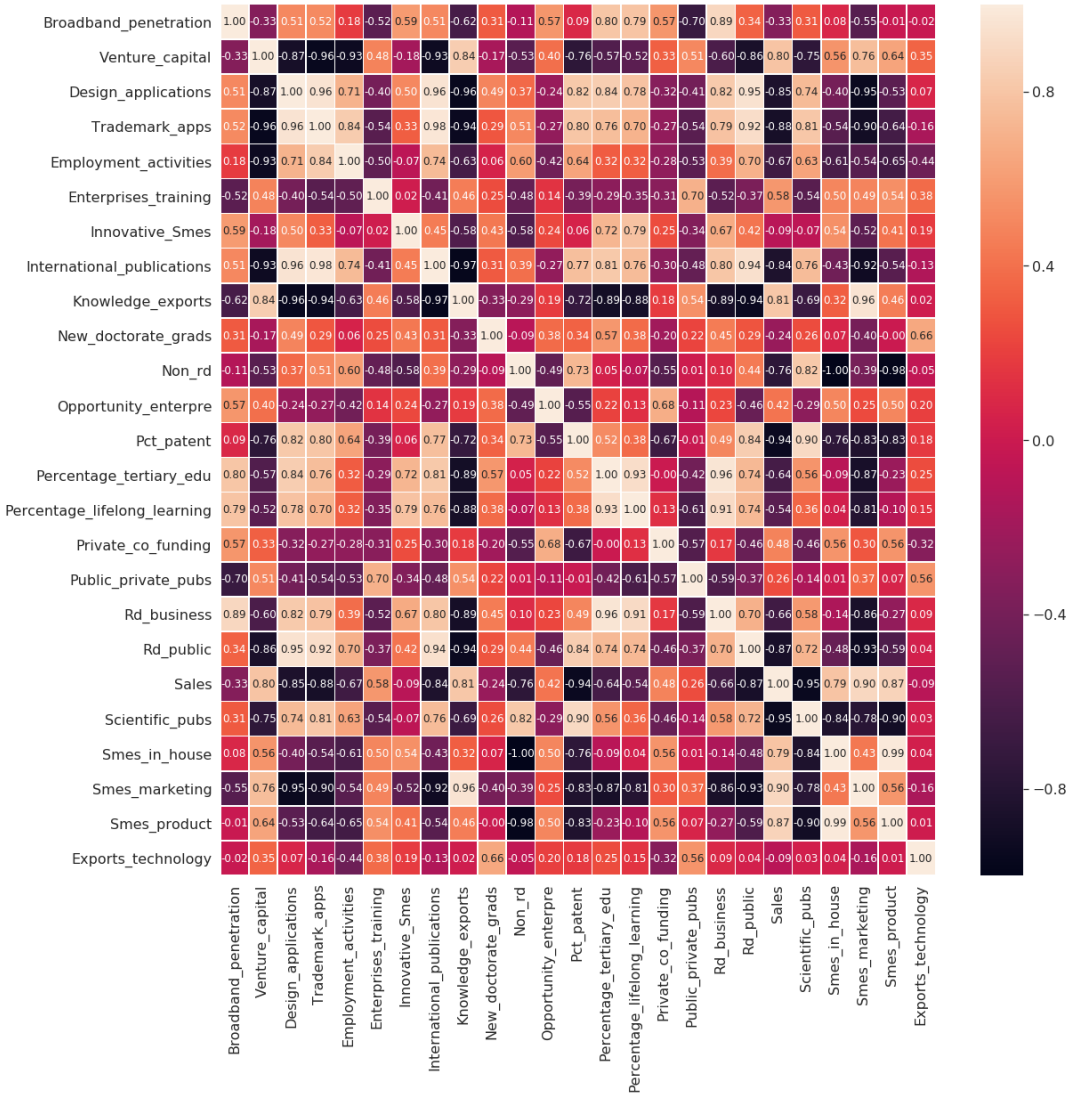

FIGURE 5.2: Correlation Heat-map: Indicators-Greece

Modeling Technique

Next we describe our time series modeling technique for classification purpose. We construct a vector (v) which models the fluctuation of the summary innovation index over the time frame (2010-2017) with the following technique. We slide a window with step equal to one year over the time frame starting from the beginning and if the value of the summary innovation index in present year is higher than the value from previous year then $v_i = 1$, else $v_i = 0$. Thus, it is a binary classification problem with $X_{features}$ the indicators and label y the vector v.

Then, we train the four machine learning models, Logistic Regression, SVM (linear kernel), Random Forest Classifier and Extra Trees Classifier, using 3-fold cross validation method. We keep the model's estimate for feature importance in each of the 3 fold and then the final value of importance is calculated as the average of these for each model. We transform each model based importance value in percentage for each model. Then, we average all percentage values of indicator importance from each model to summarize the importance. Following, there are the two tables summarizing feature importance in percentage for EU (table 5.3) and Greece (table 5.4) respectively.

EU

As we observe in table 5.3, the top-five important features affecting the most the fluctuation of summary innovation value of EU are the Venture capital, Exports of medium and high technology products, Broadband penetration, Design applications, Public-private co-publications.

Indicator	Importance (%)
Venture_capital	11.49
Exports_technology	9.94
Broadband_penetration	8.58
Design_applications	8.10
Public_private_pubs	7.67
Private_co_funding	7.50
Opportunity_enterpre	7.39
Pct_patent	6.69
Smes_in_house	6.27
Trademark_apps	6.18
Innovative_Smes	5.97
Enterprises_training	5.92
Rd_public	3.83
Sales	3.81
Percentage_lifelong_learning	0.66

TABLE 5.3: Indicator Importance: EU

On the other side, the top-five important features affecting the most the fluctuation of summary innovation value of Greece are the Design applications, Venture capital, Percentage population aged 25-34 having completed tertiary education, New doctorate graduates and Innovative SMEs collaborating with others.

Greece

Indicator	Importance (%)
Design_applications	15.57
Venture_capital	14.29
Percentage_tertiary_edu	12.40
New_doctorate_grads	9.02
Innovative_Smes	7.55
Pct_patent	7.00
Exports_technology	7.00
Opportunity_enterpre	6.26
Private_co_funding	5.10
Non_rd	5.01
Public_private_pubs	4.45
Broadband_penetration	3.59
Enterprises_training	2.75

TABLE 5.4: Indicator Importance: Greece

In figure 5.3, we visualize the importance of indicators of Greece versus the EU for comparison purposes.

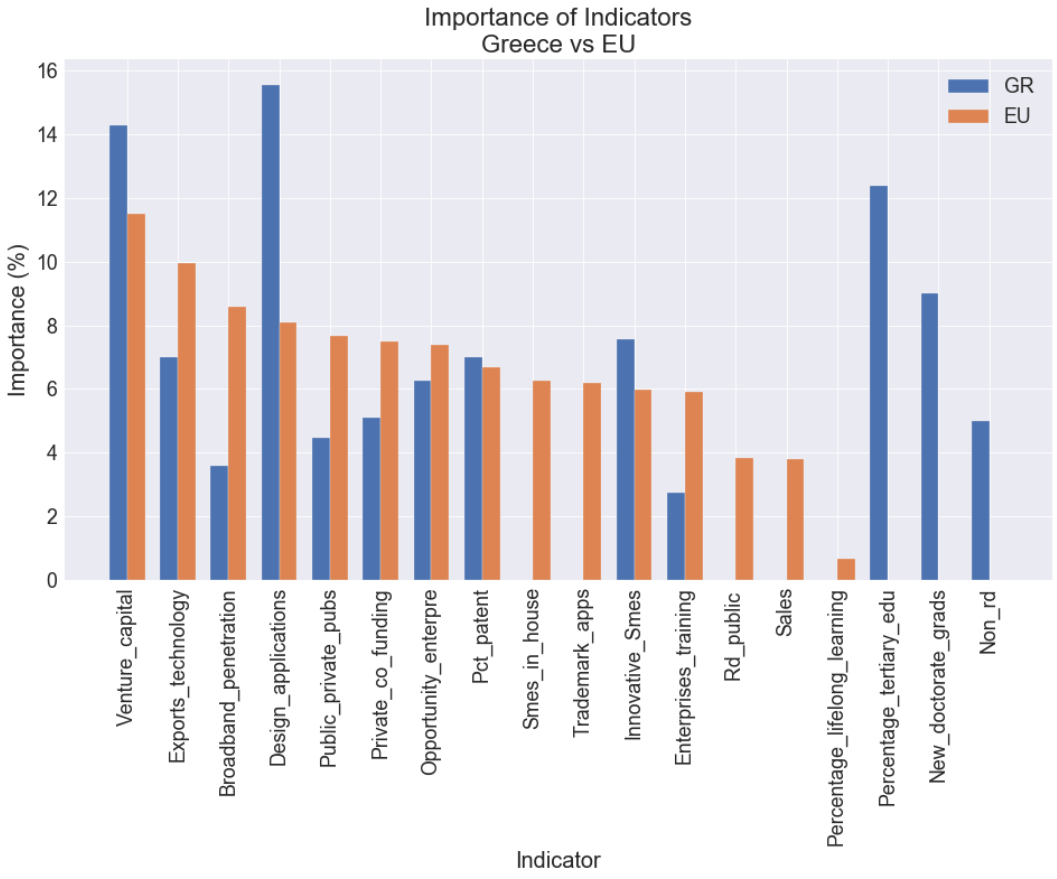

FIGURE 5.3: Importance of Indicators - Greece vs EU

Venture capital plays an important role in the innovation output of both Greece and EU. In fact, it is the fuel of start-ups and entrepreneurship. It facilitates innovations and allows them to be developed into marketable products. It enables the financing of business ideas that would otherwise not have a chance of gaining access to the necessary capital. In addition, designing innovative applications and products plays surely a fundamental role in country's innovation output.

According to our results, except venture capital and designing applications and products, EU average gives a high credit in high level features to increase innovation such as exports of medium and high technology products, broadband penetration and public-private co-publications. It is meaningful that exports of technology products play an important role to innovation output. In addition, research linkages and active collaboration activities between business sector researchers and public sector researchers resulting in academic publication is considerable in increment of innovation. Actually research is mostly what drives innovation. Also, facilities and especially high speed internet and networking consolidate the e-potential of EU. Realizing Europe's full e-potential depends on creating the conditions for electronic commerce and the Internet to flourish. Broadband penetration plays an interesting role in innovation output.

Furthermore, Greece seems to rely also a lot on well-educated people and innovative SMEs in order to increase innovation output. New doctorate graduates and people 25-34 having completed tertiary education play certainly an important role on innovation output. In fact, innovative ideas come mostly from educated people. SMEs in Greece represent 99,9% of the total private sector of the country. Specifically, micro enterprises (1-5 employees and below 1mil. revenues) represent about 96,6% of the private sector and about 56% of the total employment of the Greek economy. In those terms, SMEs are the most significant part of the Greek and European economy, affecting directly both the financial and the social aspects of economic life. Innovative SMEs collaborating with other enterprises or institutions is an important indicator for Greece's innovation. It seems that, the flow of knowledge between public research institutions and firms, and between firms and other firms is significant to innovation output.

Chapter 6

Conclusion

6.1 Results Summary

This section summarizes all the results from our analysis. We provide comparisons of Greece and EU average. We assess systematic overperformance or underperformance of Greece compared to EU. We report the trends of indicators, upwards (positive trend) or downwards (negative trend), and the importance of each indicator to total innovation output regarding Greece and EU.

In tables 6.1 and 6.2, we summarize the innovativeness of Greece compared to EU. The tables include only statistical significant values. Please note that "—" denotes not enough statistical evidence to decide (statistical insignificant), or missing values for Greece level. The two missing indicators for Greece level are below:

Foreign_doctor: Foreign doctorate students as a % of all doctorate students,

Employment_fast-growing: Employment in fast-growing enterprises (% of total employment).

However, for completion purposes we include these indicators in our final table of indicators.

Composite Indicator	Performance Score	Trend	Trend relative to EU
Summary_Innovation_Index	Lower	—	Negative
Human_Resources	Lower	Positive	—
Research_Systems	Lower	Positive	Positive
Innovation-friendly_environment	Lower	—	—
Finance_and_support	Lower	Positive	—
Firm_investments	Lower	—	Negative
Innovators	Higher	—	—
Linkages	Lower	—	—
Intellectual_assets	Lower	Positive	Positive
Employment_impacts	Lower	—	Positive
Sales_impacts	Lower	Negative	Negative

TABLE 6.1: Summary Performance of Innovativeness of Greece (Composite Indicators)

Indicator	Performance Score	Trend	Trend relative to EU
Broadband_penetration	Lower	—	Negative
Venture_capital	Lower	Negative	Negative
Design_applications	Lower	Positive	Positive
Trademark_apps	Lower	Positive	Positive
Employment_activities	Lower	—	—
Enterprises_training	Lower	—	Negative
Innovative_Smes	Higher	—	—
International_publications	Higher	Positive	Positive
Knowledge_exports	Lower	Negative	Negative
New_doctorate_grads	Lower	—	Negative
Non_rd	Higher	—	—
Opportunity_enterpre	Lower	—	—
Pct_patent	Lower	—	Positive
Percentage_tertiary_edu	Higher	Positive	Positive
Percentage_lifelong_learning	Lower	Positive	Positive
Private_co_funding	Lower	—	—
Public_private_pubs	Lower	—	—
Rd_business	Lower	Positive	—
Rd_public	Lower	Positive	Positive
Sales	Higher	Negative	Negative
Scientific_pubs	Lower	—	—
Smes_in_house	Lower	—	—
Smes_marketing	Higher	Negative	Negative
Smes_product	Lower	—	—
Exports_technology	Lower	—	—
Foreign_doctor	—	—	—
Employment_fast-growing	—	—	—

TABLE 6.2: Summary Performance of Innovativeness of Greece (Indicators)

Trendline analysis is commonly used as a forecasting tool. Thus, the above tables, provide us with statistical evidence to make predictions on innovativeness of Greece about the years to come. Below we highlight the five most important indicators, according to our methodology, affecting the fluctuation of summary innovation index. We round the percentage values to integer values for clarity.

Indicator	Importance (%)
Design_applications	16
Venture_capital	14
Percentage_tertiary_edu	12
New_doctorate_grads	9
Innovative_Smes	8

TABLE 6.3: Top-five Indicator Importance: Greece

Indicator	Importance (%)
Venture_capital	12
Exports_technology	10
Broadband_penetration	9
Design_applications	8
Public_private_pubs	8

TABLE 6.4: Top-five Indicator Importance: EU

Our data suggest that Greece should take actions to increase the innovation output of country by focusing not only on the top-five important indicators of Greece level, indicated above, but also try to follow the model of EU towards the increment of indicators highlighted in table 6.4.

6.2 Conclusion

In this research, an analysis of innovation and especially indicators of innovation regarding Greece relative to the European Union (EU) is presented. We use data from European Innovation Scoreboard version 2018. Specifically we select the normalized scores of composite indicators and simple indicators from 2010-2017 of Greece and EU average. Data charts of indicators and percentage change each year are presented in the data analysis. We compare Greece with EU average in country-level innovativeness. By utilizing statistics and hypothesis testing, we evaluate the case of Greece relative to EU. Overperformance of Greece versus EU is found in the composite indicator of Innovators and in simple indicators, namely Innovative SMEs collaborating with others, International scientific co-publications, Non-R&D innovation expenditures, Percentage population aged 25-34 having completed tertiary education, Sales of new-to-market and new-to-firm innovations and SMEs introducing marketing or organizational innovations.

In addition, we analyze and compare the linear trendlines of Greece and EU indicators in a statistical manner. The method, we used for linear regression lines is generalized least-squares (GLS) regression method with autoregressive errors (AR(1)), namely GLSAR. Greece shows positive significant trend relative to EU in composite indicators, namely Research systems, Intellectual assets and Employment impacts, and in indicators, namely Design applications, Trademark applications, International scientific co-publications, PCT patent applications, Percentage population aged 25-34 having completed tertiary education, Percentage population aged 25-64 involved in lifelong learning, R&D expenditure in the public sector. For more information on systematic overperformance or underperformance of Greece versus EU and trendline analysis, please read summary tables 6.1 and 6.2.

By employing a modeling technique on summary innovation index, we evaluate the effect of indicators on its fluctuation. We implement a model-based feature importance analysis for Greece and EU using four well known classifier models. Specifically the models are Logistic Regression, Random Forest, Extra-Trees and Support Vector Machines. Indicator correlation analysis provide us with evidence to exclude some indicators from our modeling. The fact that we have limited data instances leads us to cross-validate training of models, specifically 3-fold cross-validation. We decide to keep the estimate of feature importance of each fold an then average for each model. Finally, after transforming values of importance in percentage values for each model we summarize the percentage values. For further information on

indicator importance of EU and Greece, please read tables 5.3 and 5.4 respectively, or the more compact tables 6.4 and 6.3.

To sum up, we believe that this research study provides explanations and evidence to help the country assess its strengths and weaknesses regarding its innovation performance and as an extension its economic growth. By comparisons with EU countries (average), we display the position of Greece relative to the European Union.

6.3 Future Work

Future work concerns deeper analysis of particular mechanisms of innovation and new proposals to try different methods. Surely, we plan to use data from different and multiple data sources, as other related studies, presented in 2.4. Data sources such as interviews of executive managers or chief executives of enterprises or academics from numerous institutions will be helpful to examine more in depth particular indicators of innovation.

Closing, we have done a similar study in the regional level, analyzing and comparing the innovation performance of regions using the indicators provided by the Regional Innovation Scoreboard database. Our aim is to publish this study soon.

Appendix A

Appendix

In this Appendix we include all visualizations of time series data regarding the composite indicators and simple indicators. There is a compact definition for each indicator above each figure. For further information on indicators, please look on chapter 2, specifically section 2.1 where composite indicators and simple indicators are explained.

A.1 Composite Indicators Charts

Definition 1 *The Summary Innovation Index summarizes the range of different indicators of innovation and measures the total innovation performance.*

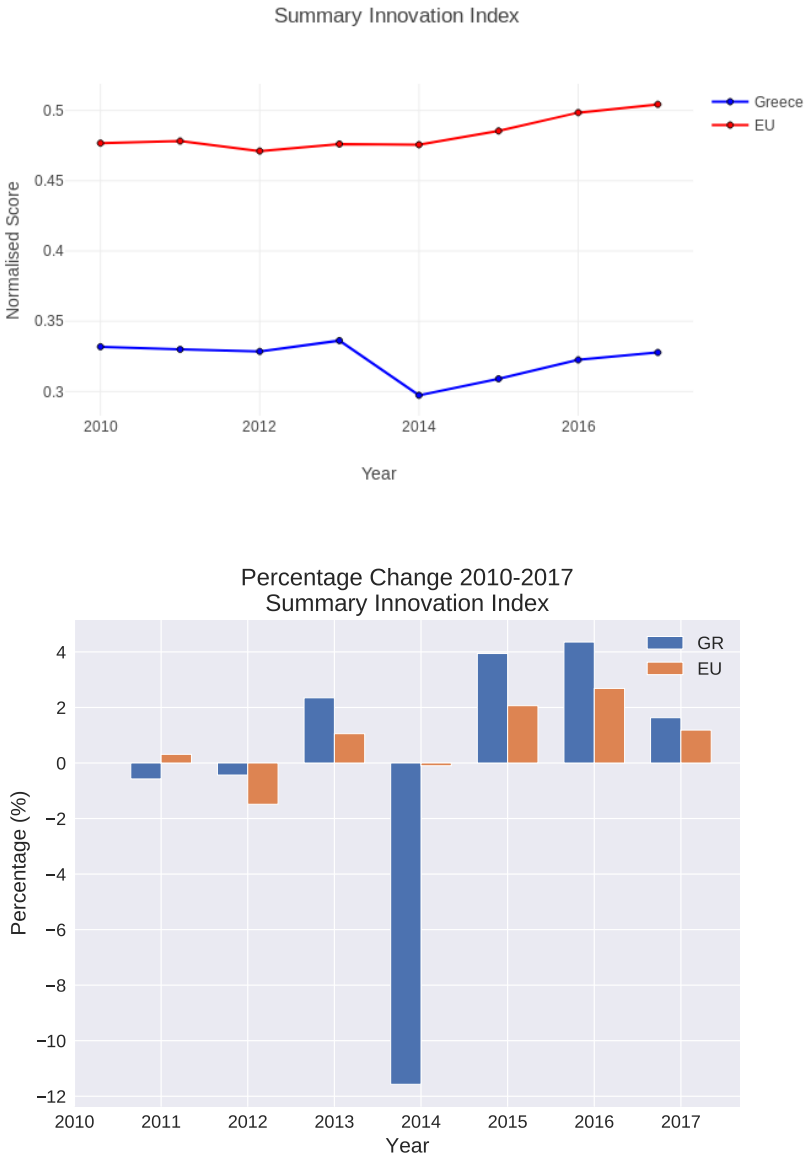

FIGURE A.1: Summary Innovation Index (on the top) and percentage change (at the bottom).

Definition 2 *The Human resources composite indicator calculates the availability of a high-skilled and educated workforce.*

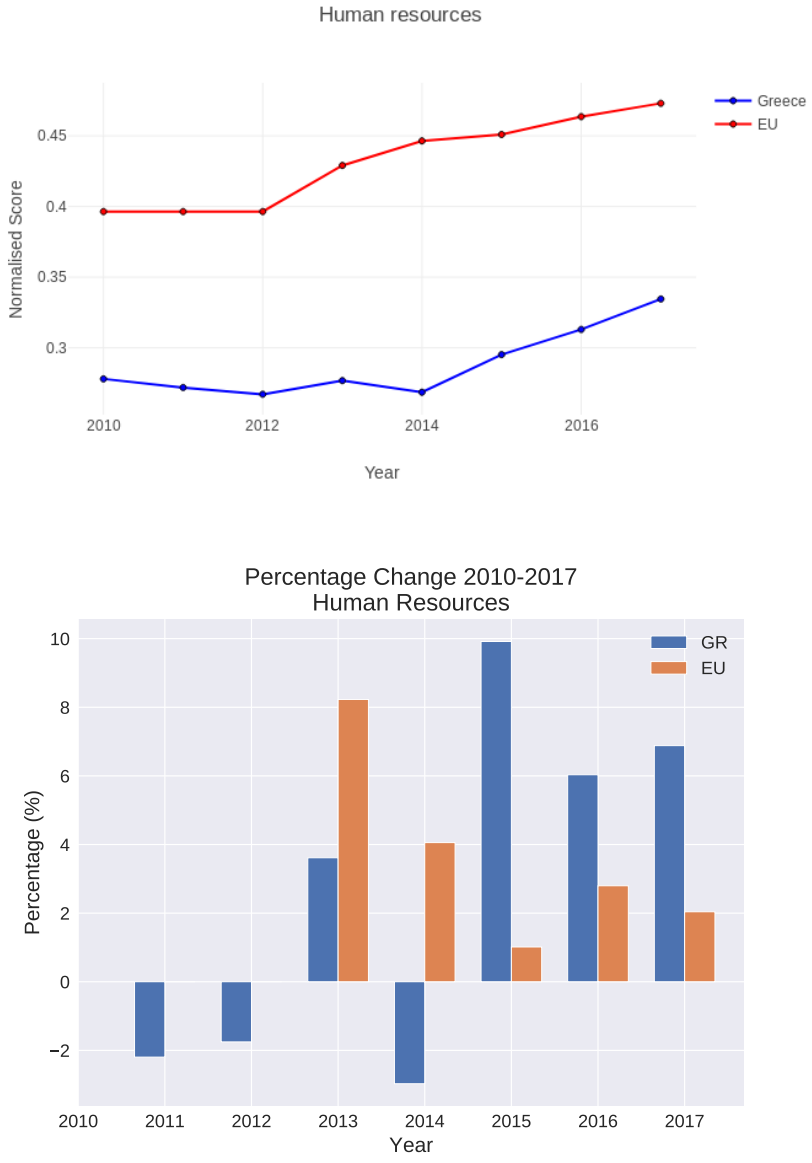

FIGURE A.2: Human resources (on the top) and percentage increase (at the bottom).

Definition 3 *Research systems includes gauges the international competitiveness of the science.*

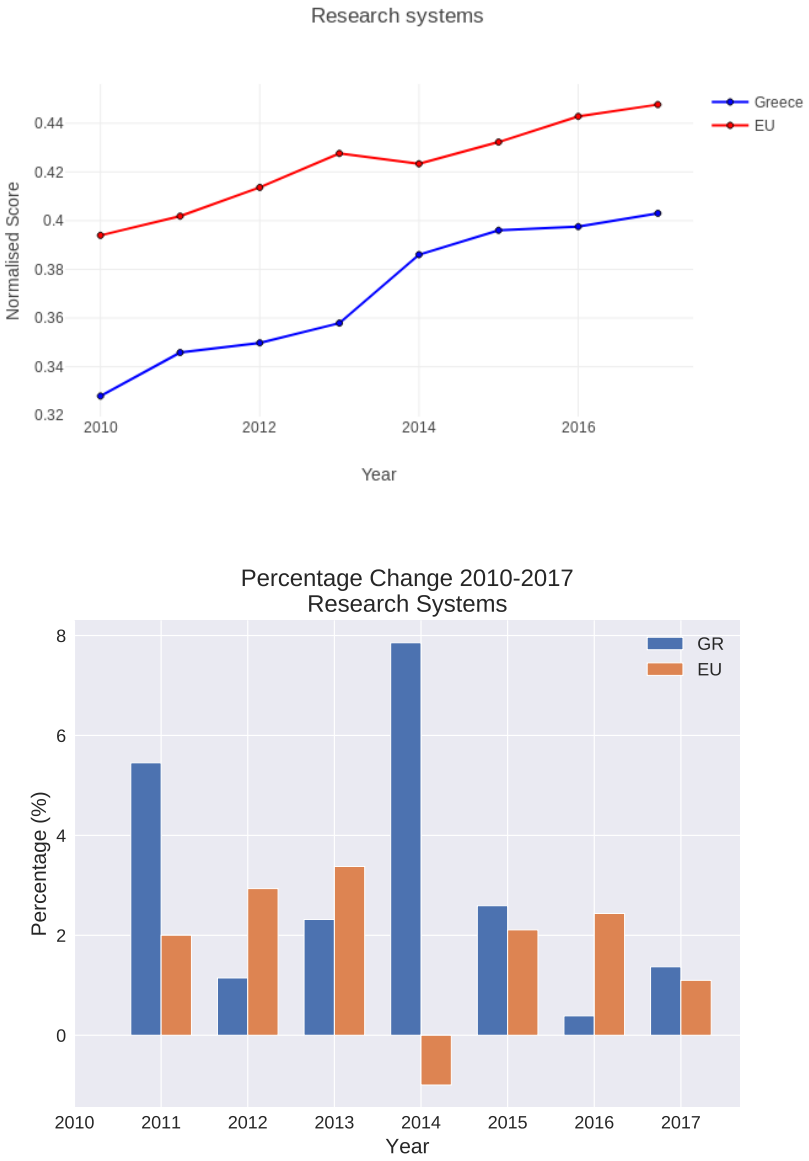

FIGURE A.3: Research systems (on the top) and percentage increase (at the bottom).

Definition 4 *Innovation-friendly environment captures the environment in which enterprises operate.*

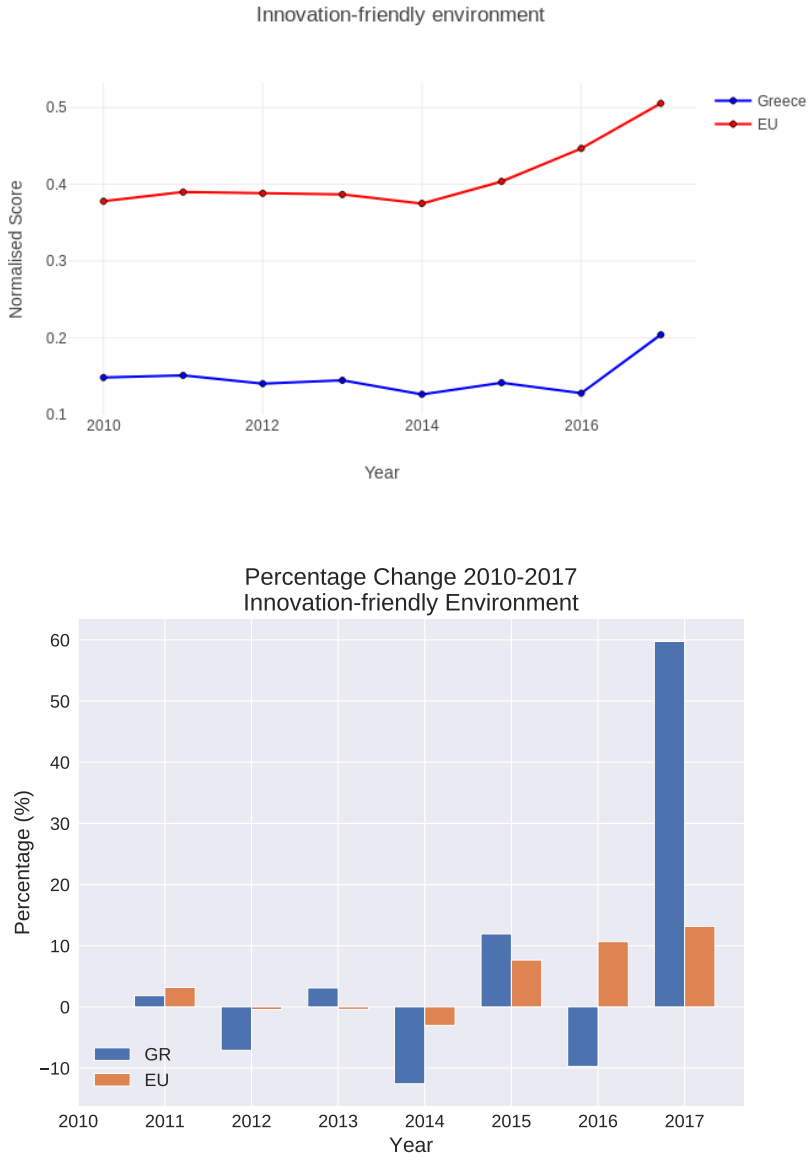

FIGURE A.4: Innovation-friendly environment (on the top) and percentage increase (at the bottom).

Definition 5 *Finance and support measures the availability of finance for innovation projects and the support of governments for research and innovation activities.*

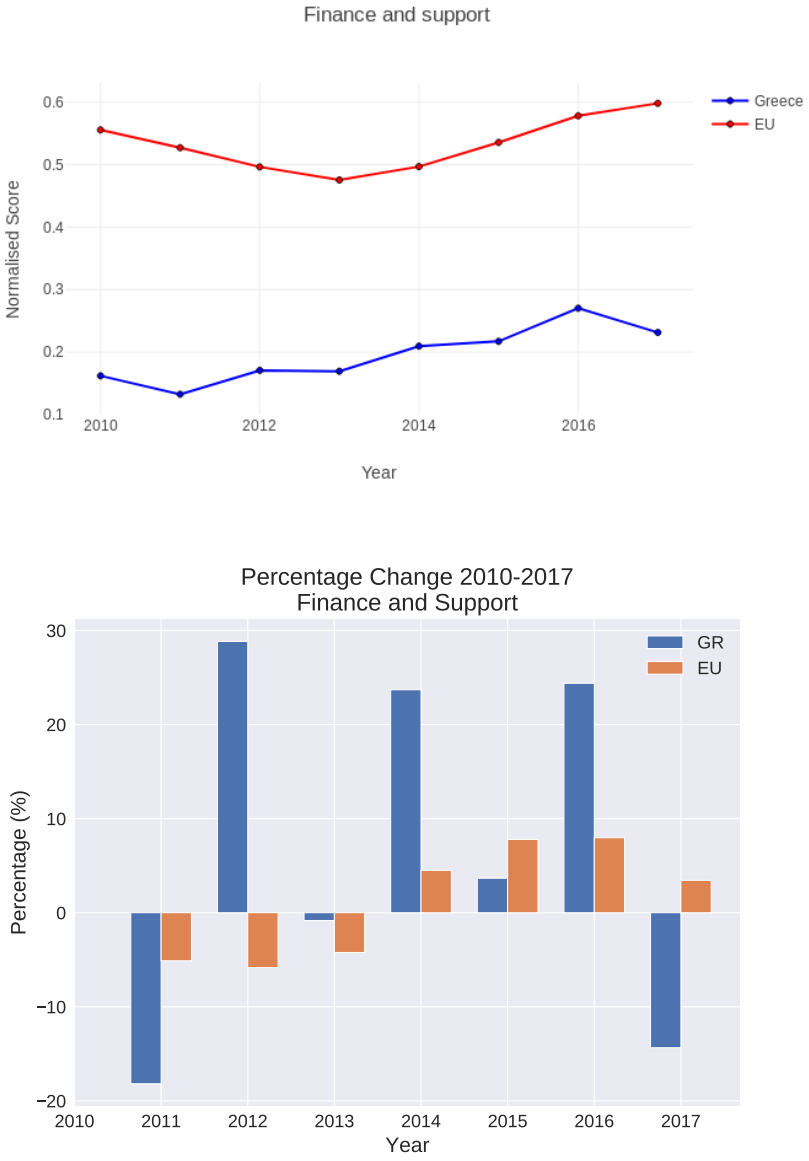

FIGURE A.5: Finance and support (on the top) and percentage increase (at the bottom).

Definition 6 *Firm investments include both R&D and non-R&D investments that firms make to generate innovations, and the efforts enterprises make to upgrade the ICT skills of their personnel.*

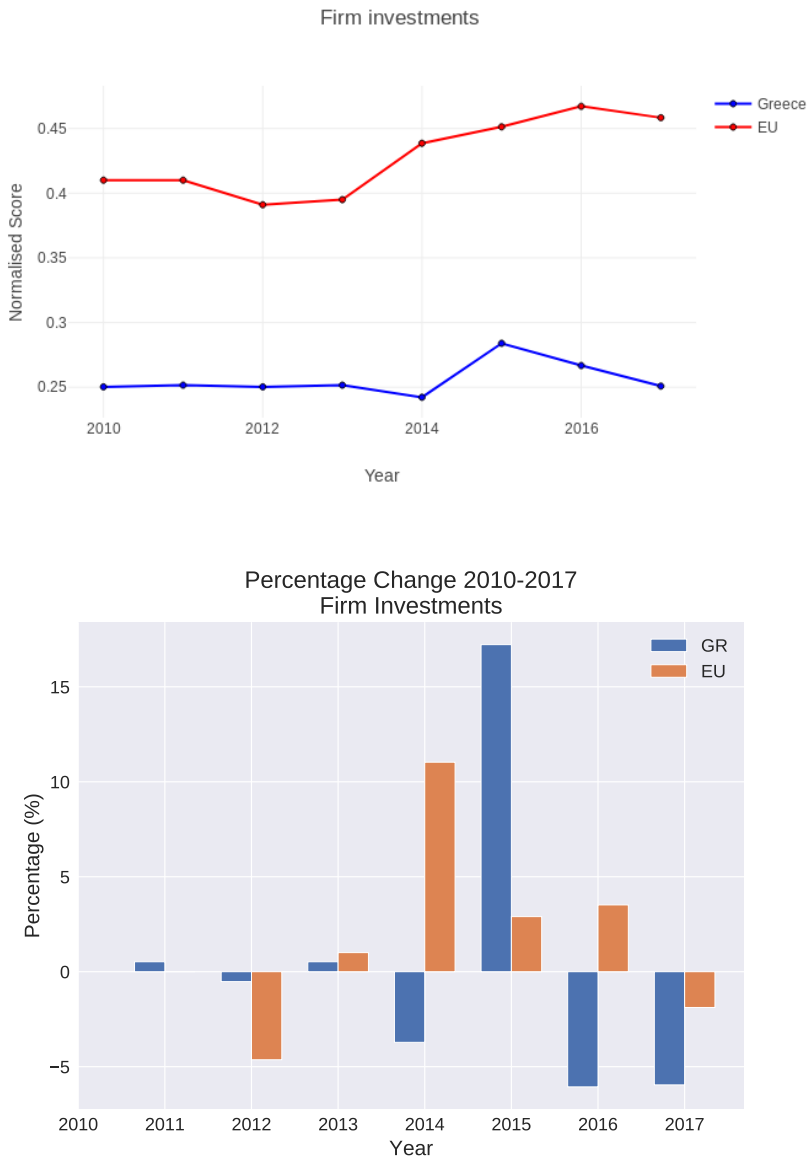

FIGURE A.6: Firm Investments (on the top) and percentage increase (at the bottom).

Definition 7 *Innovators composite indicator measures the share of firms that have introduced innovations onto the market or within their organizations, covering both product and process innovators, marketing and organizational innovators, and SMEs that innovate in-house.*

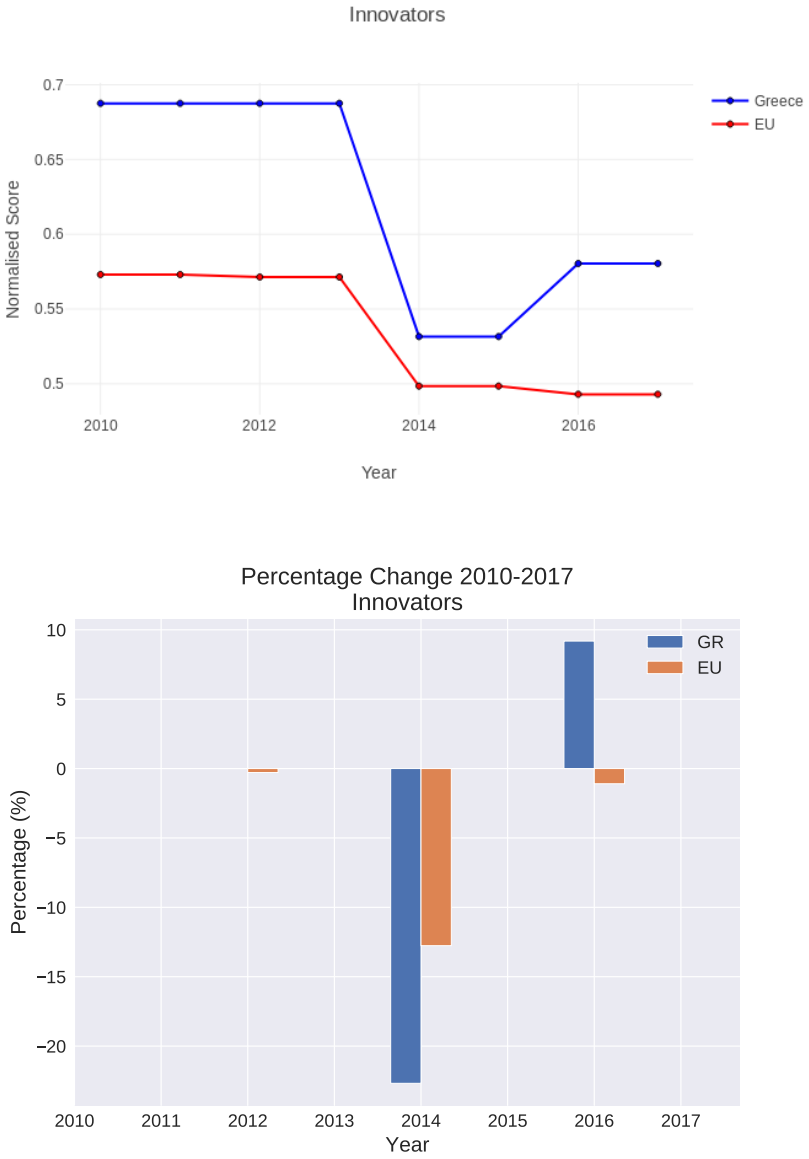

FIGURE A.7: Innovators (on the top) and percentage increase (at the bottom).

Definition 8 *Linkages composite indicator measures the innovation capabilities by looking at collaboration efforts between innovating firms, research collaboration between the private and public sector, and the extent to which the private sector finances public R&D activities.*

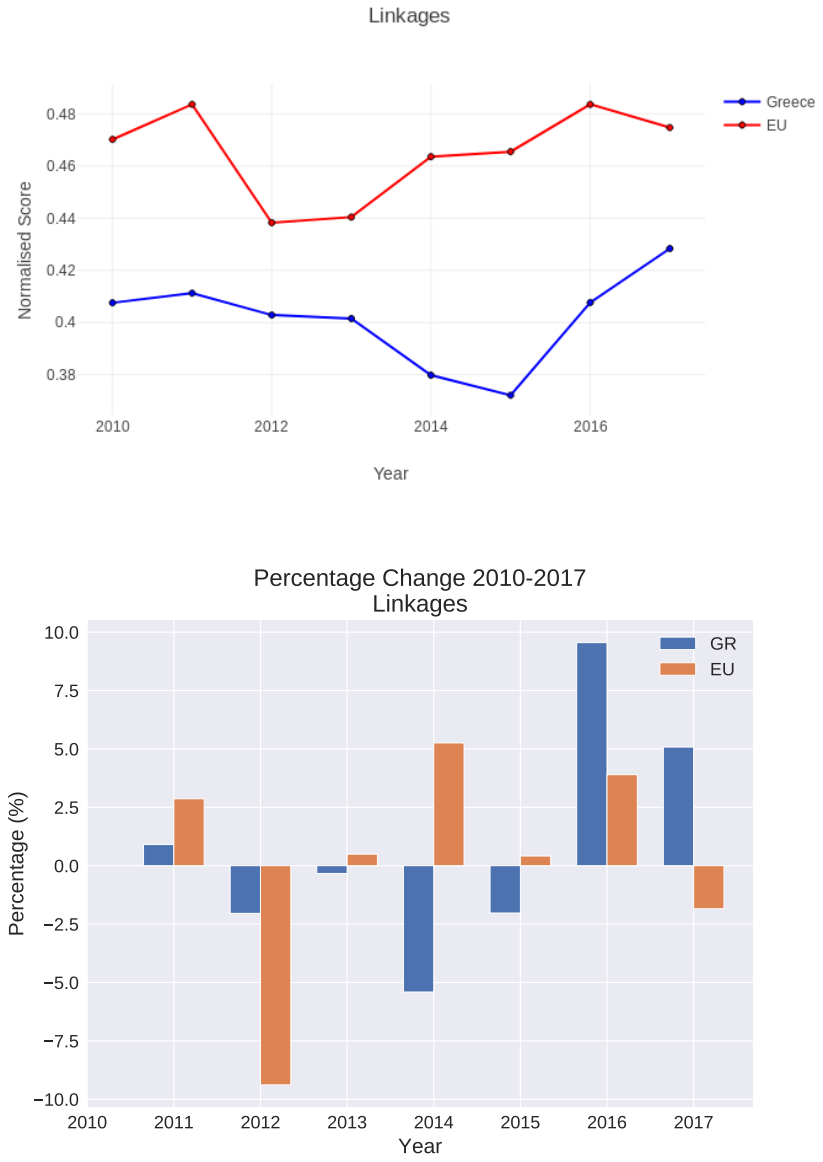

FIGURE A.8: Linkages (on the top) and percentage increase (at the bottom).

Definition 9 *Intellectual assets captures different forms of Intellectual Property Rights (IPR) generated in the innovation process.*

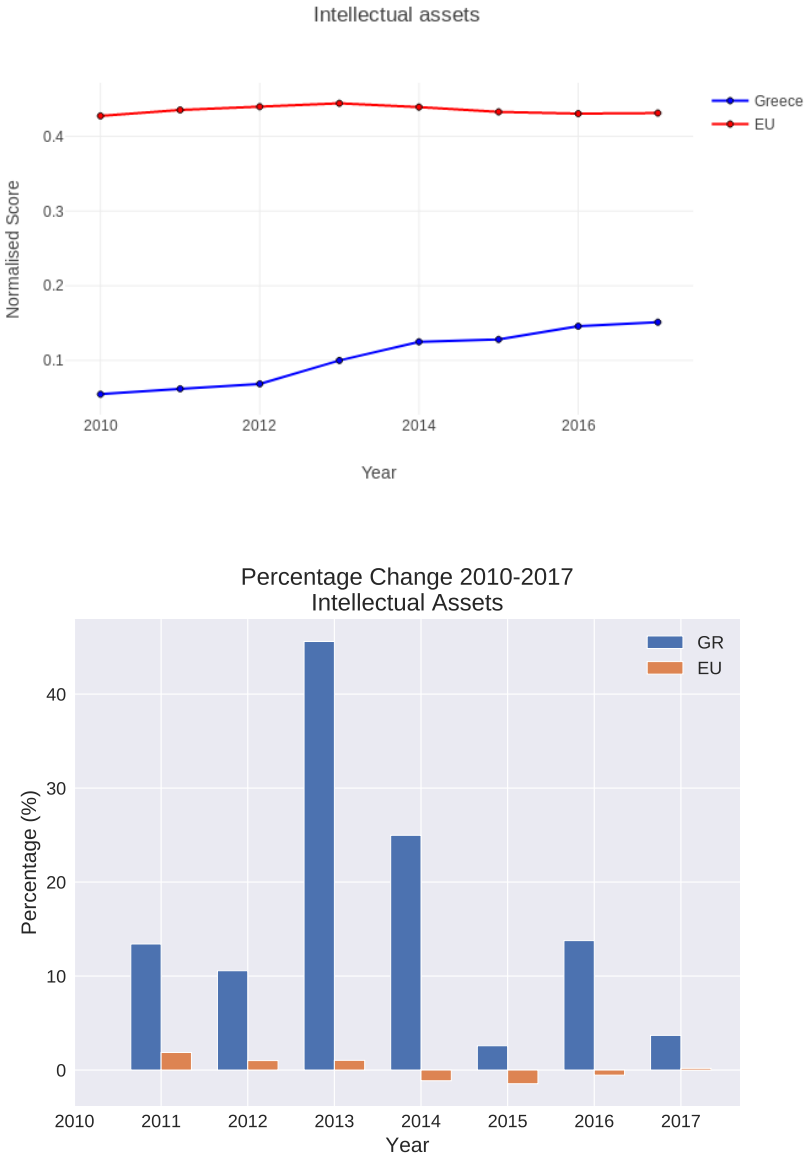

FIGURE A.9: Intellectual assets (on the top) and percentage increase (at the bottom).

Definition 10 *Employment impacts measures the impact of innovation on employment.*

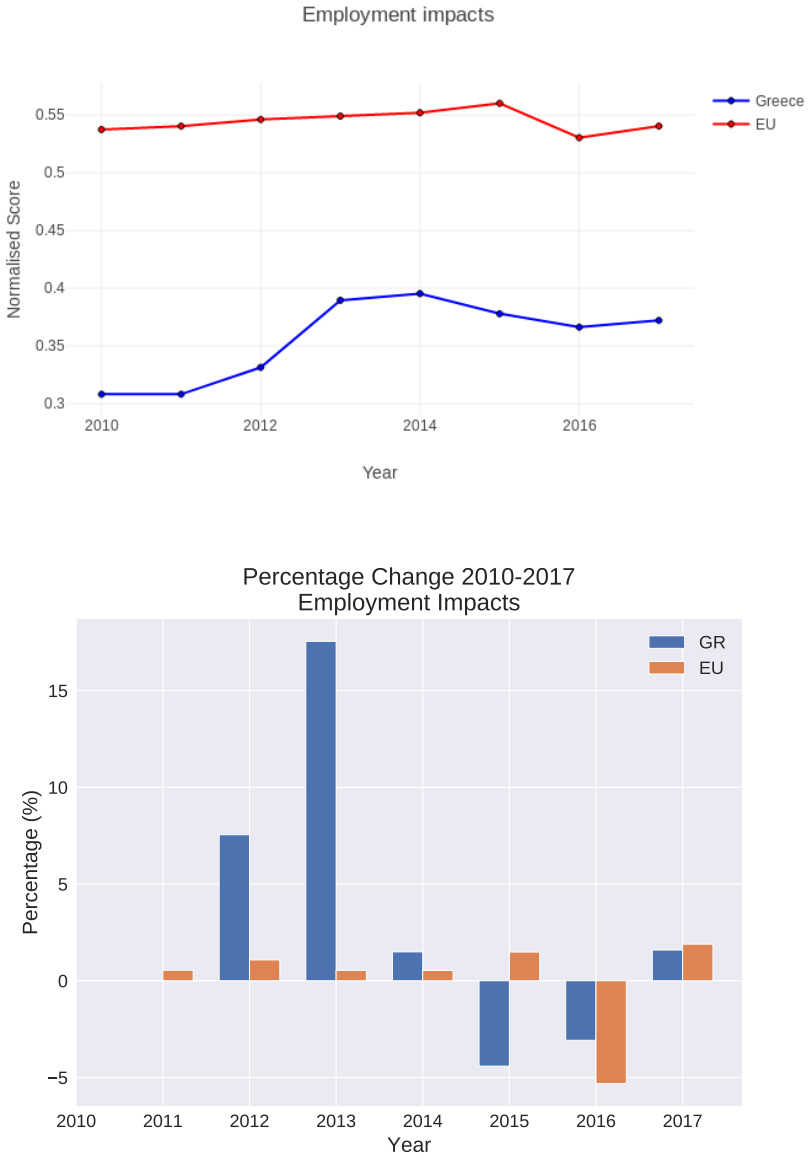

FIGURE A.10: Employment impacts (on the top) and percentage increase (at the bottom).

Definition 11 *Sales impacts measures the economic impact of innovation.*

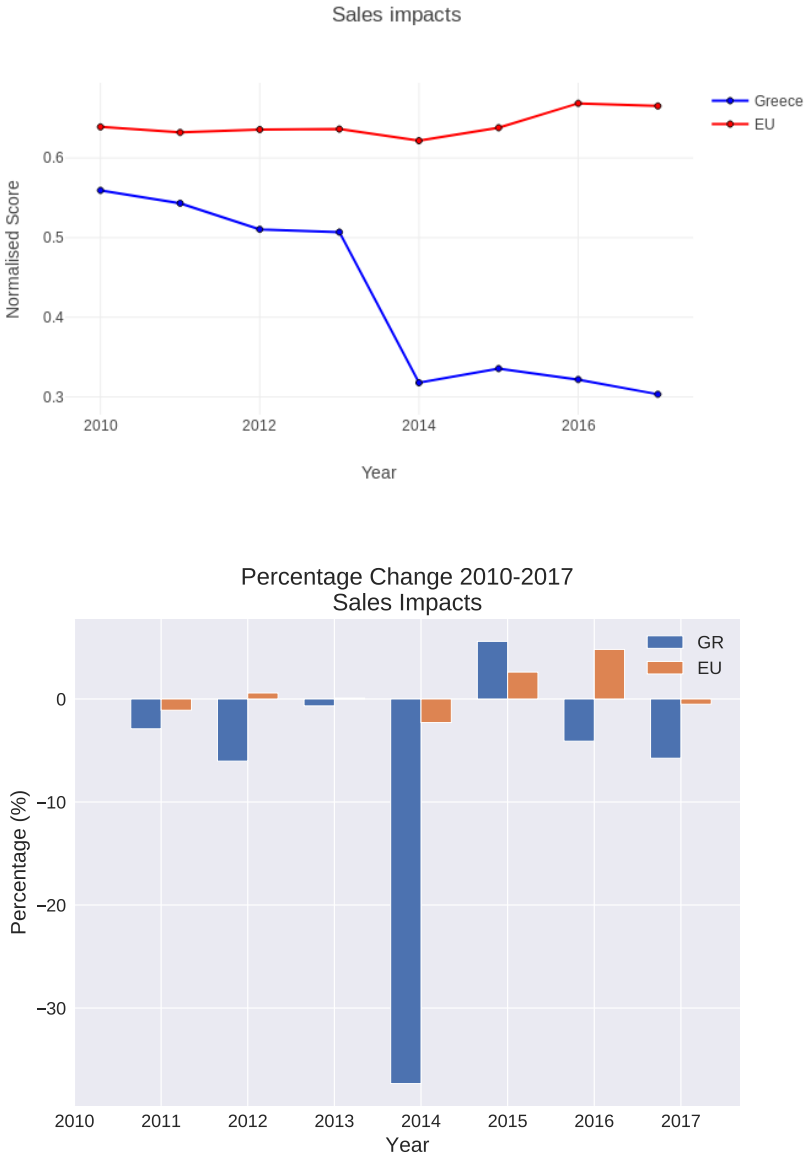

FIGURE A.11: Sales impacts (on the top) and percentage increase (at the bottom).

A.2 Indicators Charts

Definition 12 *"Broadband penetration" indicator has the number of enterprises with a maximum contracted download speed of the fastest fixed internet connection of at least 100 Mb/s and total number of enterprises as numerator and denominator respectively.*

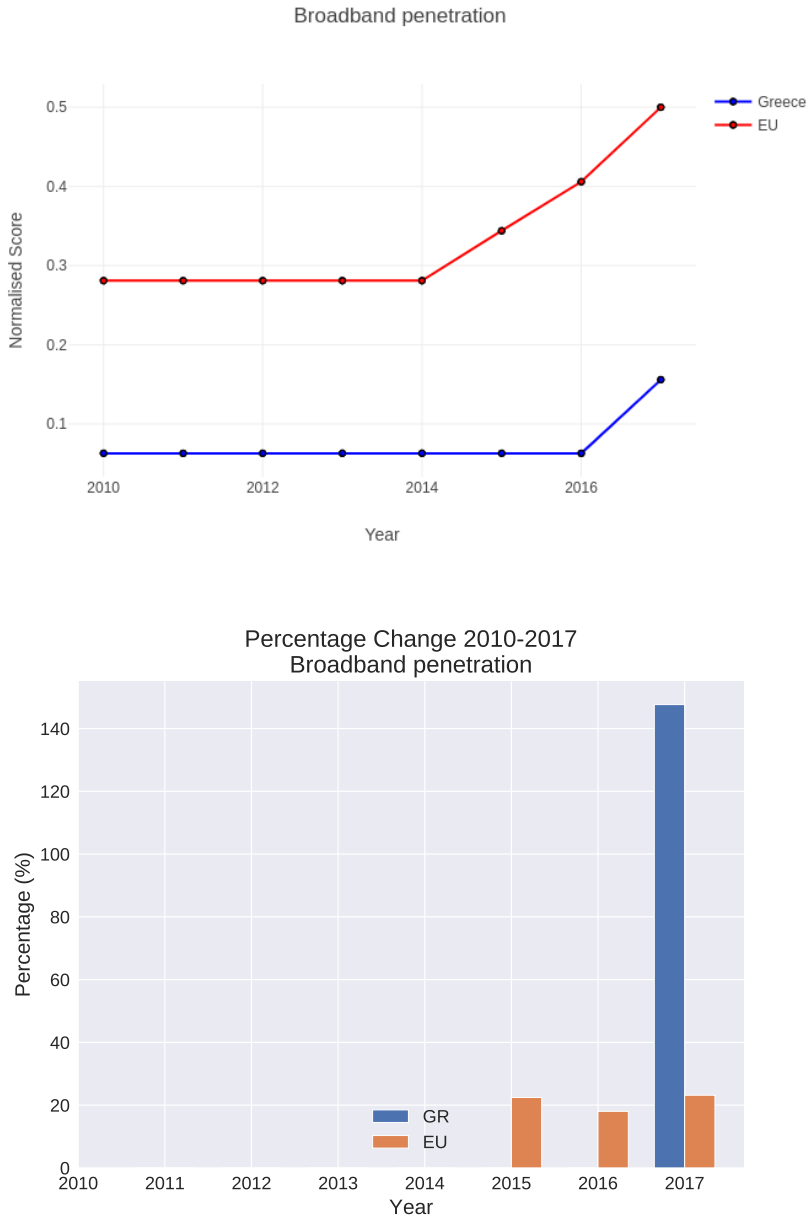

FIGURE A.12: Broadband penetration (on the top) and percentage increase (at the bottom).

Definition 13 *"Design applications per billion GDP (in PPS)" is an indicator which shows the number of individual designs applied for at European Union Intellectual Property Office divided by GDP in Product in Purchasing Power Standard.*

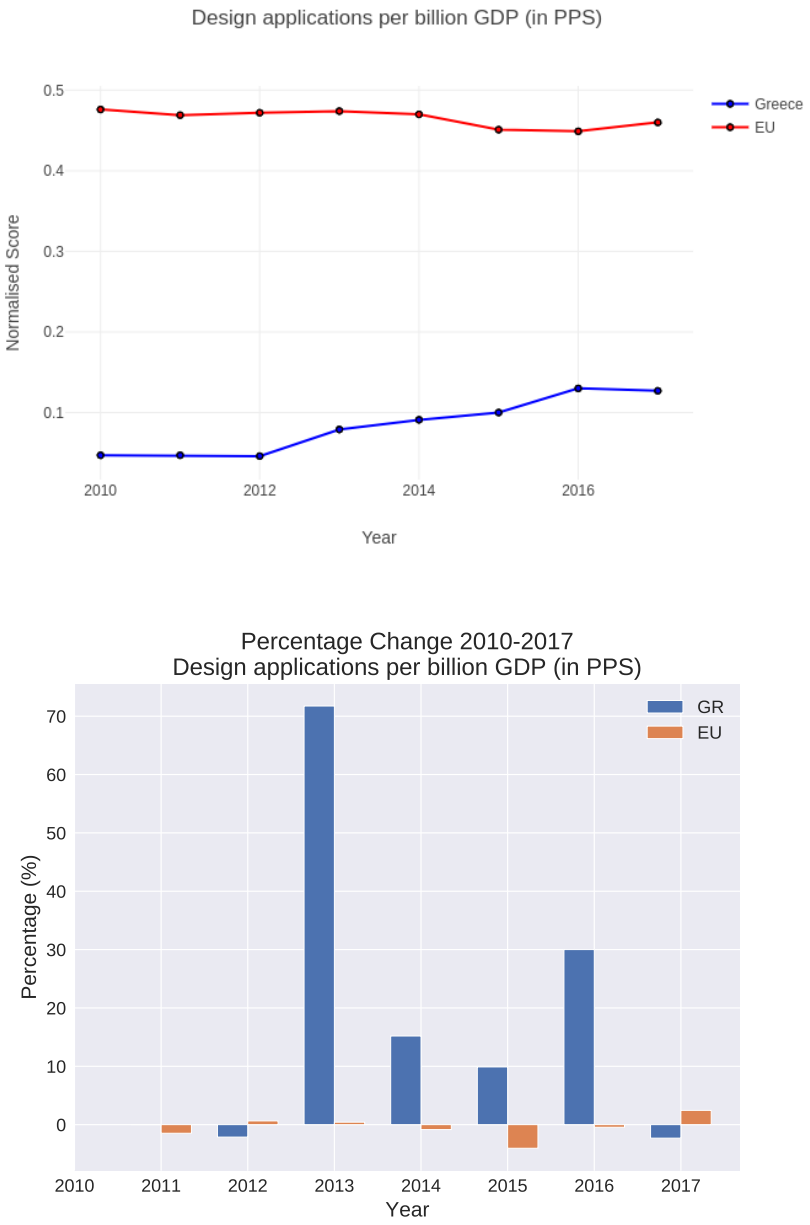

FIGURE A.13: Design applications per billion GDP (in PPS) (on the top) and percentage increase (at the bottom).

Definition 14 *"Employment in knowledge-intensive activities (% of total employment)" is an indicator which shows the number of employed persons in knowledge-intensive activities in business industries divided by the total employment.*

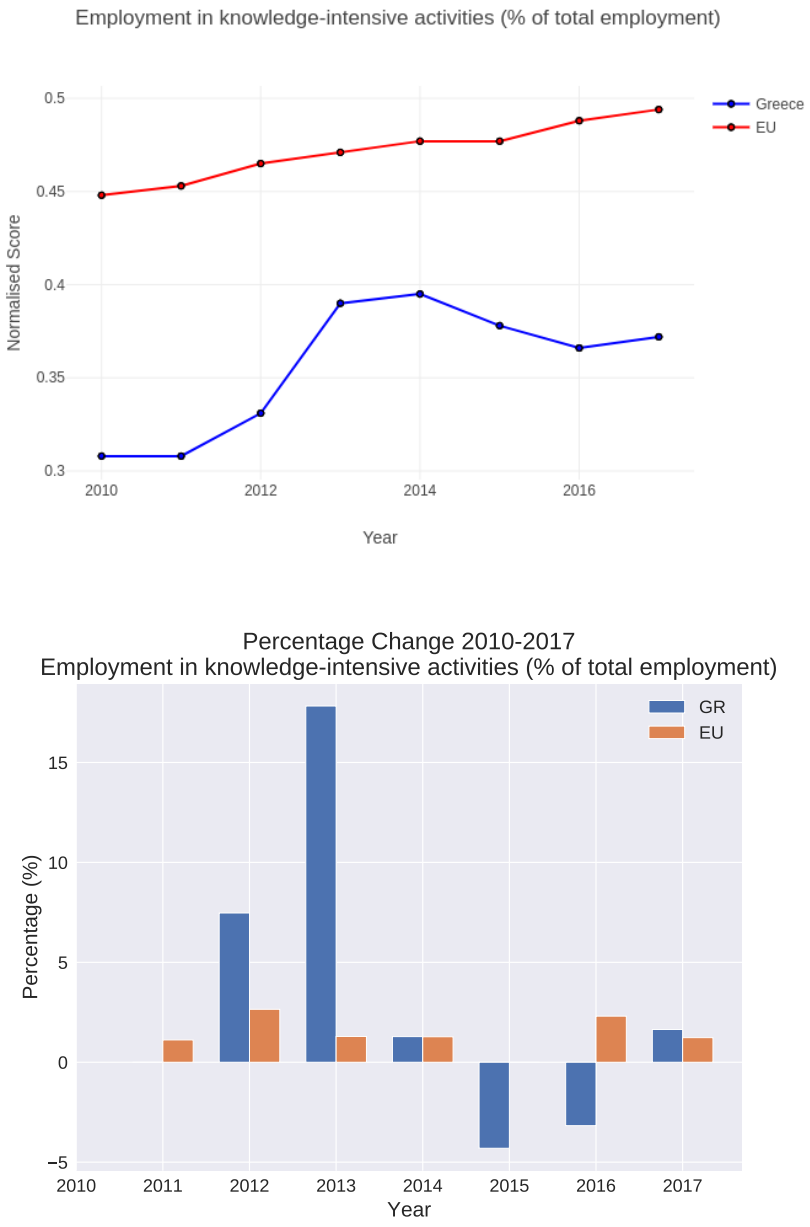

FIGURE A.14: Employment in knowledge-intensive activities (% of total employment) (on the top) and percentage increase (at the bottom).

Definition 15 *"Enterprises providing training to develop or upgrade ICT skills of their personnel" is an indicator which is a fraction of the number of enterprises that provided any type of training to develop Information Communication Technology (ICT) related skills of their personnel divided by the total number of enterprises.*

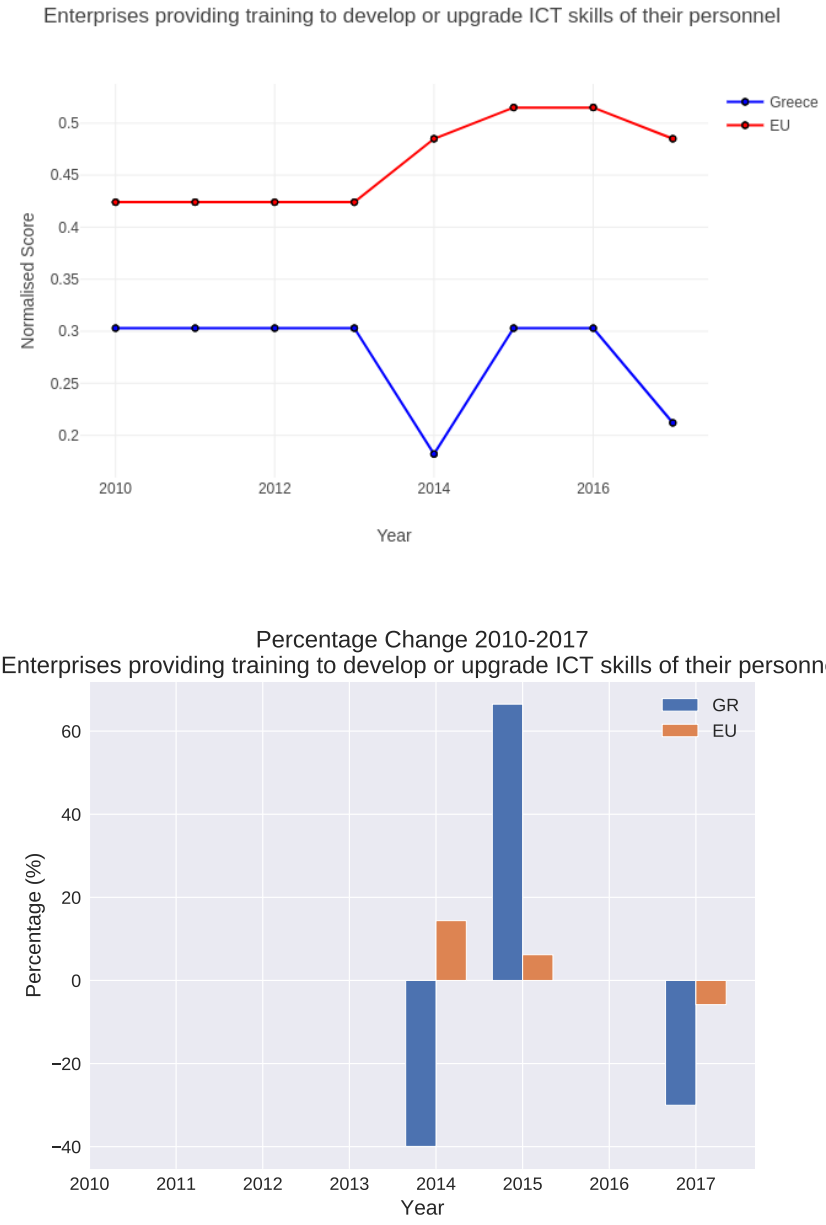

FIGURE A.15: Enterprises providing training to develop or upgrade ICT skills of their personnel (on the top) and percentage increase (at the bottom).

Definition 16 *"Exports of medium and high technology products as a share of total product exports" is the value of medium and high tech exports, in national currency and current prices, including exports, divided by the value of total product exports.*

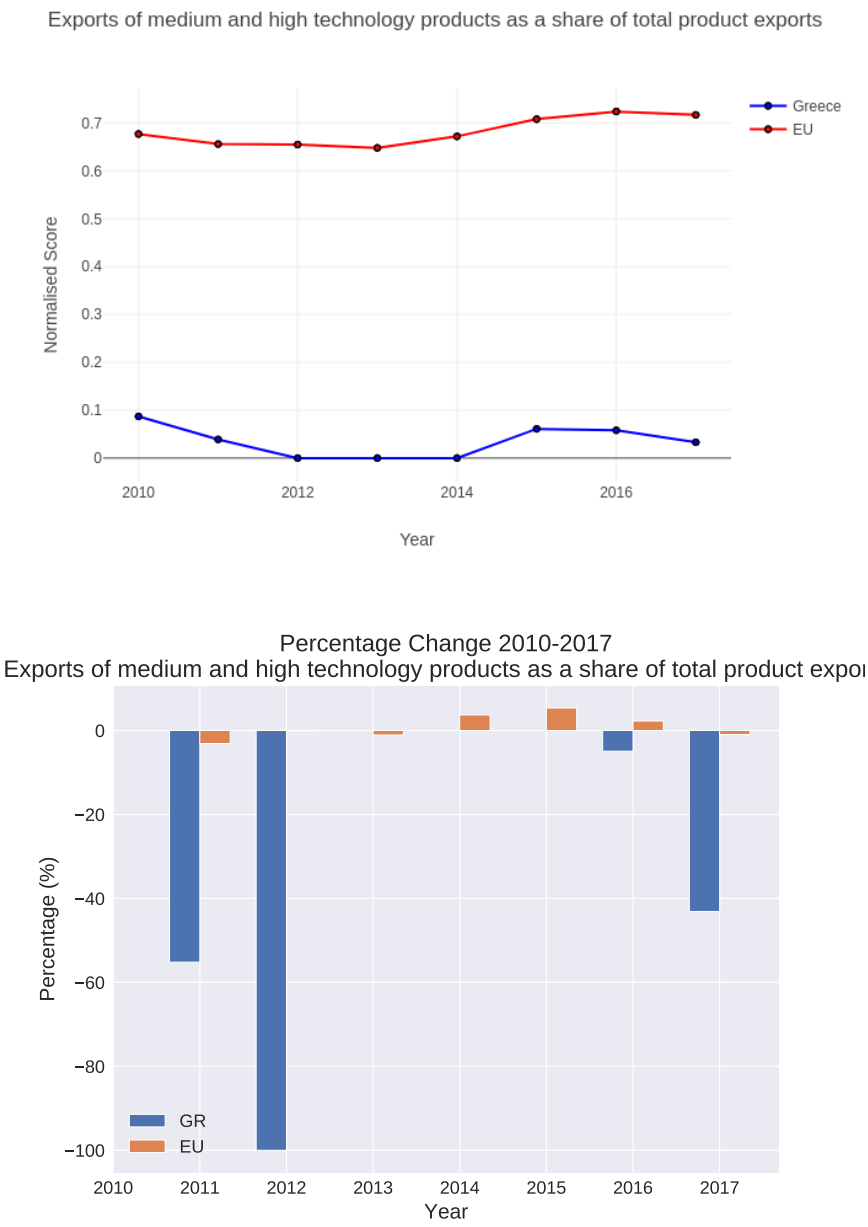

FIGURE A.16: Exports of medium and high technology products as a share of total product exports (on the top) and percentage increase (at the bottom).

Definition 17 *"Innovative SMEs collaborating with others (% of SMEs)" is an indicator which shows the number of SMEs with innovation co-operation activities divided by the total number of SMEs.*

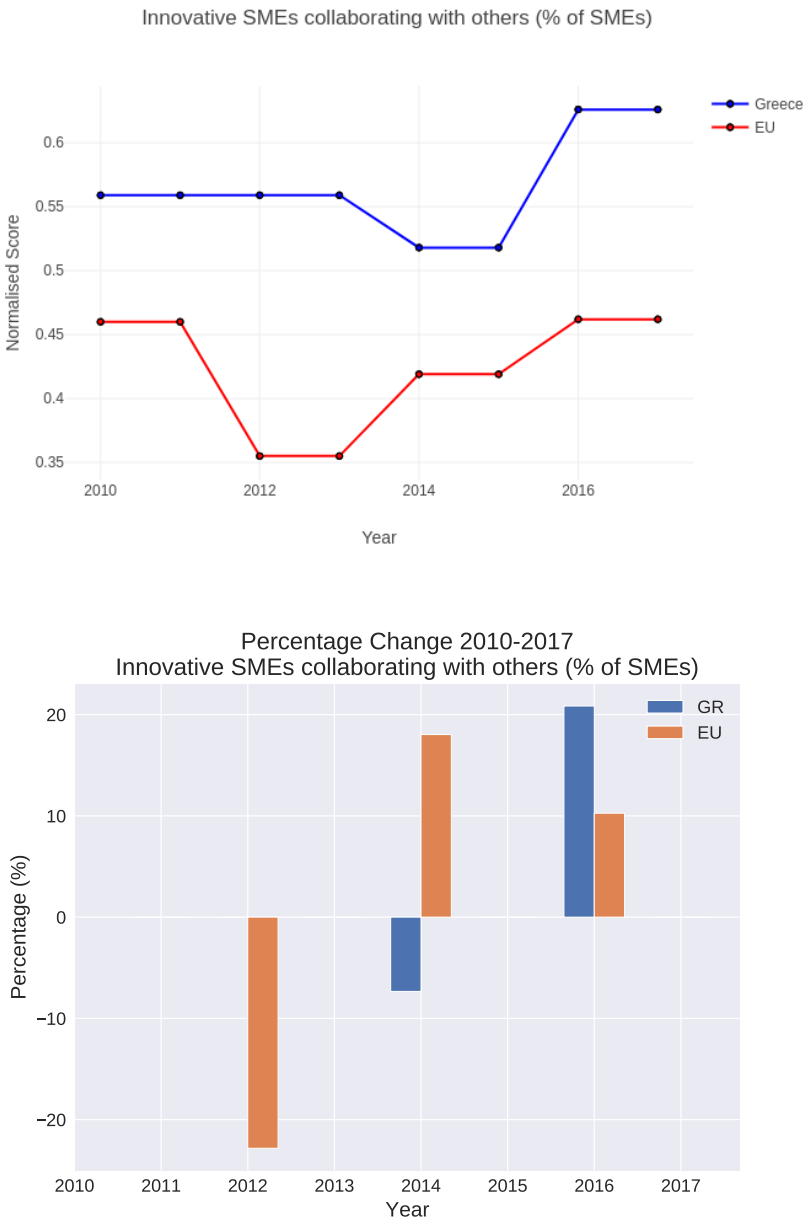

FIGURE A.17: Innovative SMEs collaborating with others (% of total employment) (on the top) and percentage increase (at the bottom).

Definition 18 *"International scientific co-publications per million population" is a fraction which has as numerator the number of scientific publications with at least one co-author based abroad (where abroad is non-EU for the EU28) and denominator the total population.*

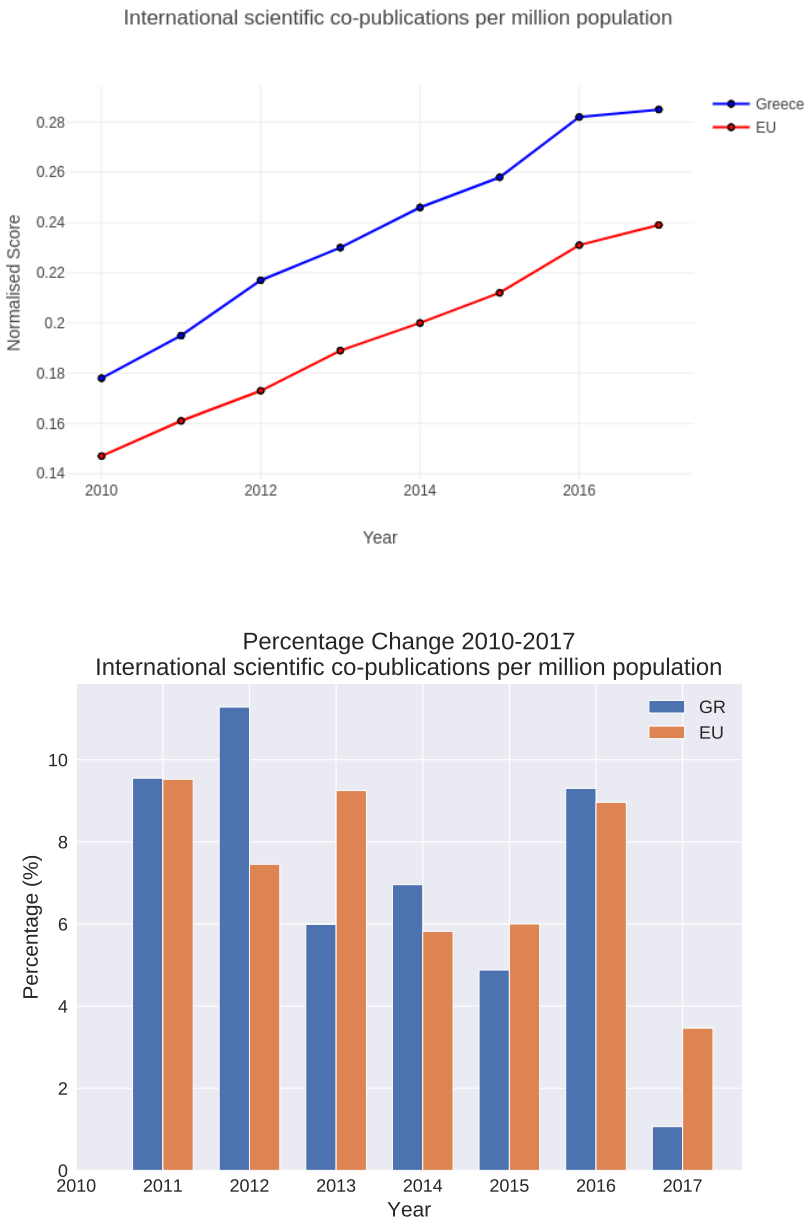

FIGURE A.18: International scientific co-publications per million population (on the top) and percentage increase (at the bottom).

Definition 19 *"Knowledge-intensive services exports as % of total services exports" is a fraction which has as numerator the exports of knowledge-intensive services and denominator the total value of services exports.*

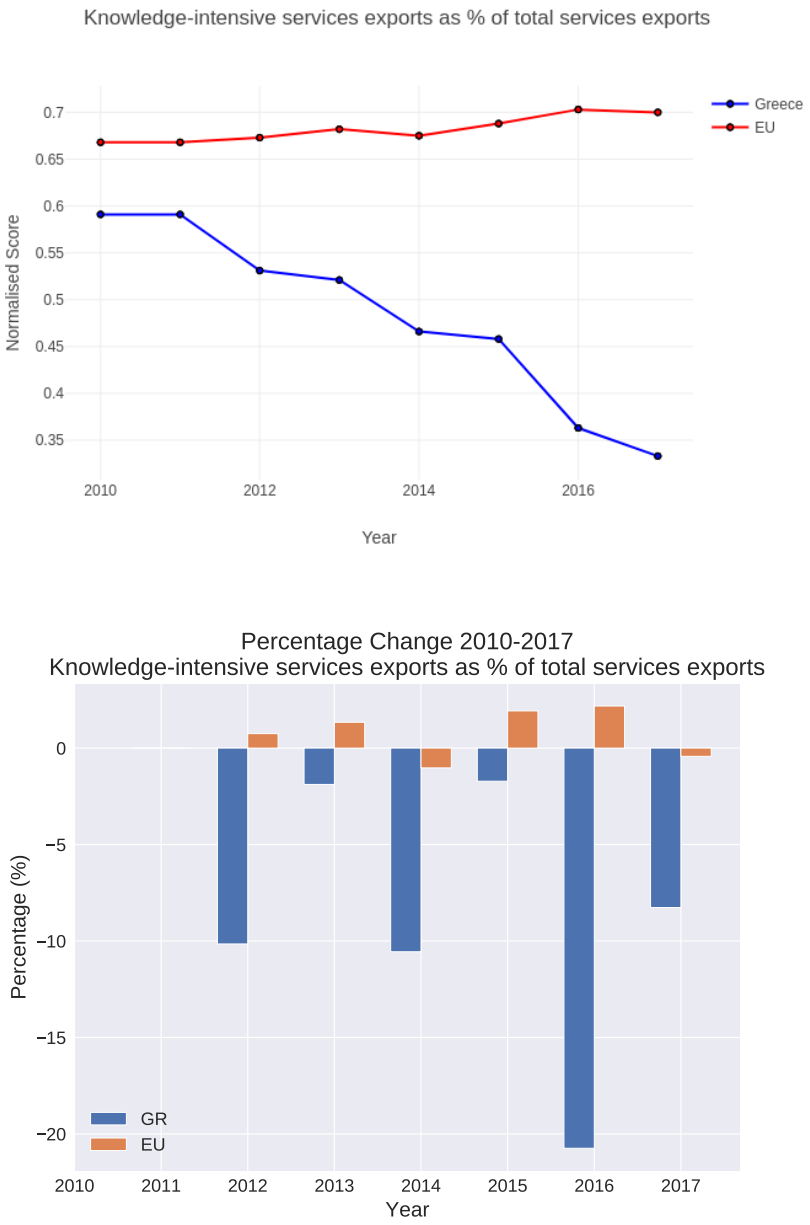

FIGURE A.19: Knowledge-intensive services exports as % of total services exports (on the top) and percentage increase (at the bottom).

Definition 20 *"New doctorate graduates per 1000 population aged 25-34" is a fraction which has as numerator the number of doctorate graduates and denominator the population between and including 25 and 34 years.*

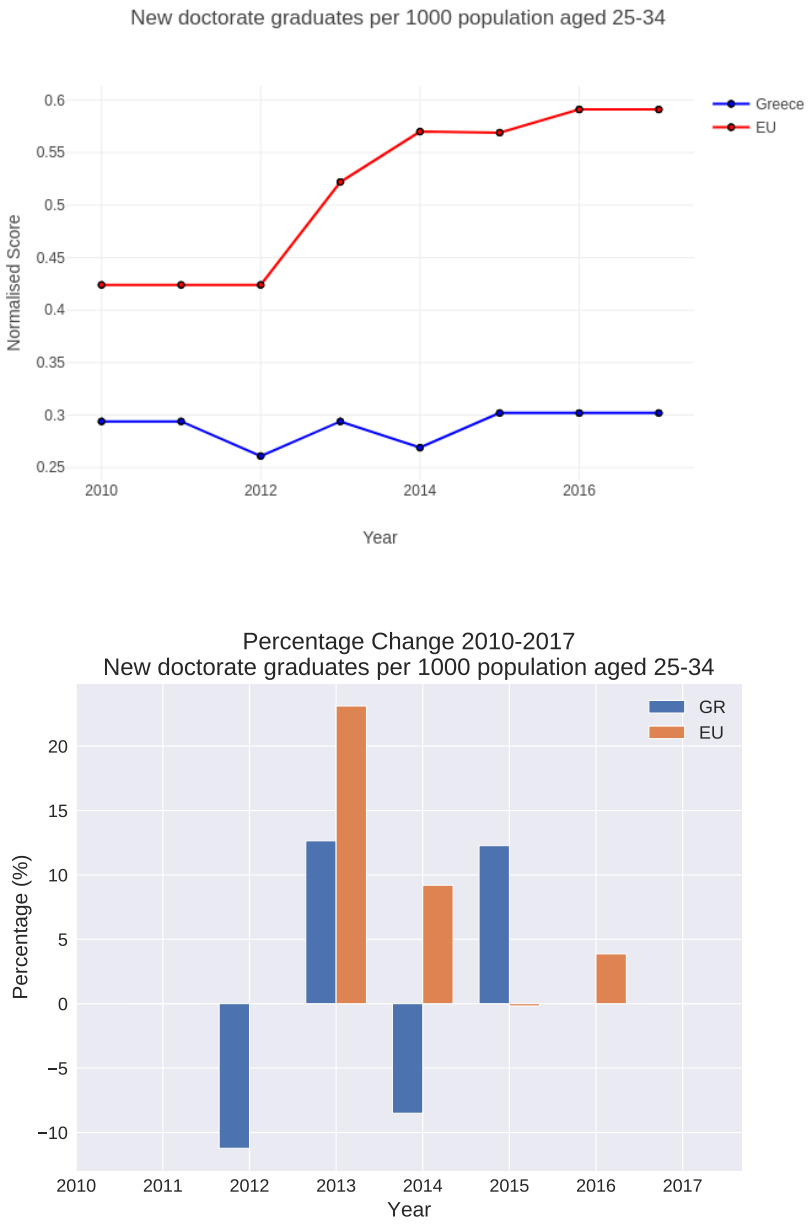

FIGURE A.20: New doctorate graduates per 1000 population aged 25-34 (on the top) and percentage increase (at the bottom).

Definition 21 *"Non-R&D innovation expenditures (% of turnover)" is a fraction of the sum of total innovation expenditure for enterprises, excluding intramural and extramural R&D expenditures divided by total turnover for all enterprises.*

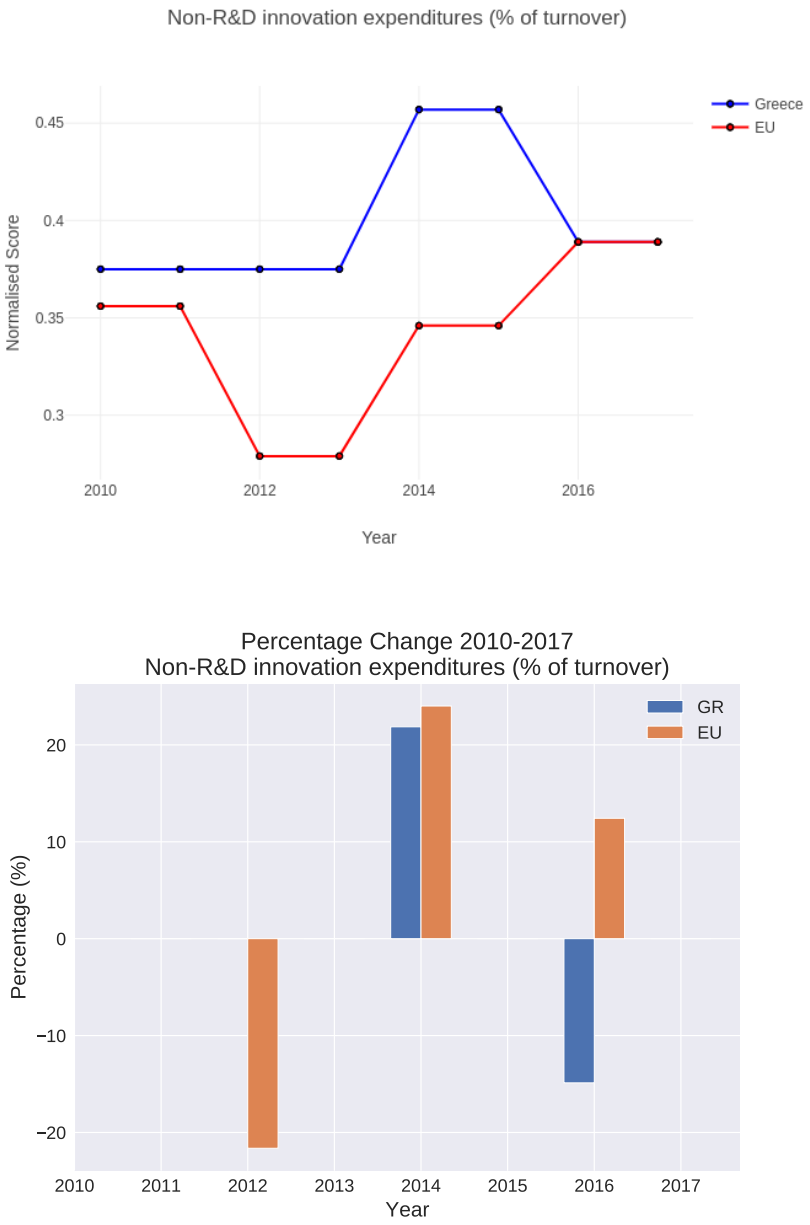

FIGURE A.21: Non-R&D innovation expenditures (% of turnover) (on the top) and percentage increase (at the bottom).

Definition 22 *"Opportunity-driven entrepreneurship (Motivation Index)" is calculated as the ratio between the share of persons involved in improvement-driven entrepreneurship and the share of persons involved in necessity-driven entrepreneurship.*

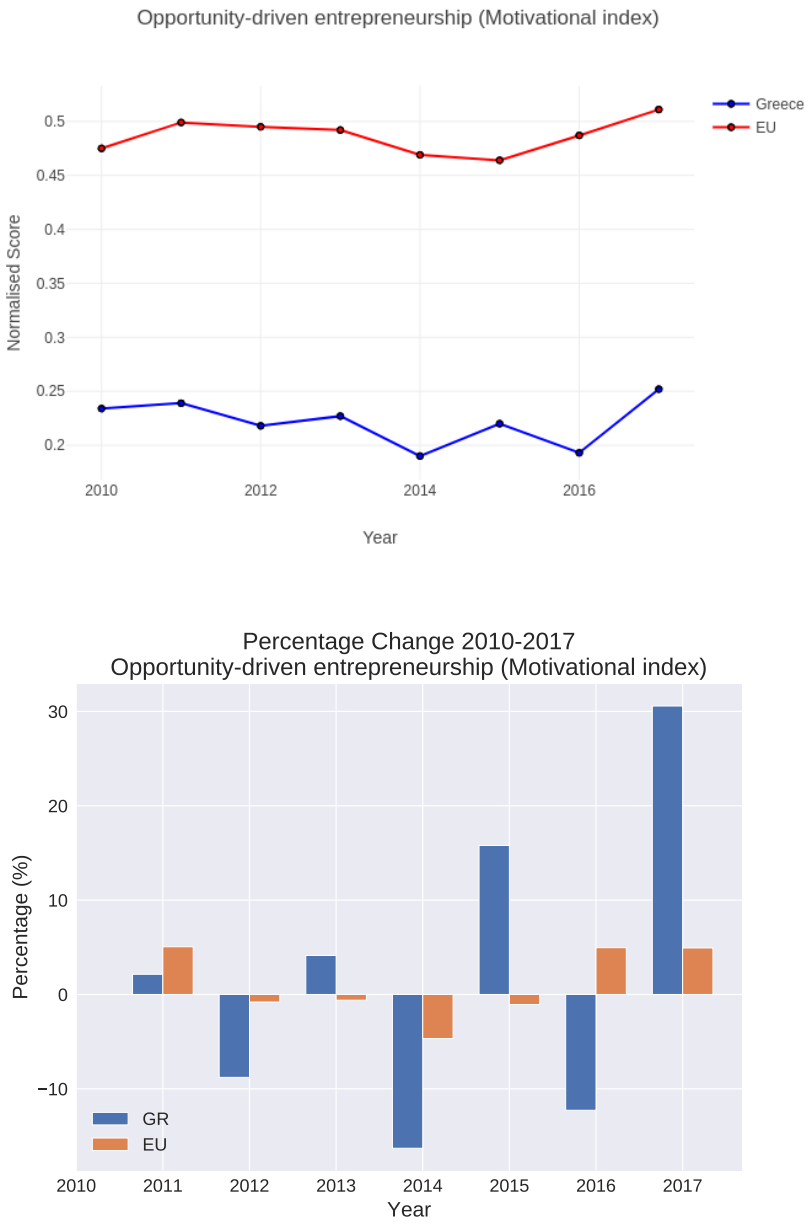

FIGURE A.22: Opportunity-driven entrepreneurship (Motivation Index) (on the top) and percentage increase (at the bottom).

Definition 23 *"PCT patent applications per billion GDP (in PPS)" is a fraction which has as numerator the number of patent applications filed under the Patent Cooperation Treaty (PCT), at international phase, designating the European Patent Office (EPO) and denominator the GDP in Purchasing Power Standard.*

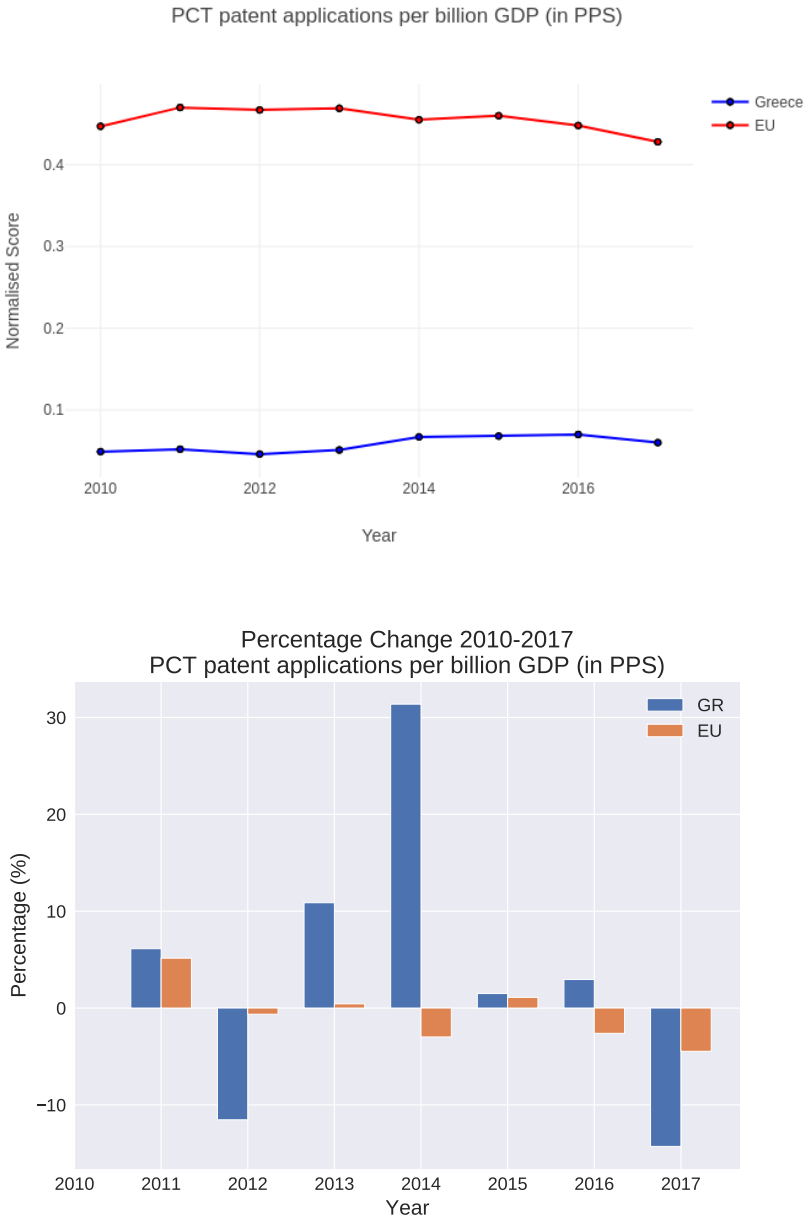

FIGURE A.23: PCT patent applications per billion GDP (in PPS) (on the top) and percentage increase (at the bottom).

Definition 24 *"Percentage population aged 25-34 having completed tertiary education" is a fraction which has as numerator the number of persons in age class with some form of post-secondary education and denominator the population between and including 25 and 34 years.*

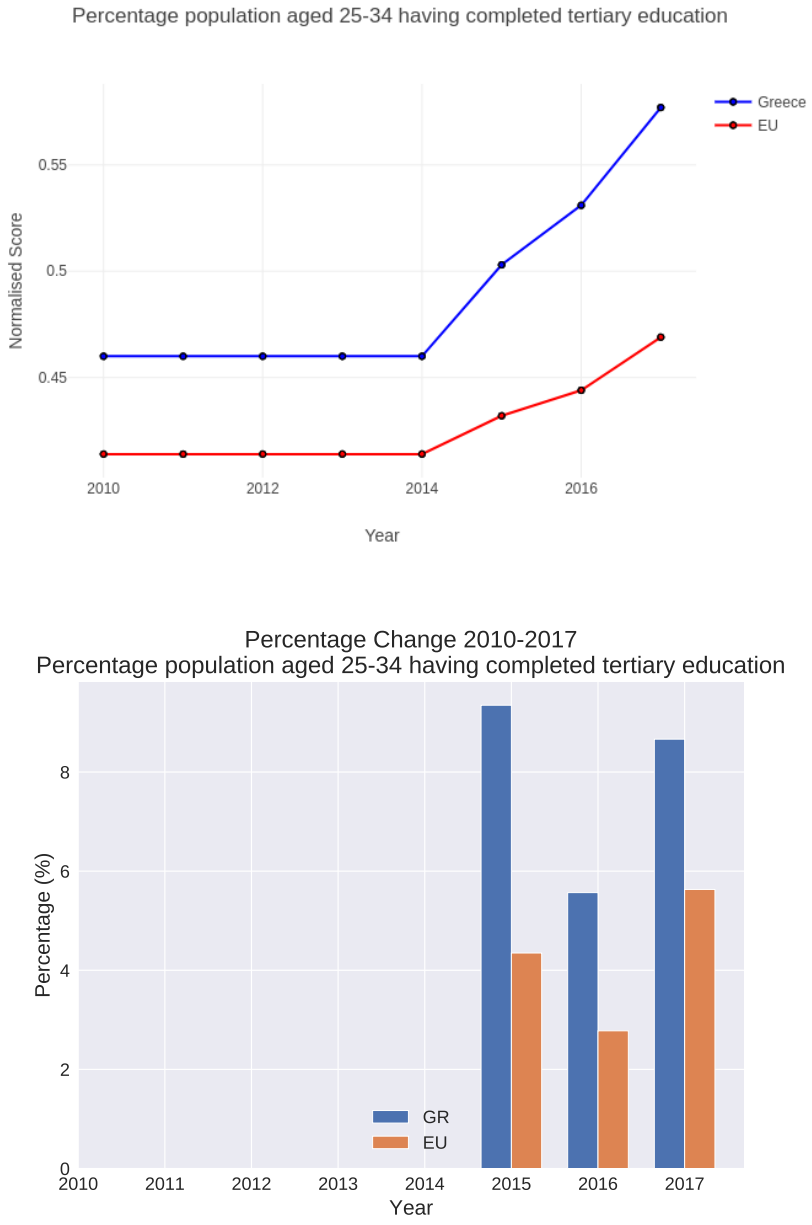

FIGURE A.24: Percentage population aged 25-34 having completed tertiary education (on the top) and percentage increase (at the bottom).

Definition 25 *"Percentage population aged 25-64 involved in lifelong learning" indicator is a fraction which has as numerator the target population for lifelong learning statistics referring to all persons in private households aged between 25 and 64 years and denominator the total population of the same age group, excluding those who did not answer the question concerning participation in (formal and non-formal) education and training.*

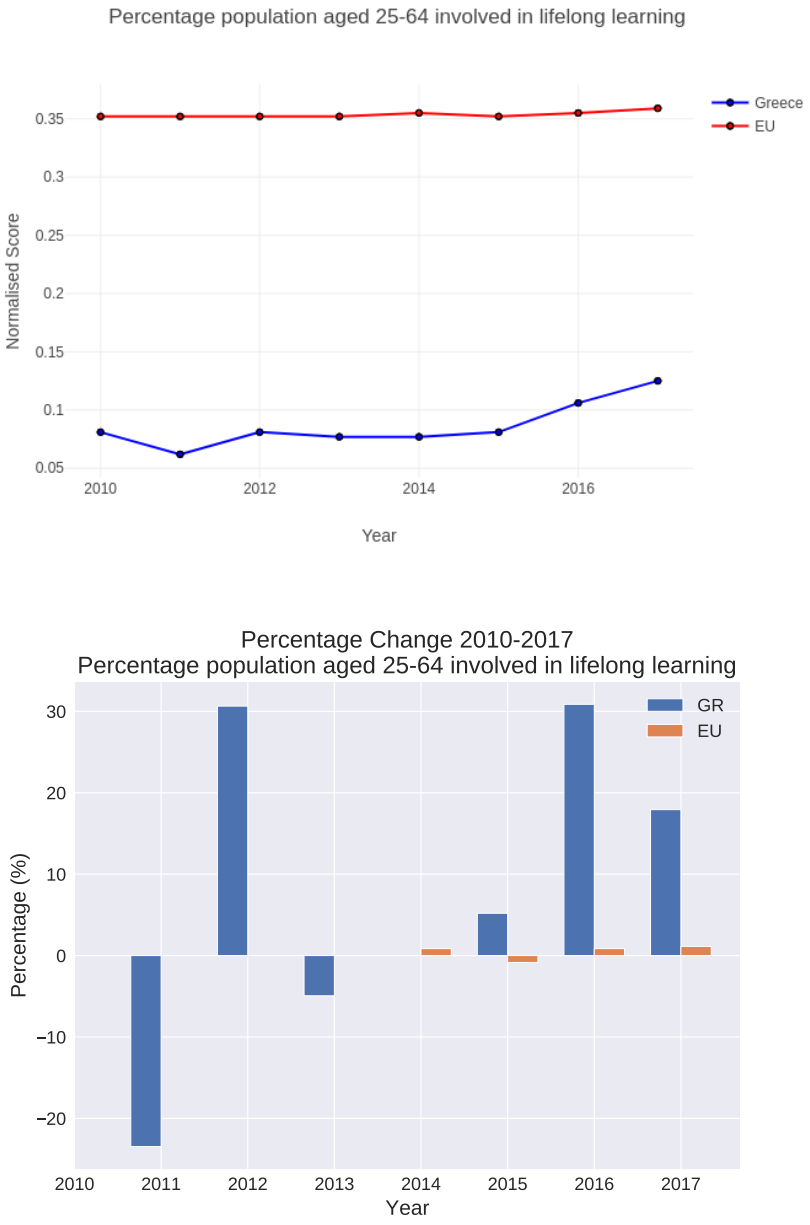

FIGURE A.25: Percentage population aged 25-64 involved in lifelong learning (on the top) and percentage increase (at the bottom).

Definition 26 *"Private co-funding of public R&D expenditures (percentage of GDP)" indicator represents all R&D expenditures in the government sector and the higher education sector financed by the business sector divided by Gross Domestic Product (GDP).*

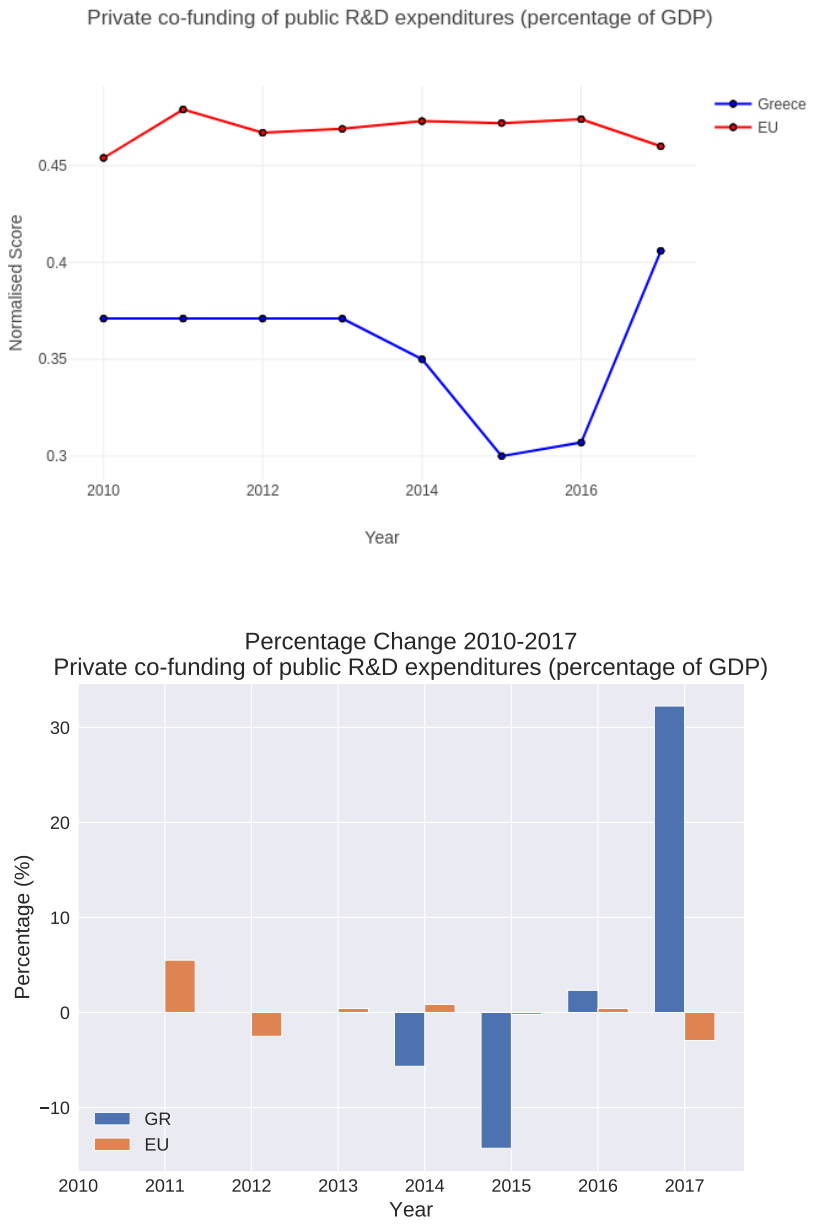

FIGURE A.26: Private co-funding of public R&D expenditures (percentage of GDP) (on the top) and percentage increase (at the bottom).

Definition 27 *"Public-private co-publications per million population" indicator is the number of public-private co-authored research publications divided by total population.*

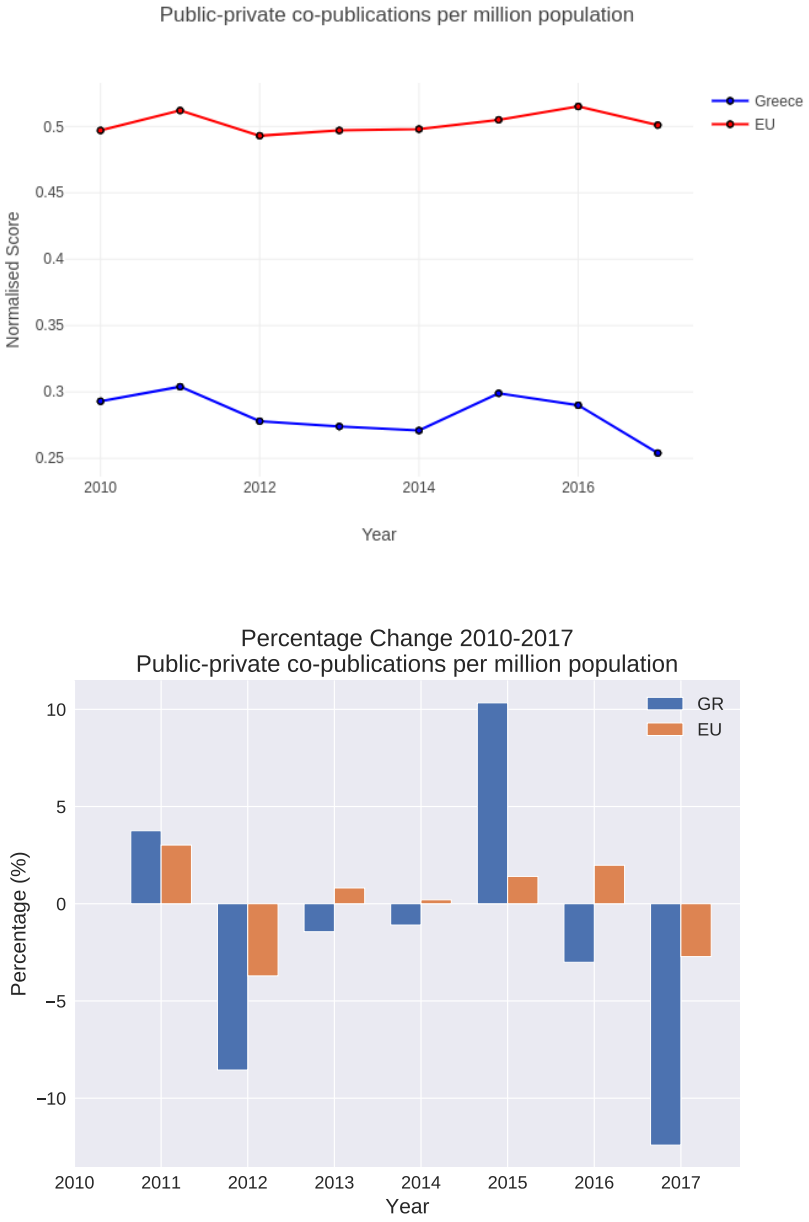

FIGURE A.27: Public-private co-publications per million population (on the top) and percentage increase (at the bottom).

Definition 28 *"R&D expenditure in the business sector (% of GDP)" indicator has all R&D expenditures in the business sector and GDP as nominator and denominator respectively.*

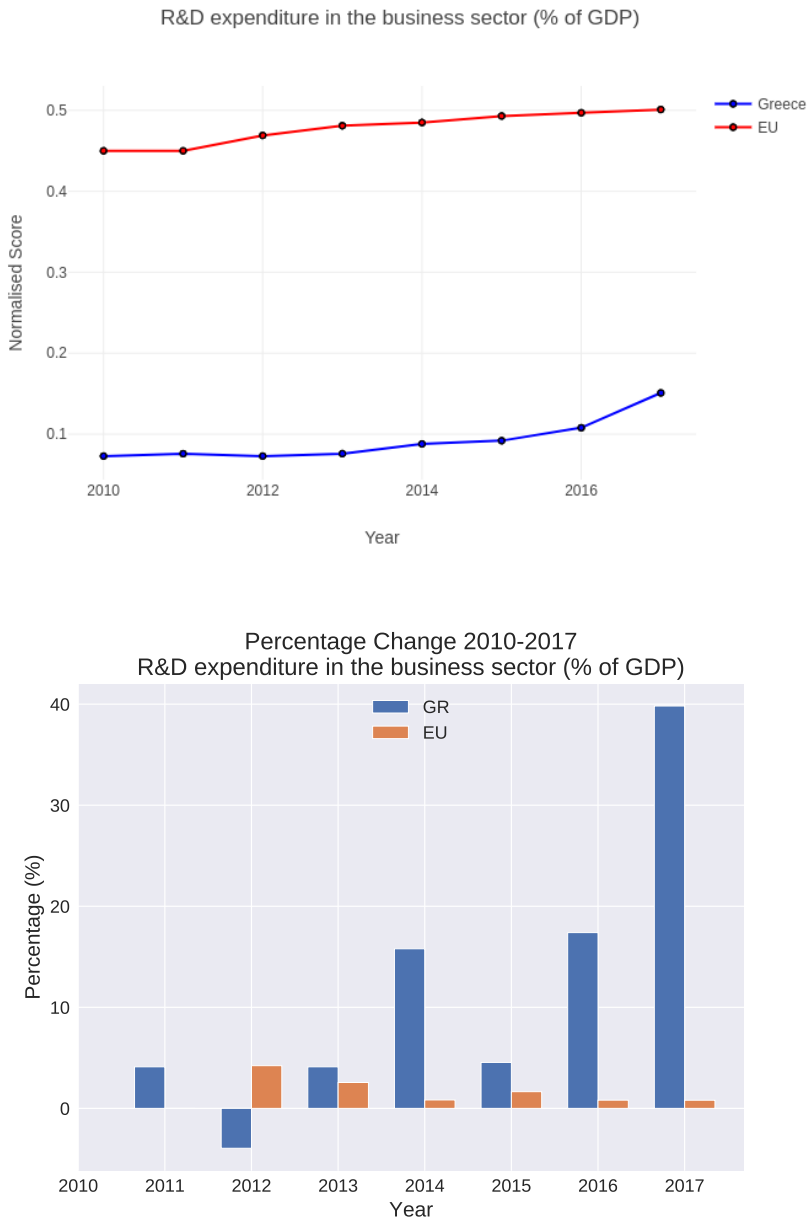

FIGURE A.28: R&D expenditure in the business sector (% of GDP) (on the top) and percentage increase (at the bottom).

Definition 29 *"R&D expenditure in the public sector (% of GDP)" indicator is a fraction which has as numerator all R&D expenditures in the government sector and the higher education sector and denominator the Gross Domestic Product (GDP).*

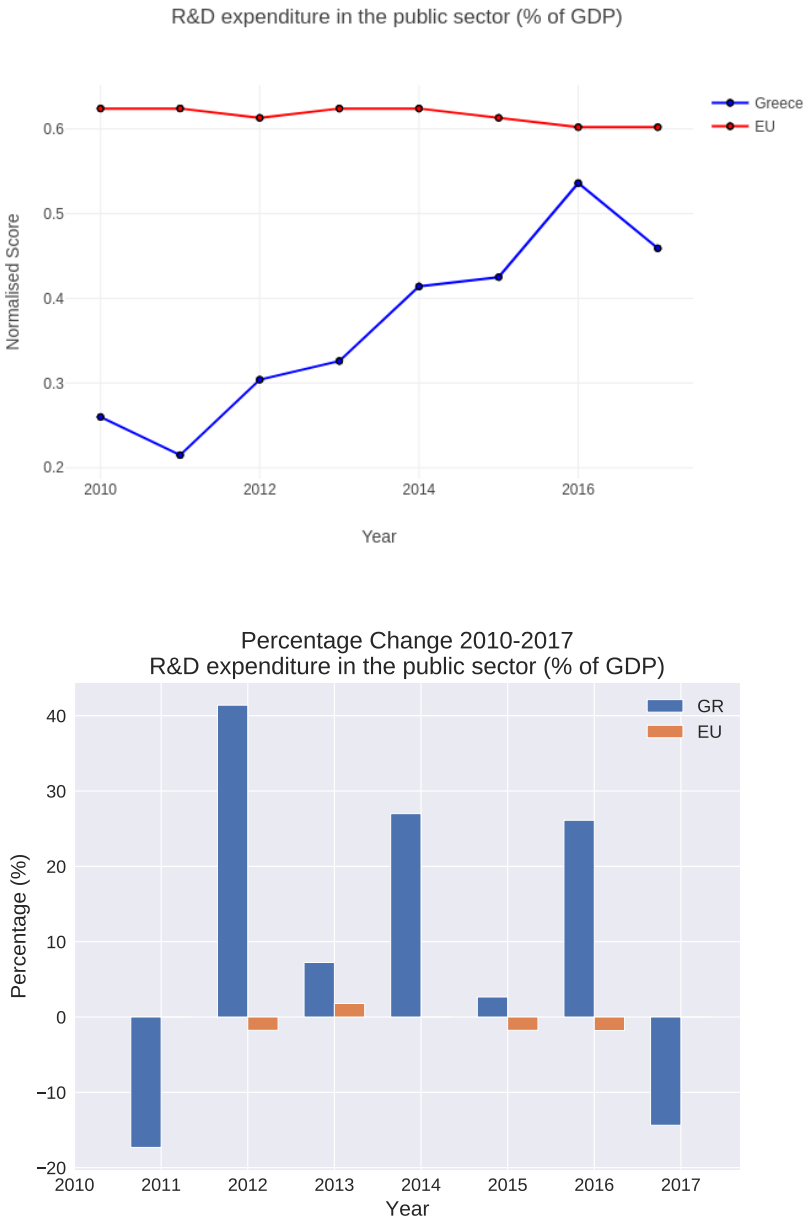

FIGURE A.29: R&D expenditure in the public sector (% of GDP) (on the top) and percentage increase (at the bottom).

Definition 30 *"Sales of new-to-market and new-to-firm innovations as % of turnover" indicator is the sum of total turnover of new or significantly improved products, either new-to-the-firm or new-to-the-market, for all enterprises divided by total turnover for all enterprises.*

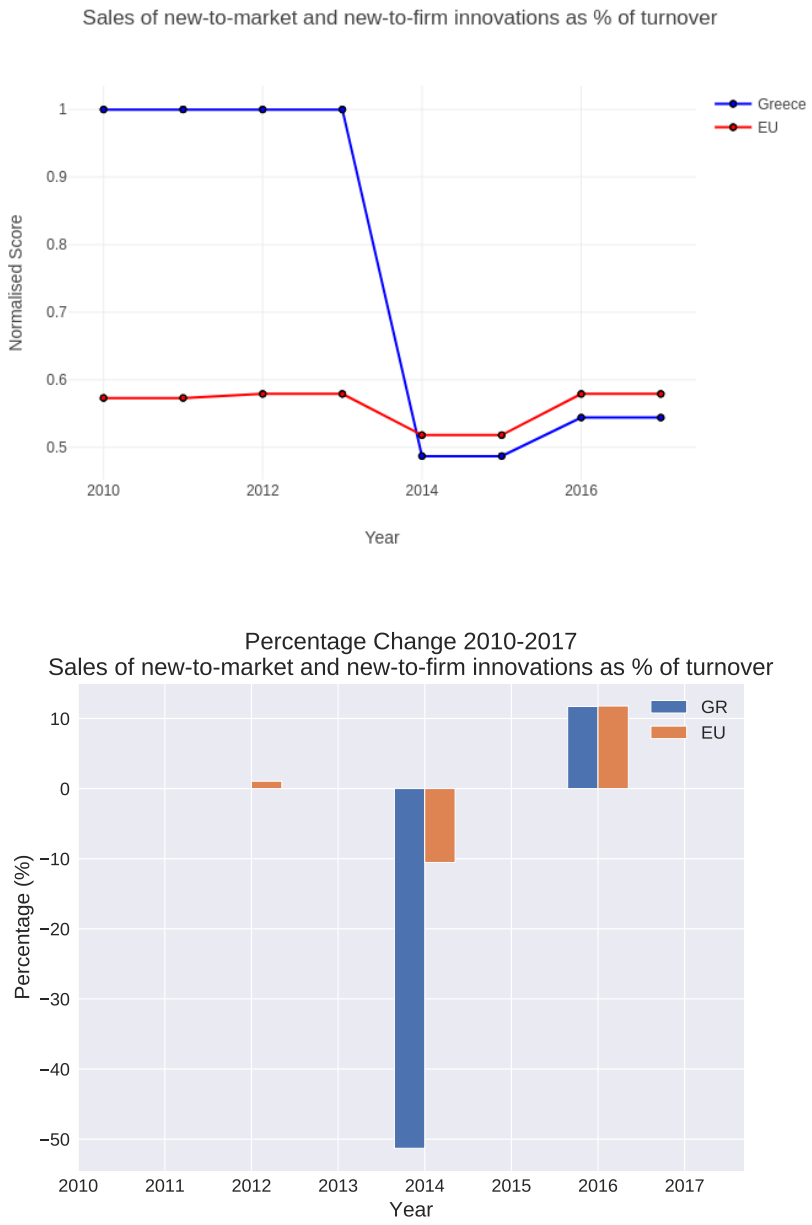

FIGURE A.30: Sales of new-to-market and new-to-firm innovations as % of turnover (on the top) and percentage increase (at the bottom).

Definition 31 *"Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country" indicator has the number of scientific publications among the top-10% most cited publications worldwide and total number of scientific publications as numerator and denominator respectively.*

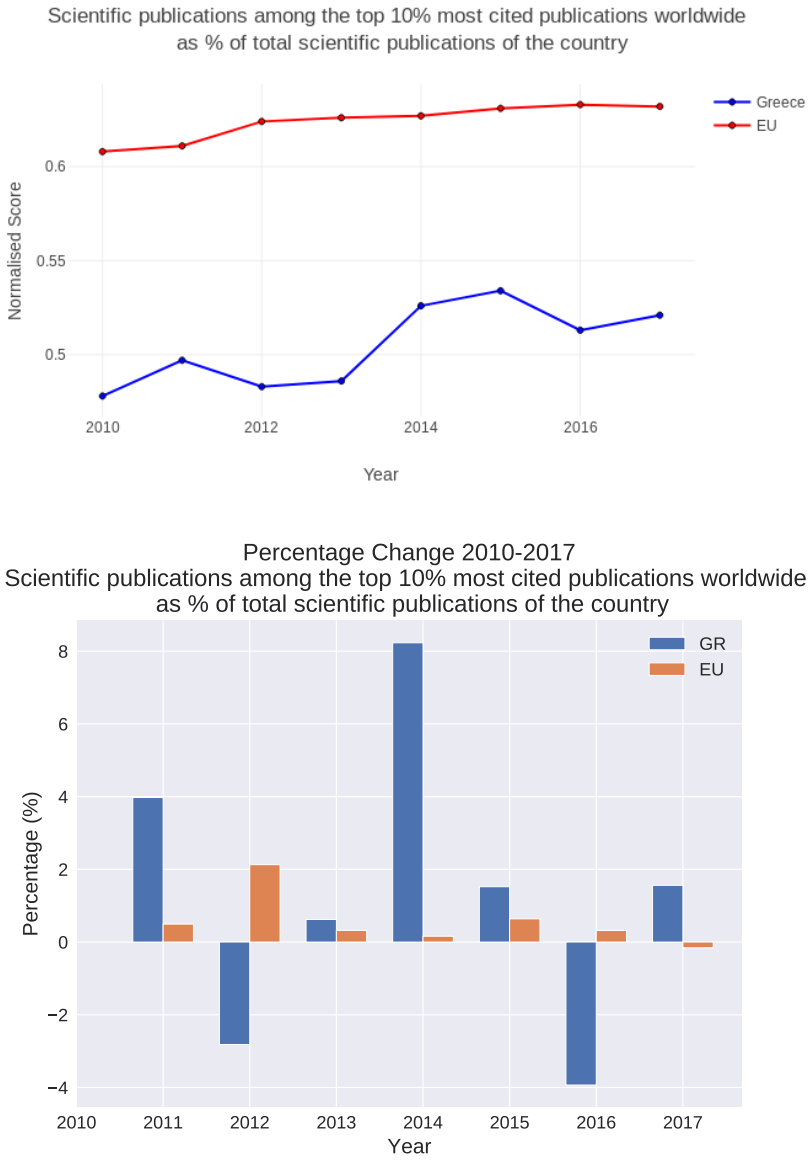

FIGURE A.31: Scientific publications among the top 10% most cited publications worldwide as % of total scientific publications of the country (on the top) and percentage increase (at the bottom).

Definition 32 *"SMEs innovating in-house as % of SMEs" is the number of SMEs with in-house innovation activities divided by the total number of SMEs.*

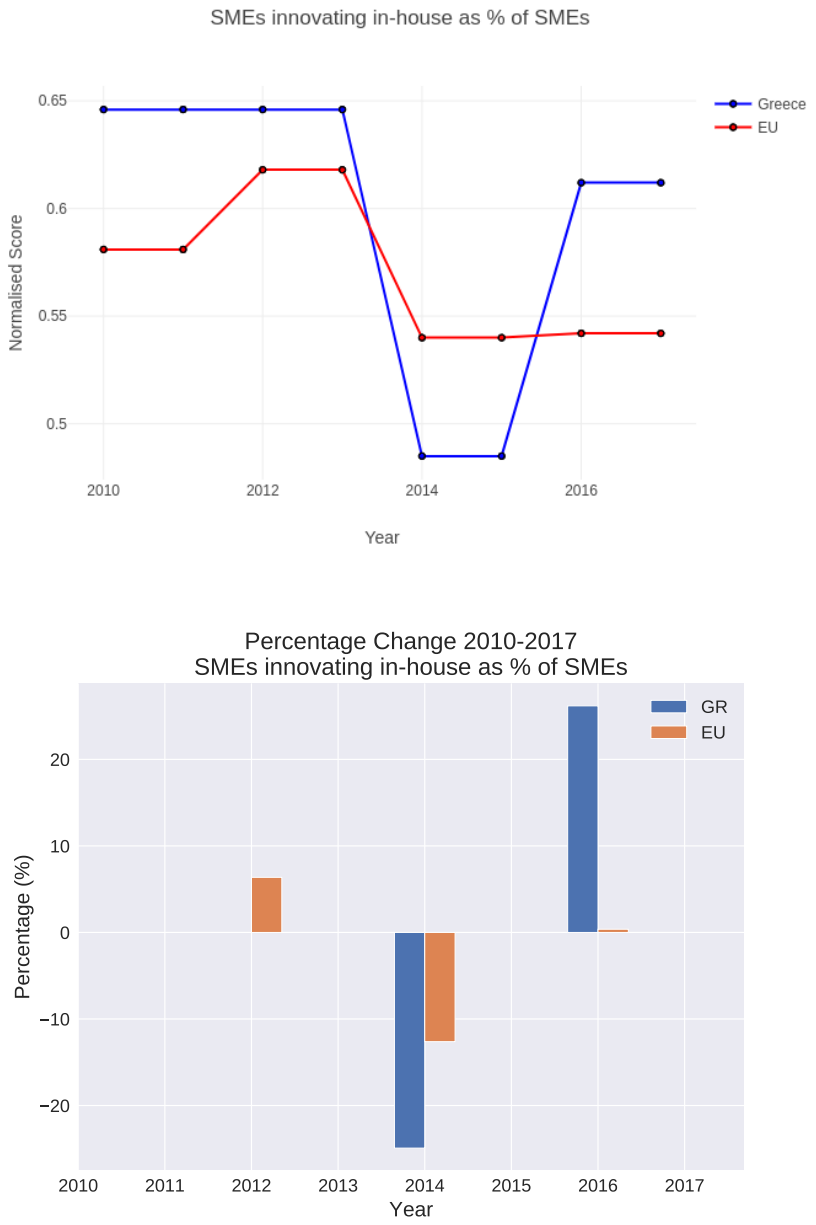

FIGURE A.32: SMEs innovating in-house as % of SMEs (on the top) and percentage increase (at the bottom).

Definition 33 *"SMEs introducing marketing or organisational innovations as % of SMEs" is the number of SMEs who introduced at least one new organizational innovation or marketing innovation divided by the total number of SMEs.*

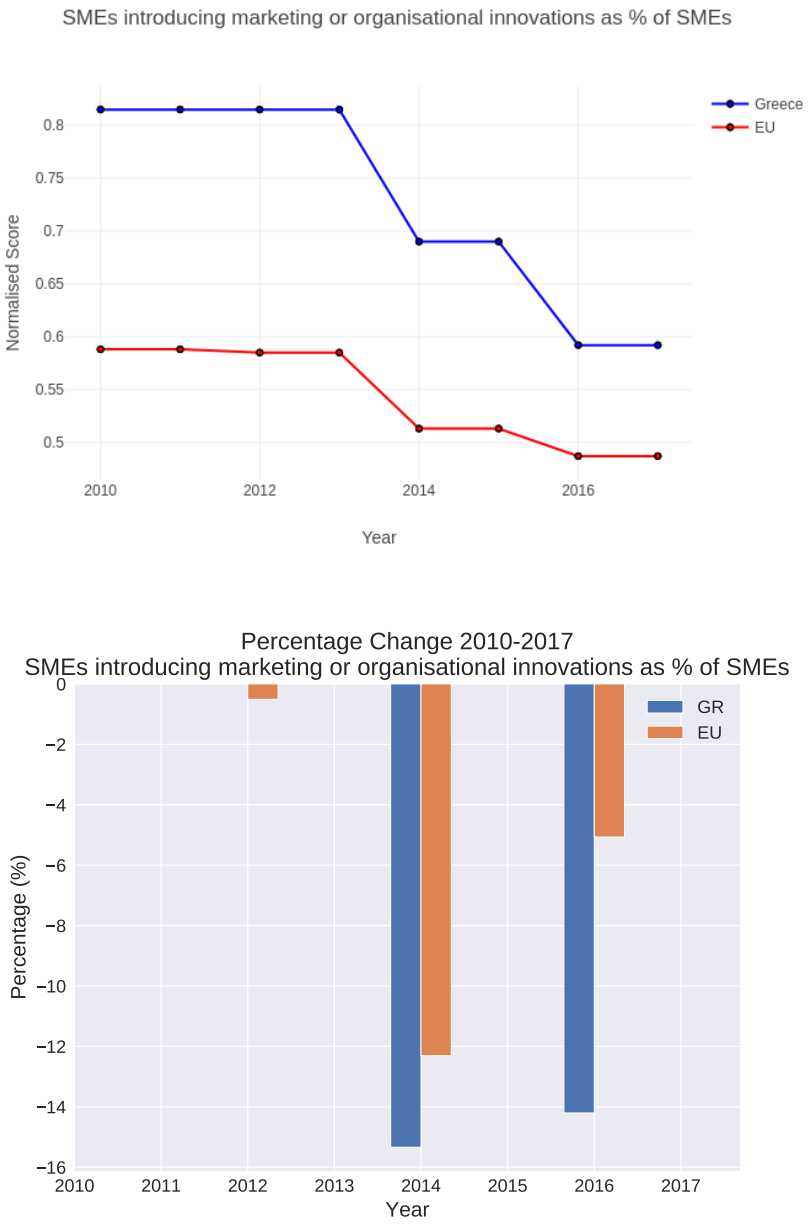

FIGURE A.33: SMEs introducing marketing or organisational innovations as % of SMEs (on the top) and percentage increase (at the bottom).

Definition 34 *"SMEs introducing product or process innovations as % of SMEs" indicator is a fraction of the number of Small and medium-sized enterprises (SMEs) who introduced at least one product innovation or process innovation either new to the enterprise or new to their market and the total number of SMEs.*

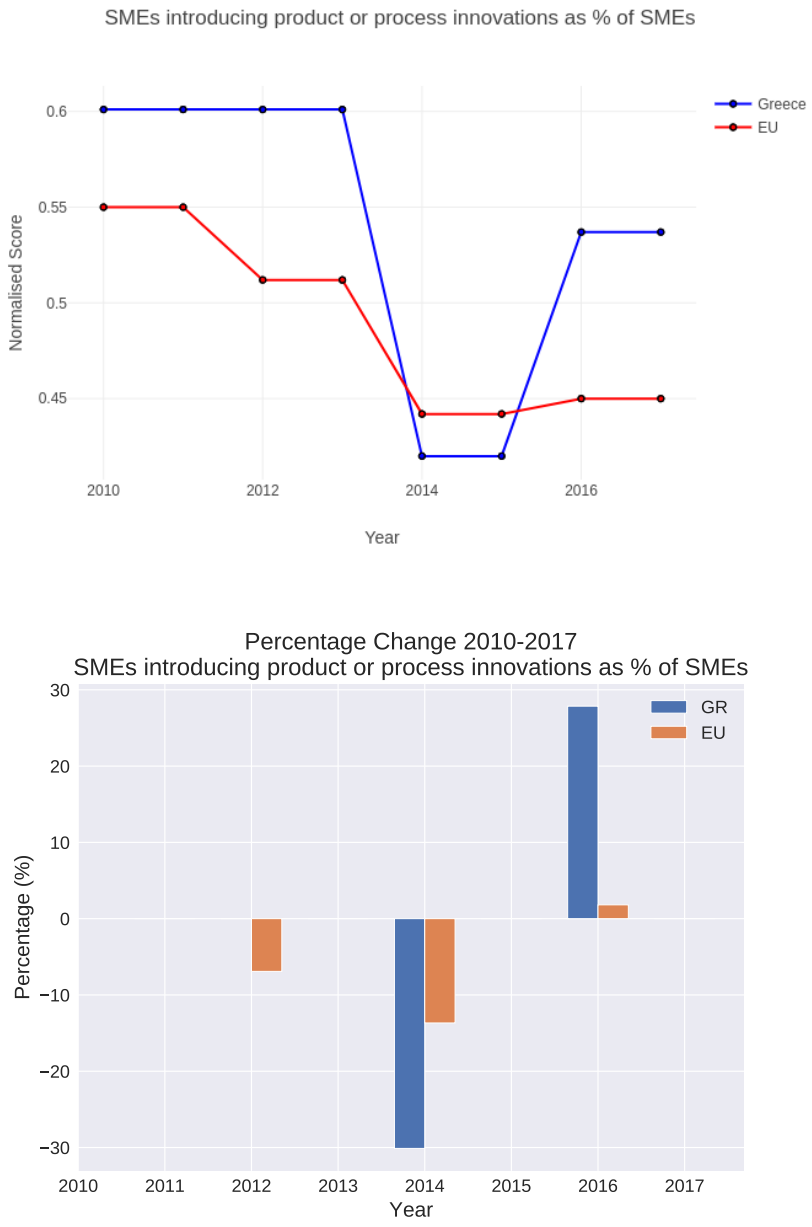

FIGURE A.34: SMEs introducing product or process innovations as % of SMEs (on the top) and percentage increase (at the bottom).

Definition 35 *"Trademark applications per billion GDP (in PPS)" is the number of trademark applications applied for at European Union Intellectual Property Office plus number of trademark applications applied for at World Intellectual Property Office ("yearly Madrid applications by origin") divided by GDP in in Purchasing Power Standard.*

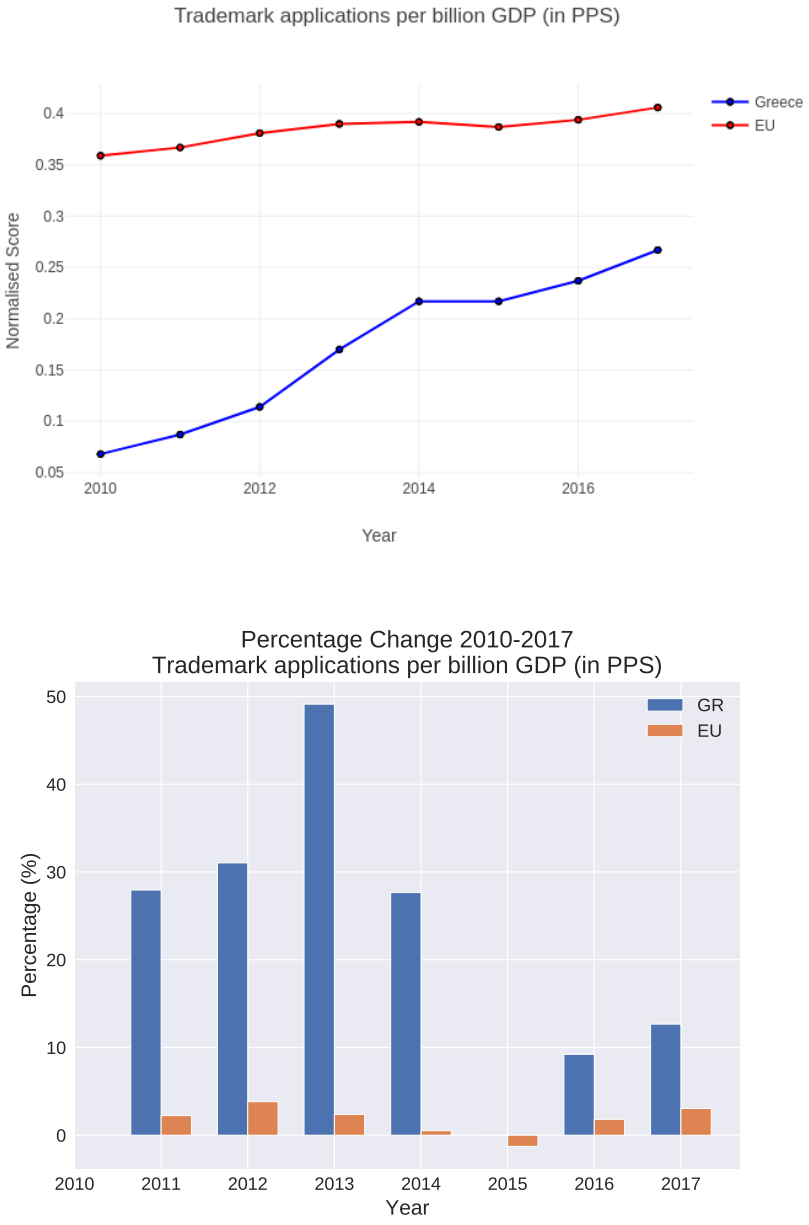

FIGURE A.35: Trademark applications per billion GDP (in PPS) (on the top) and percentage increase (at the bottom).

Definition 36 *"Venture Capital (% of GDP)" is a fraction which has as numerator the venture capital expenditures and denominator the GDP.*

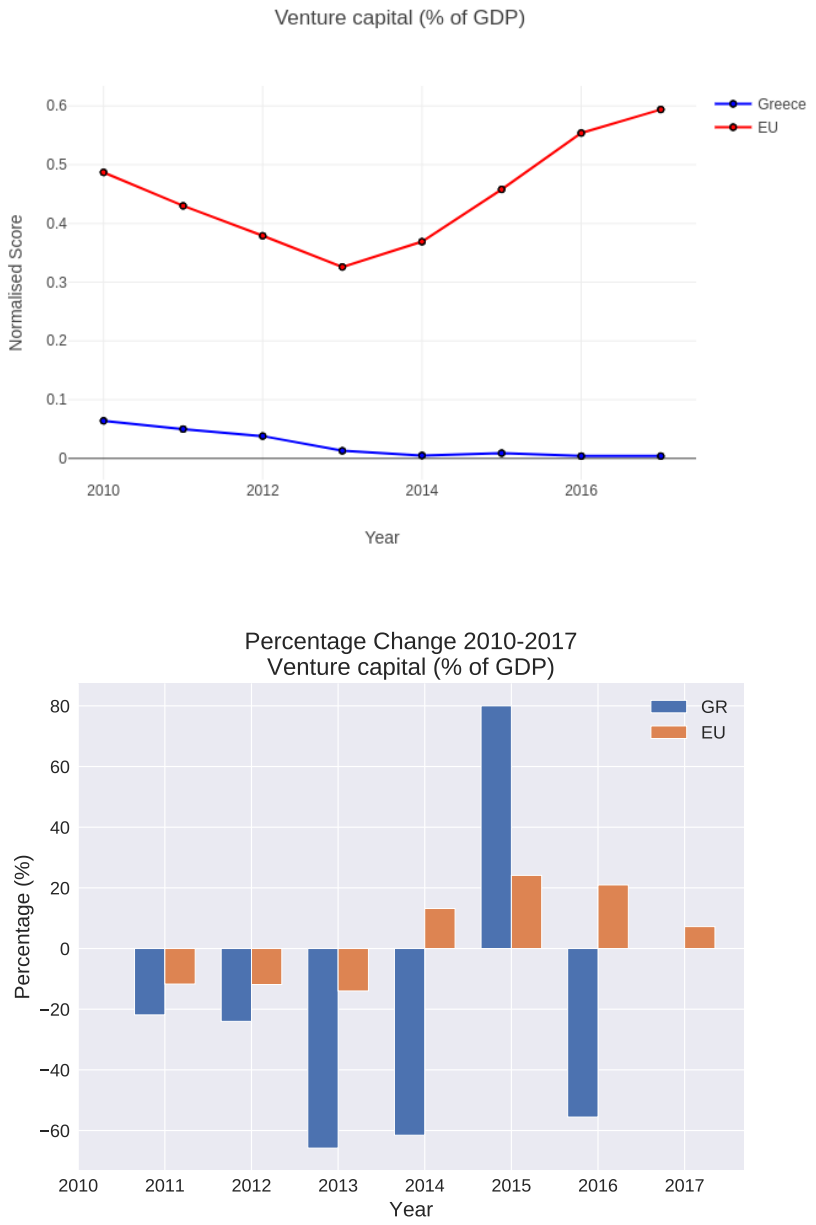

FIGURE A.36: Venture Capital (% of GDP) (on the top) and percentage increase (at the bottom).

Bibliography

[1] P. F. P. F. Drucker, *The practice of management*, 1st ed. New York: Harper & Row, 1954, p. 404, ISBN: 9780060913168. [Online]. Available: https://www.worldcat.org/title/practice-of-management/oclc/230717.

[2] J. C. Henderson and C. M. Lentz, "Learning, Working, and Innovation: A Case Study in the Insurance Industry", *Journal of Management Information Systems*, vol. 12, no. 3, pp. 43–64, Dec. 1995, ISSN: 0742-1222. DOI: 10.1080/07421222.1995.11518090. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/07421222.1995.11518090.

[3] J. A. Schumpeter and R. Opie, *The theory of economic development; an inquiry into profits, capital, credit, interest, and the business cycle*, Harvard University Press, 1934, p. 255, ISBN: 9780674879904. [Online]. Available: http://www.hup.harvard.edu/catalog.php?isbn=9780674879904.

[4] R. Carnegie and Business Council of Australia., *Managing the innovating enterprise : Australian companies competing with the world's best*. Business Library, 1993, p. 427, ISBN: 1863501517. [Online]. Available: https://catalogue.nla.gov.au/Record/1573090.

[5] H. Tohidi and M. M. Jabbari, "The important of Innovation and its Crucial Role in Growth, Survival and Success of Organizations", *Procedia Technology*, vol. 1, pp. 535–538, Jan. 2012, ISSN: 2212-0173. DOI: 10.1016/J.PROTCY.2012.02.116. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221201731200117X.

[6] *Industrial policy | Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/industry/policy_en.

[7] *Innobarometer | Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/industry/innovation/facts-figures/innobarometer_en.

[8] *Innovation | Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/industry/innovation_en.

[9] *Monitoring innovation | Internal Market, Industry, Entrepreneurship and SMEs*. [Online]. Available: https://ec.europa.eu/growth/industry/innovation/facts-figures_en.

[10] European Commission, *DocsRoom - European Commission EIS 2018 Methodology Report*, 2018. [Online]. Available: https://ec.europa.eu/docsroom/documents/30081.

[11] Y. Dodge, F. H. C. F. H. C. Marriott, and International Statistical Institute., *The Oxford dictionary of statistical terms*. Oxford University Press, 2003, p. 498, ISBN: 9780199206131. [Online]. Available: https://global.oup.com/academic/product/the-oxford-dictionary-of-statistical-terms-9780199206131?cc=gr&lang=en&.

[12] Stanford University. and Center for the Study of Language and Information (U.S.), *Stanford encyclopedia of philosophy*. Stanford University, 1997. [Online]. Available: https://plato.stanford.edu/entries/statistics/.

[13] "Inferential and Descriptive Statistics", in *Encyclopedia of Epidemiology*, 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. DOI: 10.4135/9781412953948.n226. [Online]. Available: http://methods.sagepub.com/reference/encyc-of-epidemiology/n226.xml.

[14] B. L. WELCH, "THE GENERALIZATION OF 'STUDENT'S' PROBLEM WHEN SEVERAL DIFFERENT POPULATION VARLANCES ARE INVOLVED", *Biometrika*, vol. 34, no. 1-2, pp. 28–35, Jan. 1947, ISSN: 0006-3444. DOI: 10.1093/biomet/34.1-2.28. [Online]. Available: https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/34.1-2.28.

[15] K. Ito and Nihon Sugakkai., *Encyclopedic dictionary of mathematics*. MIT Press, 1993, p. 2148, ISBN: 0262590204.

[16] P. G. Hoel, S. C. Port, and C. J. Stone, *Introduction to statistical theory*. Houghton-Mifflin, 1971, p. 237, ISBN: 0395046378.

[17] P. J. Veazie, "Understanding Statistical Testing", *SAGE Open*, vol. 5, no. 1, Mar. 2015, ISSN: 2158-2440. DOI: 10.1177/2158244014567685. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2158244014567685.

[18] A. P. Husser, "Correlation Analysis", in *The International Encyclopedia of Communication Research Methods*, Hoboken, NJ, USA: John Wiley & Sons, Inc., Nov. 2017, pp. 1–2. DOI: 10.1002/9781118901731.iecrm0048. [Online]. Available: http://doi.wiley.com/10.1002/9781118901731.iecrm0048.

[19] B. Ratner, "The correlation coefficient: Its values range between +1/-1, or do they?", *Journal of Targeting, Measurement and Analysis for Marketing*, vol. 17, no. 2, pp. 139–142, Jun. 2009, ISSN: 1479-1862. DOI: 10.1057/jt.2009.5. [Online]. Available: http://link.springer.com/10.1057/jt.2009.5.

[20] "Pearson's Correlation Coefficient", in *Encyclopedia of Public Health*, Dordrecht: Springer Netherlands, 2008, pp. 1090–1091. DOI: 10.1007/978-1-4020-5614-7{_}2569. [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-4020-5614-7_2569.

[21] M. Shadmani, S. Marofi, and M. Roknian, "Trend Analysis in Reference Evapotranspiration Using Mann-Kendall and Spearman's Rho Tests in Arid Regions of Iran", *Water Resources Management*, vol. 26, no. 1, pp. 211–224, Jan. 2012, ISSN: 0920-4741. DOI: 10.1007/s11269-011-9913-z. [Online]. Available: http://link.springer.com/10.1007/s11269-011-9913-z.

[22] B. H. Baltagi, "Generalized Least Squares", in *Econometrics*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 235–251. DOI: 10.1007/978-3-662-04693-7{_}9. [Online]. Available: http://link.springer.com/10.1007/978-3-662-04693-7_9.

[23] M. Hallin, "Gauss-Markov Theorem in Statistics", in *Wiley StatsRef: Statistics Reference Online*, Chichester, UK: John Wiley & Sons, Ltd, Sep. 2014. DOI: 10.1002/9781118445112.stat07536. [Online]. Available: http://doi.wiley.com/10.1002/9781118445112.stat07536.

[24] J. Staudenmayer and J. P. Buonaccorsi, *Measurement Error in Linear Autoregressive Models*. DOI: 10.2307/27590617. [Online]. Available: https://www.jstor.org/stable/27590617.

[25] I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection", *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003, ISSN: ISSN 1533-7928. [Online]. Available: http://jmlr.csail.mit.edu/papers/v3/guyon03a.html.

[26] M. L. Bermingham, R. Pong-Wong, A. Spiliopoulou, C. Hayward, I. Rudan, H. Campbell, A. F. Wright, J. F. Wilson, F. Agakov, P. Navarro, and C. S. Haley, "Application of high-dimensional feature selection: evaluation for genomic prediction in man", *Scientific Reports*, vol. 5, no. 1, p. 10 312, Sep. 2015, ISSN: 2045-2322. DOI: 10.1038/srep10312. [Online]. Available: http://www.nature.com/articles/srep10312.

[27] G. G. M. James, *An introduction to statistical learning : with applications in R*, ISBN: 1461471370.

[28] S. J. S. J. Russell, P. Norvig, and E. Davis, *Artificial intelligence : a modern approach*, p. 1132, ISBN: 0136042597.

[29] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sep. 2000, ISBN: 9780471722144. DOI: 10.1002/0471722146. [Online]. Available: http://doi.wiley.com/10.1002/0471722146.

[30] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification", *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, 2008, ISSN: 1532-4435. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1442794.

[31] J. R. Quinlan, "Induction of Decision Trees", *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, ISSN: 0885-6125. DOI: 10.1023/A:1022643204877. [Online]. Available: http://dx.doi.org/10.1023/A:1022643204877.

[32] T. K. Ho, "Random Decision Forests", in *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ser. ICDAR '95, Washington, DC, USA: IEEE Computer Society, 1995, pp. 278–, ISBN: 0-8186-7128-9. [Online]. Available: http://dl.acm.org/citation.cfm?id=844379.844681.

[33] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees", *Mach Learn (*, 2006. DOI: 10.1007/s10994-006-6226-1. [Online]. Available: https://orbi.uliege.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf.

[34] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 2000, p. 314, ISBN: 9780387987804.

[35] V. Vapnik, "An overview of statistical learning theory", *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999, ISSN: 10459227. DOI: 10.1109/72.788640. [Online]. Available: http://ieeexplore.ieee.org/document/788640/.

[36] L. Hyvärinen, "Innovativeness and its Indicators in Small- and Medium-sized Industrial Enterprises", *International Small Business Journal: Researching Entrepreneurship*, vol. 9, no. 1, pp. 64–79, Oct. 1990, ISSN: 0266-2426. DOI: 10.1177/026624269000900106. [Online]. Available: http://journals.sagepub.com/doi/10.1177/026624269000900106.

[37] C. BENEKI, D. GIANNIAS, and G. MOUSTAKAS, "INNOVATION AND ECONOMIC PERFORMANCE: the case of Greek SMEs", *Regional and Sectoral Economic Studies*, vol. 12, no. 1, pp. 43–54, 2012. [Online]. Available: https://ideas.repec.org/a/eaa/eerese/v12y2012i1_3.html.

[38] A. Eleftherios, E. Nikolaos, G. Antonios, and T. Anastasios, "RD activity and operating performance of small and medium-sized enterprises (SMEs): The case of a small open economy", *Journal of Accounting and Taxation*, vol. 8, no. 4, pp. 40–50, Sep. 2016, ISSN: 2141-6664. DOI: 10.5897/JAT2016.0233. [Online]. Available: http://academicjournals.org/journal/JAT/article-abstract/2BFE02660733.

[39] V. Souitaris, "Strategic Influences of Technological Innovation in Greece", *British Journal of Management*, vol. 12, no. 2, pp. 131–147, Jun. 2001, ISSN: 1045-3172. DOI: 10.1111/1467-8551.00190. [Online]. Available: http://doi.wiley.com/10.1111/1467-8551.00190.

[40] B. Asheim, A. Isaksen, C. Nauwelaers, and F. Tödtling, *Regional Innovation Policy for Small-Medium Enterprises*. Edward Elgar Publishing, 2003, ISBN: 9781781009659. DOI: 10.4337/9781781009659. [Online]. Available: https://www.elgaronline.com/view/1843763982.xml.

[41] P. Hajek and R. Henriques, "Modelling innovation performance of European regions using multi-output neural networks", *PLOS ONE*, vol. 12, no. 10, M. C. Díaz Roldán, Ed., e0185755, Oct. 2017, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0185755. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0185755.

[42] P. Hajek and J. Stejskal, "Predicting the innovation activity of chemical firms using an ensemble of decision trees", in *2015 11th International Conference on Innovations in Information Technology (IIT)*, IEEE, Nov. 2015, pp. 35–39, ISBN: 978-1-4673-8509-1. DOI: 10.1109/INNOVATIONS.2015.7381511. [Online]. Available: http://ieeexplore.ieee.org/document/7381511/.

[43] M. de la Paz-Marín, P. Campoy-Muñoz, and C. Hervás-Martínez, "Non-linear multiclassifier model based on Artificial Intelligence to predict research and development performance in European countries", *Technological Forecasting and Social Change*, vol. 79, no. 9, pp. 1731–1745, Nov. 2012, ISSN: 0040-1625. DOI: 10.1016/J.TECHFORE.2012.06.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0040162512001485?via%3Dihub.

[44] S.-C. Chien, T.-Y. Wang, and S.-L. Lin, "Application of neuro-fuzzy networks to forecast innovation performance – The example of Taiwanese manufacturing industry", *Expert Systems with Applications*, vol. 37, no. 2, pp. 1086–1095, Mar. 2010, ISSN: 0957-4174. DOI: 10.1016/J.ESWA.2009.06.107. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741740900596X.

[45] T.-Y. Wang and S.-C. Chien, "Forecasting innovation performance via neural networks—a case of Taiwanese manufacturing industry", *Technovation*, vol. 26, no. 5-6, pp. 635–643, May 2006, ISSN: 0166-4972. DOI: 10.1016/J.TECHNOVATION.2004.11.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166497204002135.

[46] S. Saberi and R. M. Yusuff, "Neural network application in predicting advanced manufacturing technology implementation performance", *Neural Computing and Applications*, vol. 21, no. 6, pp. 1191–1204, Sep. 2012, ISSN: 0941-0643. DOI: 10.1007/s00521-010-0507-0. [Online]. Available: http://link.springer.com/10.1007/s00521-010-0507-0.

[47] N. Klimova, O. Kozyrev, and E. Babkin, *Innovation in Clusters*. Cham: Springer International Publishing, 2016, ISBN: 978-3-319-21108-4. DOI: 10.1007/978-3-319-21109-1. [Online]. Available: http://link.springer.com/10.1007/978-3-319-21109-1.

[48] E. Roszko-Wójtowicz and J. Białek, "Diverse approaches to the multidimensional assessment of innovation in the European union", *Acta Oeconomica*, vol. 68, no. 4, pp. 521–547, Dec. 2018, ISSN: 0001-6373. DOI: 10.1556/032.2018.68.4.3. [Online]. Available: https://www.akademiai.com/doi/10.1556/032.2018.68.4.3.

[49] K. Kalapouti, K. Petridis, C. Malesios, and P. K. Dey, "Measuring efficiency of innovation using combined Data Envelopment Analysis and Structural Equation Modeling: empirical study in EU regions", *Annals of Operations Research*, pp. 1–24, Dec. 2017, ISSN: 0254-5330. DOI: 10.1007/s10479-017-2728-4. [Online]. Available: http://link.springer.com/10.1007/s10479-017-2728-4.

[50] Jan van den Ende and Timo van Balen, "Innovativeness of the Netherlands relative to EU countries", Rotterdam School of Management Erasmus University, Rotterdam, the Netherlands, Tech. Rep., 2017.

[51] E. Roszko-Wójtowicz and J. Białek, "EVALUATION OF THE EU COUNTRIES' INNOVATIVE POTENTIAL - MULTIVARIATE APPROACH", *Statistics in Transition. New Series*, vol. 18, no. 1, pp. 167–180, 2017, ISSN: 1234-7655. DOI: 10.21307/stattrans-2016-064. [Online]. Available: https://www.exeley.com/statistics_in_transition/doi/10.21307/stattrans-2016-064.

[52] *Welcome to Python.org*. [Online]. Available: https://www.python.org/.

[53] *Anaconda Python/R Distribution - Anaconda*. [Online]. Available: https://www.anaconda.com/distribution/.

[54] *Python Data Analysis Library — pandas: Python Data Analysis Library*. [Online]. Available: https://pandas.pydata.org/.

[55] *SciPy.org — SciPy.org*. [Online]. Available: https://www.scipy.org/.

[56] *NumPy — NumPy*. [Online]. Available: https://www.numpy.org/.

[57] *Matplotlib: Python plotting — Matplotlib 3.1.0 documentation*. [Online]. Available: https://matplotlib.org/.

[58] *Modern Analytic Apps for the Enterprise - Plotly*. [Online]. Available: https://plot.ly/.

[59] *Project Jupyter | Home*. [Online]. Available: https://jupyter.org/.

[60] R. Paternoster, R. Brame, P. Mazerolle, and A. Piquero, "USING THE CORRECT STATISTICAL TEST FOR THE EQUALITY OF REGRESSION COEFFICIENTS", Tech. Rep. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.9930&rep=rep1&type=pdf.