CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Economic Studies

Jan Malecha

Innovation Indicator Analysis in the European Union: A Machine Learning Approach

Bachelor thesis

Prague 2019

Author: Jan Malecha

Supervisor: Petr Pleticha, MSc.

Academic Year: 2018/2019

Bibliographic note

MALECHA, Jan. *Innovation Indicator Analysis in the European Union: A Machine Learning Approach*. Prague, 2019. 49 pp. Bachelor thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies. Thesis supervisor Petr Pleticha, MSc.

Abstract

The European Commission annually publishes a European Innovation Scoreboard (EIS) as a tool to measure the innovation performance of the EU Member States. This thesis extends the analysis published in the EIS 2018 in two different manners. The first part, a clustering analysis, examines the partition of the EU Member States to innovation performance groups. The thesis comes with a unique scheme of partition created by using hierarchical clustering. A comparison with the existing scheme shows that the general trends are similar in both schemes. The only main exception is the differentiation of the British Isles and Luxembourg apart from the other high performing countries. The proposed scheme provides insight about the within-cluster similarities, such as the similarity of Finland, Sweden and Denmark and their relative distinction from France, although they belong to one cluster. The second part, a regression analysis, attempts to examine the impact of innovations on real labour productivity. Contrary to existing literature, we do not find a statistically significant relationship between productivity and the components of the EIS. Additionally, the analysis is extended by the lasso estimation that provides a variable selection. The latter approach improves our findings and identifies four EIS indicators with positive impact on labour productivity.

Keywords

innovation, innovation performance, machine learning, hierarchical clustering, lasso, composite indicator, European Union, European Innovation Scoreboard

Abstrakt

Evropská komise každoročně vydává Evropský srovnávací přehled inovací, který je nástrojem k měření inovační výkonnosti členských států Evropské unie (EU). Tato práce rozšiřuje analýzy publikované v Evropském srovnávacím přehledu inovací 2018 ve dvou směrech. První část zkoumá, pomocí shlukové analýzy, rozdělení členských států EU do skupin podle inovační výkonnosti. V této práci přicházíme s unikátním schématem řešení tohoto rozdělení za použití hierarchického shlukování. Při porovnání našeho schématu se schématem stávajícím se ukazuje, že hlavní vzory v obou schématech jsou velmi podobné. Jedinou výjimkou je odlišnost Velké Británie, Irska a Lucemburska od ostatních nadprůměrně výkonných států. Námi navrhované schéma navíc přináší informace o podobnosti států uvnitř jednotlivých výkonnostních skupin, například uvádí podobnost mezi Finskem, Švédskem a Dánskem a jejich relativní odlišnost od Francie, přestože jsou v jedné skupině. V druhé části se pomocí regresní analýzy pokoušíme zkoumat efekt inovací na reálnou lidskou produktivitu. Na rozdíl od existující literatury nenacházíme statisticky významný vztah mezi produktivitou a indikátory Evropského srovnávacího přehledu inovací. Proto následně rozšiřujeme analýzu o penalizační metodu, označovanou jako lasso metoda, která provádí výběr proměnných z modelu. Tento přístup zlepšuje naše výsledky a jmenuje čtyři inovační indikátory, které značí pozitivní efekt na lidskou produktivitu.

Klíčová slova

inovace, inovační výkonnost, strojové učení, hierarchické shlukování, lasso, souhrnný indikátor, Evropská unie, Evropský srovnávací přehled inovací

Range of thesis: 74 316

Declaration of Authorship

1. The author hereby declares that he compiled this thesis independently, using only the listed resources and literature.

2. The author hereby declares that all the sources and literature used have been properly cited.

3. The author hereby declares that the thesis has not been used to obtain a different or the same degree.

Prague, 10.5.2019 Jan Malecha

Acknowledgments

I would like to express my gratitude to my supervisor, Petr Pleticha MSc., who provided me with valuable insights and helped me to shape this thesis. Last, but not least, I am thankful to my family and friends for their support during the studies.

Institute of Economic Studies
Bachelor thesis proposal

Author	Jan Malecha
Supervisor	Petr Pleticha, MSc.
Proposed topic	Innovation Indicator Analysis in the European Union: A Machine Learning Approach

Topic specification

Innovations are an important determinant of economic growth. But what is innovation? What does the term "innovation" mean? How can we measure innovation? In the European Union, researchers from the Joint Research Centre attempt to answer these questions. They annually publish a report called the European Innovation Scoreboard (EIS). The EIS report summarizes innovation performance in the EU during the previous year. Even though the EIS provides a thorough analysis in a related field, it offers many opportunities for further research. The purpose of this thesis would be to extend the analysis presented by the *European Innovation Scoreboard 2018.*

Research areas

i. Clustering

One of the frequently mentioned components of the EIS is a composite indicator called the Summary Innovation Index. The SII score allows the innovation performance of individual EU Member States to be compared using a single number. The usage of such an approach and the construction of the SII itself is often questioned by researchers. Schibany & Streicher (2008) discuss the strengths and weaknesses of the EIS and provide a review of the selected indicators. Another study from Edquist et al. (2018) critically asses the meaningfulness of the SII under its construction and provides an alternative indicator to describe innovation performance. I would like to use hierarchical clustering to construct an alternative partition of the EU Member States and challenge the status quo.

ii. Regression

With regression analysis, I want to study the impacts of innovation on labour productivity with usage of EIS indicators. Previous research by Griliches (1979) suggests a significant and positive impact of innovations on productivity, these findings had subsequently become established. More recent evidence (Hall, 2011) also study the relationship between innovation and productivity and affirms the previous findings. In my thesis, I would like to verify these findings on the most recent data using the EIS indicators as dependent variables. Thus, I would be able to conclude which indicators support the positive impact of innovations on productivity.

Methodology

In my thesis, I will use mostly machine learning techniques, from both unsupervised and supervised learning. From clustering techniques, an agglomerative hierarchical clustering will be applied on the dataset published together with the European Innovation Scoreboard 2018.

In the regression part, I will use panel data methods to estimate impacts of innovations on productivity. Additionally, I would like to introduce a penalized regression technique, particularly the lasso estimation to perform a variable selection. Such selection should stress out the most relevant indicators from the EIS in the context of productivity.

Outline

Abstract

1. Introduction

2. Literature Review

3. Methodology

4. Data description

5. Empirical Model

6. Conclusion

7. Bibliography

References

Edquist, C., Zabala-Iturriagagoitia, J. M., Barbero, J., & Zofío, J. L. (2018). On the meaning of innovation performance: Is the synthetic indicator of the Innovation Union Scoreboard flawed? *Research Evaluation*, *27*(3), 196–211.

European Innovation Scoreboard 2018. (2018). European Commission.

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics*, *10*(1), 92–116.

Hall, B. H. (2011). *Innovation and productivity*. National bureau of economic research.

Schibany, A., & Streicher, G. (2008). The European Innovation Scoreboard: drowning by numbers? *Science and Public Policy*, *35*(10), 717–732.

Contents

1 INTRODUCTION 3

2 LITERATURE REVIEW 5

2.1 EIS Preface 5

2.2 Existing Literature 7

2.3 Summary Innovation Index 10

3 METHODOLOGY 11

3.1 Hierarchical Clustering 11

3.2 Panel Data Methods 14

3.3 Limitations of the Estimation Methods 16

3.4 The Lasso 18

3.5 Model 19

4 DATA DESCRIPTION 21

5 EMPIRICAL RESULTS 26

5.1 Hierarchical Clustering 26

5.2 Regression 38

6 CONCLUSION 45

7 BIBLIOGRAPHY 47

8 APPENDIX 50

List of Figures

Figure 5.1 - Dendrogram manual 27

Figure 5.2 - Euclidean distance comparison 28

Figure 5.3 - Manhattan distance comparison 28

Figure 5.4 - Dendrogram comparison for $k = 4$ 31

Figure 5.5 - Dendrogram comparison for increasing k 33

Figure 5.6 - Mojena's rule 35

Figure 5.7 - Phylogenetic tree - Manhattan, complete, $k = 4$ 36

Figure 5.8 - RLP.pp model lasso path 43

Figure 5.9 - RLP.pp model cross-validation 44

Figure 8.1 - The correlation matrix 50

List of Tables

Table 2.1 - EIS 2018 performance groups 10

Table 3.1 - Distance measures $D(x, y)$ 12

Table 4.1 - The innovation dimensions of the EIS 21

Table 4.2a - The EIS indicators (the first part) 22

Table 4.3b - The EIS indicators (the second part) 23

Table 5.1 - The last four countries entering the principal cluster 29

Table 5.2 - Variance-inflation factor 39

Table 5.3 - RLP.pp regression results 42

Table 8.1 - List of abbreviations used 50

Table 8.2 - List of labels and codes of the EU Member States 51

Table 8.3 - The Summary Innovation Index score 2018 52

1 Introduction

Innovations are generally recognized as an important factor of productivity growth and consequently of economic growth. Besides, the term innovation could be labelled as one of the most frequent buzzwords in the world of business in the 21st century. In the European Union (EU) innovation performance is tracked by an annual report called the European Innovation Scoreboard (EIS) that provides an extensive assessment of the state of innovations in the EU and also provides a composite index named the Summary Innovation Index (SII) that evaluates the state of innovation systems of the EU Member States. The existing literature (Edquist et al., 2018; Kuhlman et al., 2017; Pełka, 2018) stress out the non-existence of a unique definition of the term "innovation" or "innovation performance", that becomes even more serious when one tries to measure it. Consequently, researchers consider the term "innovation" as ill-defined.

In this thesis, we decided to focus on innovation performance in the EU using the status quo set by the *European Innovation Scoreboard 2018 (EIS 2018)*. In contrast with the *EIS 2018*, this paper selects a machine learning approach as a research method. The text intends to widen existing literature by applying modern methods, similar intention have many recent studies (Edquist et al., 2018; Kuhlman et al., 2017; Szymańska & Zalewska, 2018). Analogous approach should produce new, beneficial and relevant outcomes, which will contribute to the existing literature. The main division of machine learning is on unsupervised learning and on supervised learning. In this thesis both types of machine learning are applied as the thesis consists of two distinct parts - clustering (unsupervised learning) and regression (supervised learning).

Firstly, the clustering section consists of hierarchical clustering analysis. In this section, we examine the partition of the EU Member States by the composite index. Using hierarchical clustering we challenge the status quo scheme and develop an alternative approach of the partition. Consequently, we provide a comparison of possible outcomes of our analysis. The aim is to test if unsupervised learning will provide any results that are not displayed in the segmentation currently used by the EIS but could be potentially interesting or significant. Moreover, we examine if the selected

number of clusters by the *EIS 2018* is meaningful from the perspective of unsupervised learning.

Secondly, in the regression section, we begin with panel data analysis of innovation performance in the EU in the second decade of the 21st century and examine its impacts on labour productivity. In the regression part, we want to test hypotheses such as:

- Which indicators from the EIS have a positive effect on labour productivity?
- Which indicators from the EIS are the most relevant in the context of labour productivity?

Despite the first hypothesis being tested several times (e.g. Castellani et al., 2016; or Griliches, 1979), we see an opportunity to verify their findings on the most recent data. To address the second question, we apply shrinkage methods, so that we aim at selecting the most relevant indicators in the context of labour productivity. Such an approach could be substantial for policy makers if they wanted to address innovations more efficiently.

This thesis is divided into five sections; this is the end of the first section. The second section summarizes the existing literature. An overview of the used methodology is described in the third section. In the fourth section, we carry out description of our dataset, which is used in the following fifth section where we analysed dataset using derived methods. The last, sixth, section summarizes the research results.

2 Literature Review

Firstly, we will put the topic of the European Innovation Scoreboard (EIS) into context, since the EIS is essential for our further work. Secondly, we will summarize the existing literature in this field. Lastly, we introduce the Summary Innovation Index more thoroughly, because we will work with it in the clustering part.

2.1 EIS Preface

The EIS is an annual report of the European Commission (EC), prepared by experts from the Joint Research Centre and invited experts, that is issued since 2001. The EIS provides a comparative analysis of innovation performance in the EU countries and other European countries. On top of that, the EIS presents a comparison of EU Member States with selected developed countries as the United States, Japan, etc., to describe trends in innovation in the global world. As stated in the establishing communication of European innovation reports called *Innovation in a knowledge-driven economy* (2000), the original objective of the EIS is setting the broad policy lines for enhancing innovations in the EU.

Origins of the EIS date back to March 2000, when the Lisbon European Council took place and the European leaders defined their goals for the next decade. Given the EU's aim to become 'the most competitive and dynamic knowledge-based economy in the world' by 2010, instruments were needed to measure the progress towards this target. (*Innovation in a knowledge-driven economy,* 2000)

Innovation in a knowledge-driven economy (2000) is the building block of innovation overviews; it contains a review of previous progress of activities enhancing innovation in the European countries. The study includes a draft of scoreboard's outline and defines the main areas that should be investigated by future scoreboards. All subsequent reports and research papers build their work on conclusions of the preliminary communication *Innovation in a knowledge-driven economy* (2000). A year after, the first official scoreboard *2001 Innovation Scoreboard* has been published; it provided the first complete benchmarking overview of the EU member states. Furthermore, *2001 Innovation Scoreboard* presents the first version of the Summary

Innovation Index (SII). Since 2001, new scoreboards are annually published with the latest 17th edition - the EIS 2018, published last year.

A significant milestone in the EIS development was report from 2005. The EIS 2005 (Sajeva, Gatelli, Tarantola, & Hollanders, 2005), was already published in cooperation of the Joint Research Centre (JRC) and the United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT), which had begun to cooperate with EIS preparation. The EIS report from 2005 (Sajeva et al., 2005) revises the selection of indicators of EIS as well as computation of the summary index, concerning the selected indicators, authors carry out a statistical correlation analysis and principal components analysis (PCA). To conclude, they provide a final list of 26 indicators, which meet the requirements of both analyses. In the context of the SII, the report provides a robustness analysis examining results of different weighting schemes of indicators, which shows the stability of country benchmark rankings under different weighting schemes. Consequently, Sajeva et al. (2005) justify using the simple weighting scheme, i.e. unweighted average of all indicators. Conclusions drawn in report from 2005 remain relevant for later versions and thus provide important findings for our work.

The EIS is a persistently developing overview focusing on European innovations. Not only mentioning methodology development or country expansion, but even some key formalities, e.g. between years 2010-2015 it was called the Innovation Union Scoreboard. Each new version comes with extension or corrections, reacting on the latest challenges connected to innovation systems.

Another component of the EIS report is the Summary Innovation Index (SII), which is a composite indicator obtained as an unweighted average of scaled data of all 27 indicators. The SII is frequently a target of critique by researches as it reduces the complexity of innovation problematics to a single number that could be used for reckless comparison. (Edquist et al., 2018; Schibany & Streicher, 2008) Therefore, countries can easily be ranked by the SII. Such simplification had been made for straightforward interpretation but does not provide a full view on the performance of a country’s innovation system.

2.2 Existing Literature

First, few papers concerning the EIS are presented and then we move to more general topics in the context of innovations and their impacts.

Schibany & Streicher (2008) critically discuss the EIS as the most relevant benchmarking tool in the field of European innovation policy; the paper examines strengths and weaknesses of the EIS indicators and describe indicator's limitation. According to Schibany & Streicher (2008), indicators give a good overview of the innovation system performance of the EU countries as it is not restricted only on R&D (research and development) and include many other aspects such as education or entrepreneurship. Moreover, they stress out the importance of benchmarking for policy makers, since benchmark defines a country's status quo and identifies comparative strengths and weaknesses.

On the other hand, Schibany & Streicher (2008) in their review of a selection of the EIS indicators suggest that the selection seems "eclectic, even somewhat arbitrary", as they mention ignorance of indicator's mutual interactions and mention an absence of a hierarchical ranking of indicators. In conclusion, Schibany & Streicher (2008) state that the selection should be based on a conceptual analysis rather than on a simple statistical correlation analysis (Sajeva et al., 2005). Furthermore, they refer to very heterogeneous temporal behaviour of the EIS indicators. As some indicators, e.g. human capital indicators, are structural by nature only over a long period of time, whereas others have high short-term fluctuations in majority countries, e.g. venture capital indicator. Besides, several indicators are related to business cycle developments. Schibany & Streicher (2008) concludes that due to the possibility of various cycles and development among different countries, one should be careful with the country comparison. On top of that, Schibany & Streicher (2008) point out the multicollinearity issue among the EIS indicators, since some indicators are highly correlated, and this could consequently bias the result of the benchmarking via the SII.

A recent study from Edquist et al. (2018) critically examines if the SII is a meaningful measure of innovation performance. Edquist et al. (2018) argue that the EIS mixes input and output indicators and for the composite indicator computes an average of both. Moreover, they lack explicit definition of innovation performance, which

should be the object of measurement. Consequently, Edquist et al. (2018) do not consider the EIS with current methodology as a meaningful measure of innovation performance in Europe.

Furthermore, they provide an alternative approach for benchmarking innovation performance, using a productivity-based measure based on a robust nonparametric Data Envelopment Analysis (DEA) techniques with differentiation of input and output indicators. Edquist et al. (2018) conclude that the DEA approach yields more relevant results of innovation performance, but they admit that those results are harder to interpret, which is one of the main goals of the SII.

An interesting outcome of this critical review by Edquist et al. (2018), is their case study, which shows unforeseen results of Sweden (that in the original SII constantly ranks among best innovators) as one of the worst performing countries in the output-oriented overview. To conclude, the study identifies Sweden's innovation system as inefficient and underperforming.

Another recent study by Kuhlman et al. (2017) applies machine learning approach for measuring innovation, when they study data from the World Economic Forum (WEF) that captures the level of innovation in 150 countries. In their work (Kuhlman et al., 2017) unsupervised learning is used for revealing groups of related indicators. The study then uses supervised learning for a Group lasso predictive model, where indicators are grouped as a result of hierarchical clustering algorithm. According to Kuhlman et al. (2017), fostering innovation in countries worldwide is a way towards a more prosperous and sustainable world. Also, Kuhlman et al. (2017) stress out the difficulty of quantifying innovation, it goes hand in hand with ill-definition of innovation. Aim of their work is to help policy makers and business leaders to understand better how to foster innovation. With this intention Kuhlman et al. (2017) train a predictive model and create the Open Innovation Index as a new composite indicator, which results from the lasso model.

At first glance, it may seem that Kuhlman et al. (2017) use very similar techniques to those which we specified before as the methodology of this thesis, thus we should state how these approaches differ. However, the papers use these techniques completely differently. Hierarchical clustering is used in this thesis for a complex hierarchical

analyses, whereas Kuhlman et al. (2017) only looks for similar patterns of their indicator. Concerning the regression part, Kuhlman et al. (2017) create a predictive model and construct a new composite indicator. While in this thesis we test classical hypotheses in the context of innovations.

In the seminal paper from Griliches (1979) the relation between R&D and productivity is studied and his work has provided robust evidence of a positive and significant impact of R&D on productivity at the firm-level. Since then, this consensus has been widely held by researches (e.g. Castellani et al., 2016; Guellec & Van Pottelsberghe de la Potterie, 2004; Hall, 2011).

More recent evidence is presented by Hall (2011), where Hall summarizes subsequent literature following Griliches (1979). Moreover, Hall (2011) in his work distinguishes between the innovation of product and process and concludes a substantial positive impacts of product innovation on revenue productivity, whereas the impacts of process innovation remain ambiguous.

Guellec & Van Pottelsberghe de la Potterie (2004) examine the impacts of various sources of investments (public R&D, business R&D, etc.) on productivity growth on data of 16 countries from 1980 to 1998. They reach the conclusion that the impacts of business R&D are the most significant of all. Later on, we could also attempt to estimate the most significant R&D source. Hence various sources of investments are part of our data.

Alternatively, recent evidence of transatlantic (US/EU) productivity gap is presented by Castellani et al. (2016), when they study either the level of business R&D level and the capability to translate business R&D into productivity gains. Castellani et al. (2016) conclude that not only the level of US R&D expenditure is higher, but US firms have higher ability to translate R&D investments into productivity gains. The fact that the EU is lagging behind the US is also repeatedly pointed out by the European Innovation Scoreboard (e.g. *EIS 2018*).

2.3 Summary Innovation Index

The EIS report classifies the EU Member States into four performance groups based on their score calculated by the Summary Innovation Index. Table 2.1. summarizes the distribution of Member States among different classes. The EIS methodology of distinguishing individual classes is relatively straightforward. The country's score of composite indicator, the SII, is compared with the EU average and then assigned into the appropriate group. According to the EIS Methodology Report (2018), for determining performance group membership the following classification scheme is used:

- Innovation Leaders are all countries with a relative performance in 2017 more than 20% above the EU average in 2017
- Strong Innovators are all countries with a relative performance in 2017 between 90% and 120% of the EU average in 2017
- Moderate Innovators are all countries with a relative performance in 2017 between 50% and 90% of the EU average in 2017
- Modest Innovators are all countries with a relative performance in 2017 below 50% of the EU average in 2017

Table 2.1 - EIS 2018 performance groups

Innovation Leaders	Sweden, Denmark, Finland, the Netherlands, the United Kingdom, Luxembourg
Strong Innovators	Germany, Belgium, Ireland, Austria, France, Slovenia
Moderate Innovators	The Czech Republic, Portugal, Malta, Spain, Estonia, Cyprus, Italy, Lithuania. Hungary, Greece, Slovakia, Latvia, Poland, Croatia
Modest Innovators	Bulgaria, Romania

Source: EIS Methodology Report 2018

Hence, group segmentation is made only by comparison of the SII result to EU mean and followingly to assigned to a predetermined category; the methodology may seem as too trivial for any relevant conclusions. Therefore, we see an opportunity to analyse this problem by applying machine learning.

3 Methodology

The first part of this chapter describes a clustering technique of unsupervised learning. The second part starts with an overview of panel data methods, then mentions limitations of the Ordinary Least Squares (OLS) estimation and consequently introduce shrinkage methods.

3.1 Hierarchical Clustering

Unsupervised methods of machine learning have no supposed outcome that is being predicted. Instead, one wants to discover such patterns in the data that were not suspected. The aim of cluster analysis is to group data that has not been labelled, classified or categorized. In other words, clustering finds similarities or patterns in data without any prior knowledge.

We select to use one of the most common cluster analyses - hierarchical clustering. In our analyses, we do not use the resulting score of the Summary Innovation Index (SII) but the dataset from which the SII is computed, and on that dataset, we apply hierarchical clustering. From resulting dendrograms, we observe possibilities for cutting the tree to create groups of countries. Using hierarchical clustering, we speculate about the optimal number of clusters, since we are not primarily concerned with the status quo number of clusters. In conclusion, our findings of hierarchical clustering are summarized.

Hierarchical clustering is a sub-group method of clustering analysis that does not require any prior knowledge of the data used. As the name suggests, it is an algorithm that builds a hierarchy of clusters. In general, two strategies of hierarchical clustering can be identified - agglomerative (bottom-up) and divisive (top-down). Following Hastie, Tibshirani, & Friedman (2017) (or Mojena, 1977), with an agglomerative strategy each observation starts in its own cluster, i.e. for N observations we have N clusters. Then, the closest pair of observations, i.e. the most similar, is merged and creates new cluster by doing so, the number of clusters decreases by one, thus the algorithm has $N-1$ steps that are cyclically repeated until only one principal cluster remains. A divisive method is analogical. In the beginning, all observations start in one

cluster and at each step the algorithm split one of the existing clusters into two new clusters until N clusters are created.

For hierarchical clustering, two parameters must be specified - a distance measure and a linkage method. Firstly, the distance method defines how the distance between observation will be measured. In this work, we will work with the two most common distance measures that are summarized in Table 3.1.

Table 3.1 - Distance measures $D(x, y)$

Euclidean	**Manhattan**
$D(x,y) = \left(\frac{\sum_{i=1}^{N_d}(x_i - y_i)^2}{N_d}\right)^{1/2}$	$D(x,y) = \left(\frac{\sum_{i=1}^{N_d}\lvert x_i - y_i\rvert}{N_d}\right)$
This is the geometric distance in a multidimensional space and is usually computed from raw data (prior to any normalization). The advantage is that this measure is not affected by the addition of new objects (for example outliers). Disadvantage: this measure is affected by the difference in scale (e.g. if the same object is measured in centimeters or in meters the $D(x, y)$ is highly affected.	This distance is the average of distances across dimensions and it supplies similar results to the Euclidean distance. In this measure the effect of outliers is less pronounced (since it is not squared). The name comes from the fact that in most American cities it is not possible to go directly between two points, so the route follows the grid of roads.

Source: Tools for Composite Indicators Building (2005), page 29

Secondly, the linkage method is used to specify how distances between clusters are measured. The three most common linkage methods are *Single linkage*, *Complete linkage* and *Average linkage*. To distinguish linkage methods, we will follow Hastie et al. (2017). A measure of dissimilarity between two clusters (groups of observations) must be defined. Hastie et al. (2017) define a measure of dissimilarity between two clusters as follows:

Definition - dissimilarity $d(G, H)$. *Let G and H represent two groups of observation. The dissimilarity $d(G, H)$ between G and H is computed from the set of pairwise observation dissimilarities $d_{ii'}$ where one member of the pair i is in G and the other element $i'$is in H.*

The above definition of dissimilarity is used in Hastie et al. (2017) to accordingly define linkages methods. Single linkage (SL) takes the intergroup dissimilarity to be that of the closest pair

$$d_{SL}(G,H) = \min_{\substack{i\in G \\ i'\in H}} d_{ii'} \,. \tag{3.1}$$

Complete linkage (CL) takes the intergroup dissimilarity to be that of the furthest pair

$$d_{CL}(G,H) = \max_{\substack{i\in G \\ i'\in H}} d_{ii'} \,. \tag{3.2}$$

Finally, average linkage (AL) takes the average dissimilarity between the groups

$$d_{AL}(G,H) = \frac{1}{N_G N_H} \sum_{i\in G} \sum_{i'\in H} d_{ii'} \,, \tag{3.3}$$

where N_G and N_H are the respective number of observations in each group. When presenting empirical results, we will compare the results of these types of linkages.

The resulting dendrogram depends on the linkage method as we will see later on. Single linkage requires a small dissimilarity between two observations to combine them. Such an approach (called chaining or friend-of-a-friend) tends to construct incompact clusters. Analogically, with the complete linkage two observations are considered close if they are similar to all member of given cluster. Consequently, the average linkage present a compromise between two extreme methods as it attempts to use the positives of both approaches. In Hastie et al. (2017) or James et al. (2013), it is recommended to a researcher to use the method of linkage which produce the most interpretative results, which typically should be complete or average linkage.

To sum up, the main goal of hierarchical cluster analyses is the development of an alternative scheme that would suggest partition of the EU Member States based on their innovation performance. To develop such scheme, we will use an agglomerative hierarchical clustering, for which we stated that two parameters must be specified, i.e. the distance measure and the linkage method. Hence, our clustering analyses start by comparison of various parameter setup. Then, we proceed with a discussion of a suitable choice of a number of clusters for selected parameters, but in general,

hierarchical clustering is not a tool for proposing some optimal number of clusters in contrast with other clustering method, e.g. the k-means. However, we will show the application of the Mojena's stopping rule on our resulting dendrograms and interpret found results. Furthermore, we report the results of hierarchical clustering if we would use the status quo number of clusters ($k = 4$). Consequently, we select and describe more comprehensively the most suitable dendrogram resulting from our analyses. In contrast with the SII score, hierarchical clustering will not provide any ranking of clusters. Though by including the EU28 average observation into our analyses, an average cluster can be recognized.

3.2 Panel Data Methods

To select a correct regression method, we look at dimensions of our data. We works with dataset that contains of an annual data for the EU Member States. Therefore, the dataset has both cross-sectional and time series dimensions. Generally, such data could be treated in two ways - as an independently pooled cross-section or as a panel (longitudinal) data. There are several differences between these approaches. Firstly, for pooled cross-section, the sample differs across time (i.e. random sample) that is not true for our data since we observe the same cross-sections. Secondly, according to Baltagi (2005), regions can be viewed as heterogenous when using the panel data approach. Again, in our case there undoubtedly exist differences among the EU Member States, the SII score could be used as a proof. Consequently, we suspect that our data have time and individual effects and that is why the panel data approach should be preferred. Besides, we should note that our dataset is balanced as there are no missing values in cross-section across all years. In the following paragraphs, we summarize the most common panel data estimation methods.

Pooled OLS

Even though, we stated that pooled OLS is not a preferred method, we introduce it because pooled OLS is the basis for the employed methods. The pooled OLS only pools all cross-sectional observations together and afterwards applies standard OLS procedure on the general panel data equation

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it} \tag{3.4}$$

(describe equation 3.4)

Estimation using pooled OLS is biased and inconsistent if a_i is correlated with x_{it}. The key assumption is that there are no unique attributes of individuals within dataset and that there are no universal effects across time (Wooldridge, 2013)

First-difference

First differencing (FD) is an application of pooled OLS estimation on the first-differenced equation

$$\Delta y_{it} = \beta_0 + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta u_{it}, \tag{3.5}$$

where

$$\Delta y_{it} = y_{i,t} - y_{i,t-1}, \Delta x_{it} = x_{i,t} - x_{i,t-1} \text{ and } \Delta u_{it} = u_{i,t} - u_{i,t-1}. \tag{3.6}$$

By differencing adjacent time periods, the fixed effect is eliminated. Therefore, the key assumption of FD estimation is that the idiosyncratic errors (u_{it}) are uncorrelated with the explanatory variable in each time period, i.e. explanatory variables are strictly exogenous. Moreover, we lose the first year of observation because of the differencing, i.e. we thus have $(T-1) \times N$ observations instead of $T \times N$, where T is the number periods and N the number of cross-sectional observations.

Fixed Effects

Fixed Effects (FE) Estimation is an application of pooled OLS on the time-demeaned equation. In the initial stage we need to compute averages

$$\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}\,, \qquad \bar{x}_{ij} = \frac{1}{T}\sum_{t=1}^{T} x_{itj} \;\text{ and }\; \bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it} \tag{3.7}$$

Then, we can estimate the time-demeaned equation

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \cdots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it}, \tag{3.8}$$

where

$$\ddot{y}_{it} = y_{it} - \bar{y}_{it}, \qquad \ddot{x}_{it} = x_{i,t} - \bar{x}_{it} \;\text{ and }\; \ddot{u}_{it} = u_{it} - \bar{u}_{it}. \tag{3.9}$$

The FE transformation caused the elimination of the intercept, also by the FE estimation the constant factor a_i is "*differenced away*". In contrast with the FD method, time-demeaning in the FE method does not reduce number of observations. According to Wooldridge (2013), the FE estimation method is the most suitable when there exist unique attributes that are not the results of random variation and that are constant over time. In conclusion, we suggest that the FE estimation is the most suitable for our data.

3.3 Limitations of the Estimation Methods

In the previous section, we describe several panel data methods. All of them are based on the Ordinary Least Squares (OLS) estimation. The OLS can easily be labelled as the most common linear estimation method. Full description of the OLS estimation can be found in Wooldridge (2013) or Hastie et al. (2017). In the following section, we discuss the limitations of OLS estimation and explain why we decided to use the lasso estimation later on.

Hastie et al. (2017) state two following reasons why we are often not satisfied with the OLS estimation:

- prediction accuracy/bias-variance trade-off
- interpretation with a high number of predictors

Firstly, the issue called the prediction accuracy causes that the OLS estimates often have low bias but large variance. Subsequently, the prediction accuracy can sometimes be improved by shrinking coefficients. As a consequence, part of bias is sacrificed for a reduction of the variance to improve prediction accuracy. Secondly, we can be

concerned about interpretation when working with a large number of predictors. Then we often would like to determine a smaller subset that exhibits the strongest effects.

For a subsequent description of the bias-variance trade-off, we follow Fortmann-Roe (2012). Consider Y as the variable we are trying to predict and X as the matrix form of dependent variables. Then, the relation between these variables can be expressed as $Y = f(X) + \epsilon$ where the error term ϵ is normally distributed with a mean of zero, so $\epsilon \in N(0, \sigma_\epsilon)$.

Consequently, we may estimate a model $\hat{f}(X)$ of $f(X)$ using linear regressions or another modelling technique. In this case, the expected squared prediction error at a point x is:

$$Err(x) = E\left[\left(Y - \hat{f}(X)\right)^2\right] \tag{3.10}$$

The error from Equation (3.10) may then be decomposed into bias and variance components:

$$Err(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right] + \sigma_\epsilon^2 \tag{3.11}$$

Equation (3.11) can be rewritten as:

$$Err(x) = Bias^2 + Variance + Irreducible\ Error, \tag{3.12}$$

where the third term called irreducible error, is the noise term in the true relationship that cannot fundamentally be reduced by any model.

In our case, not only we need to tackle the complexity of the model and the bias-variance trade-off. We must also handle the issue of multicollinearity, i.e. the high correlation between multiple independent variables. The existing literature (e.g. Wooldridge, 2013) mention multicollinearity as ill-defined, due to the fact that there is no clear definition of multicollinearity boundary. Therefore, it is impossible to clearly specify how much correlation between independent variables is permissible. To detect multicollinearity in an econometric model there exists a few statistics which intend to determine the severity of the multicollinearity issue. The most common statistic is the variance inflation factor (VIF), which is the ratio of the variance of $\hat{\beta}_j$ when fitting the

full model divided by the variance of $\hat{\beta}_j$ if fit on its own. (James et al., 2013) The VIF takes value from above 1, which indicates the complete absence of collinearity. According to James et al. (2013), as a rule of thumb, a VIF value that exceeds 5 or 10 reveals collinearity. Lastly, we take in consideration advice from researches (e.g. Wooldridge, 2013) that we should not give too much emphasis to these rules or to these auxiliary statistics in general. To conclude, we set the VIF threshold to 10.

In previous paragraphs, we call for shrinking of parameters and variable selection. For these reasons, in the next section we introduce a penalized regression methods (so-called shrinkage methods/penalized regression/regularization).

3.4 The Lasso

The Least Absolute Shrinkage and Selection Operator (lasso) was firstly proposed by Tibshirani (1996) as a new method for estimation in linear models. As the name suggests the lasso estimation not only shrinks coefficients but also conducts the variable selection. Thus, the lasso combines properties of both subset selection and shrinkage methods. An alternative to the lasso estimation is another penalized regression method called Ridge regression. It works very similarly to the lasso but have a quadratic penalty (L2 norm) term.

Ridge regression could be used for treating multicollinearity when by enabling for a little bias will decrease the model variance. However, Ridge regression does not provide variable selection, it only shrinks parameters near to zero but does not eliminate them. In other words, we are unable to reduce the number of independent variables by Ridge regression, which is our aim. Thus, we do not include it in further analysis. This thesis presents description only of the lasso estimation, description of other shrinkage methods is beyond the scope of this thesis. A detailed explanation of shrinkage methods is provided in Hastie et al. (2017).

In Hastie et al. (2017) the lasso estimator is defined by:

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right\}, \tag{3.13}$$

where $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage, i.e. the larger the value of the parameter λ, the greater amount of shrinkage. If in Equation (3.13) $\lambda = 0$, then the penalty disappears and standard OLS minimization remains. On the other hand, as λ increases more and more variables shrink until they are all shrunk. This shrinking feature is very useful in order to select only the important variables. Furthermore, by the selection of variables the multicollinearity issue is solved and therefore we are able to avoid possible overfitting.

The lasso estimation is suggested for cross-sectional data. Thus, for applying the lasso on our panel dataset, we follow the same logic as in the previous sections. We manually execute a time-demeaning transformation as for FE method. We consider such approach as appropriate for treating the time-series dimension of panel data. Alternatively, we could use dummy variable regression, i.e. adding dummy for each observation and time period. However, this approach would create a large number of variables and interpretation of the model would be indistinct. Thus, we decided to execute the first method.

In the lasso estimation returns a sequence of models for the user to choose from. The most common method for such model selection is k-fold cross-validation, which is thoroughly described afterward. Cross-validation compares models by the mean squared error (MSE) of estimators that is standardly use for such purpose. Using the definition of MSE from Wooldridge (2013), if W is an estimator of θ then

$$MSE(W) = E[W - \theta]^2 = Var(W) + [Bias(W)]^2. \tag{3.14}$$

3.5 Model

This thesis examines effects of innovation indicators on productivity. We construct a model

$$RLP.pp_{it} = \beta_0 + \sum_{j=1}^{27} \beta_j EIS\_ind_{itj} + a_i + u_{it}, \tag{3.15}$$

where $RLP.pp_{it}$ is the real labour productivity, EIS_ind_j stands for each of 27 indicators included in the *EIS 2018*, a_i is the unobserved effect and u_{it} is the

idiosyncratic error. We will consider a variety of models, which we will compare and provide econometric tests. Intuitively, we expect that the fixed effect model with two-way effect (i.e. effect for individuals and across time) should perform the best. However, we are aware of the huge complexity of our model, since it consists of 27 indicators. Therefore, we expect rather an ambiguous answer to our hypothesis about positive impact of innovations on labour productivity.

As a consequence, we introduce a shrinkage method and answer the second question, concerning the most relevant EIS indicators in context of productivity. Such an approach should propose a selection of relevant innovation indicators in the context of labour productivity, then leaving the model with a more reasonable number of variables. Possibly, this could question the complexity of the *EIS 2018*.

4 Data Description

In this section, data collection for the following analysis is described. It states data sources and discusses data classification or scaling of the dataset for composite indicator. Moreover, it aims at providing a basic overview of the data to a reader.

The data for this research were downloaded from the official website[1] of the European Innovation Scoreboard, as an official dataset of the *European Innovation Scoreboard 2018*, the dataset is up to date to June 26, 2018. The vast majority of used data comes from this dataset and for the purpose of regression some additional variables were downloaded from the Eurostat website.

A methodology report is annually published together with the main EIS report. In the *EIS Methodology Report 2018*, authors thoroughly describe the origin of data, classification and definitions of indicators. Furthermore, they provide a methodology for calculating the SII and describe the process of treating missing data. The *EIS 2018* distinguishes between four types of indicators and names ten innovation dimensions, capturing in total 27 different indicators. The four main types are *Framework conditions, Investments, Innovation activities* and *Impacts*. Table 4.1. summarizes innovation dimensions of the EIS that are further divided to the individual indicators.

Table 4.1 - The innovation dimensions of the EIS

Framework conditions	Human resources, Attractive research system, Innovation-friendly environment
Investments	Finance and support, Firm investments
Innovation activities	Innovators, Linkages, Intellectual assets
Impacts	Employment impacts, Sales impacts

Source: EIS Methodology Report 2018

A list of full names of all 27 indicators is provided in Table 4.2a and Table 4.2.b. Also, the list of proposed shortcuts is presented. Moreover, information concerning data

[1] https://ec.europa.eu/growth/industry/innovation/facts-figures/scoreboards_en

sources is contained there. The most common data source is the Eurostat where the majority of the data drawn from. Data for a few indicators are collected by the Centre for Science and Technology Studies at Leiden University that collaborates with the European Commission.

Table 4.2a - The EIS indicators (the first part)

Shortcut	FRAMEWORK CONDITIONS
	Human resources
doc_grad	1.1.1 New doctorate graduates per 1000 population aged 25-34 (data source: Eurostat)
ter_edu	1.1.2 Percentage population aged 25-34 having completed tertiary education (Eurostat)
lif_lea	1.1.3 Percentage population aged 25-34 participating in lifelong learning (Eurostat)
	Attractive research systems
sci_pub	1.2.1 International scientific co-publications per million population (CWTS LU)
TOP_pub	1.2.2 Scientific publications among TOP 10% most cited publications worldwide as percentage of total scientific publications of the country (CWTS LU)
for_doc	1.2.3 Foreign doctorate students as a percentage of all doctorate students (Eurostat)
	Innovation-friendly environment
bro_pen	1.3.1 Broadband penetration (Eurostat)
opp_ent	1.3.2 Opportunity-driven entrepreneurship (Global Entrepreneurship Monitor)
	INVESTMENTS
	Finance and support
PERD	2.1.1 R&D expenditure in the public sector as a percentage of GDP (Eurostat)
VC	2.1.2 Venture capital expenditures as a percentage of GDP (Invest Europe, Eurostat)
	Firm investments
BERD	2.2.1 R&D expenditure in the business sector as a percentage of GDP (Eurostat)
noRD_ex	2.2.2 Non-R&D innovation expenditures as a percentage of turnover (Eurostat)
IT_tr	2.2.3 Enterprises providing training to develop or upgrade ICT skills of their personnel (Eurostat)

Source: European Innovation Scoreboard 2018

Note: this is only one part of the EIS indicators

Table 4.3b - The EIS indicators (the second part)

Shortcut	INNOVATION ACTIVITIES
	Innovators
SME_pro	3.1.1 SMEs with product or process innovations as a percentage of SMEs (Eurostat)
SME_org	3.1.2 SMEs with marketing or organizational innovations as a percentage of SMEs (Eurostat)
SME_inh	3.1.3 SMEs innovating in-house as a percentage of SMEs (Eurostat)
	Linkages
SME_col	3.2.1 Innovative SMEs collaborating with others as a percentage of SMEs (Eurostat)
pp_pub	3.2.2 Public-private co-publications per million population (CWTS LU)
pr_PERD	3.2.3 Private co-funding of public R&D expenditures as a percentage of GDP (Eurostat)
	Intellectual assets
pat_app	3.3.1 PCT patent applications per billion GDP in PPS (OECD, Eurostat)
tra_app	3.3.2 Trademark applications per billion GDP in PPS (EUIPO, WIPO, Eurostat)
des_app	3.3.3 Design applications per billion GDP in PPS (EUIPO, Eurostat)
	IMPACTS
	Employment impacts
emp_act	4.1.1 Employment in knowledge-intensive activities as a percentage of total employment (Eurostat)
emp_sec	4.1.2 Employment in fast-growing enterprises as a percentage of total employment (Eurostat)
	Sales impacts
tech_exp	4.2.1 Exports of medium and high-tech product as a share of total product exports (Eurostat)
ser_exp	4.2.2 Knowledge-intensive services exports as a percentage of total services exports (Eurostat)
sal_inn	4.2.3 Sales of new-to-market and new-to-firm innovations as a percentage of turnover (Eurostat)

Source: European Innovation Scoreboard 2018

Note: this is only one part of the EIS indicators

As already described before, a component of the EIS is the Summary Innovation Index (SII) - a composite indicator obtained as an unweighted average of scaled data of all 27 indicators.

Due to apparent reasons, data usage differs across particular analyses. Regarding clustering, this thesis uses data that are utilized for the composition of the Summary Innovation Index. For clarity let us call this an SII dataset. However, we do not use the SII score itself, as the score is irrelevant for the analysis. Data used for composite indicator are pre-processed by authors of the EIS. In the *EIS Methodology Report 2018*, following eight different steps of the methodology for calculating the SII are identified:

- Step 1: *Identifying and replacing outliers*
- Step 2: *Setting reference years*
- Step 3: *Imputing for missing values*
- Step 4: *Determining Maximum and Minimum scores*
- Step 5: *Transforming data with highly skewed distributions across countries*
- Step 6: *Calculating re-scaled scores*
- Step 7: *Calculating composite innovation indexes*
- Step 8: *Calculating relative-to-EU performance score*

As the aim of this thesis is not the data collection itself, we will not provide a full descriptions of those steps, all detailed information can be found in the *EIS Methodology Report 2018*. Yet, the first seven steps are relevant for our clustering analyses, we decided to stress out Step 6 and Step 7.

Step 6 describes normalisation of dataset, that is frequently mentioned in the literature (Schibany & Streicher, 2008) as a subject for consideration. According to the *EIS Methodology Report 2018*, the EIS uses so-called *minmax*-normalisation, i.e. the country score is subtracted by the minimum score of the indicator and then divided by the difference between the maximum and minimum score of the indicator.

Significance of Step 7 lies in establishing the calculation of the composite index as the unweighted average of re-scaled scores for all indicators where all indicators receive the same weight. Thus, if data are available for all 27 indicators, individual weight is 1/27 if any indicator is missing the number in denominator is reduced by the number of missing indicators. From the EU Member States, that are object of our interest, only Greece and Malta have incomplete data. In case of Greece, two indicators (for_doc and emp_sec) are missing, thus individual weight of indicator is equal to 1/25. For Malta, there is only one indicator missing (opp_ent), leaving the weight equal to 1/26. As the

Methodology Report 2018 *EIS Methodology Report 2018* states those indicators are historically not collected neither in Greece nor at Malta.

The majority of the data are available up to year 2017, unfortunately it does not hold for all indicators. Therefore, the year 2015 is selected as the latest year available for full dataset. Analogically, selection of the eldest date results by the year 2010, for which all data are available. As a result, regression analysis covers years 2010-2015, which means using six annual periods.

To observe the effects of innovation on selected exogenous variables, we downloaded additional data from the Eurostat, particularly as a productivity dependent variable we have chosen real labour productivity per person. Real labour data are collected by the Eurostat in the form of index with the year 2010 set as a benchmark (value equals 100 for each country).

5 Empirical Results

This chapter is divided into two section. In the first section, the results of hierarchical clustering are presented. The second section provides regression results. The software program used to analyse the data was R, which is a free software environment for statistical computing and graphics.[2] In our work, we used a variety of freely available packages, let us mention only the most important which are *factoextra* for clustering analyses and *glmnet* for penalized regressions. Last but not least, the vast majority of graphs is made by *ggplot2*.

5.1 Hierarchical Clustering

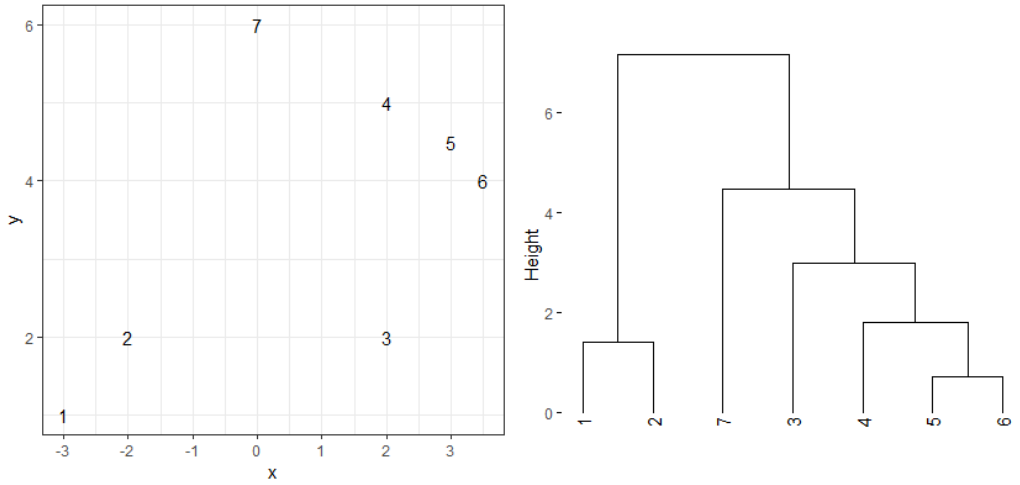
For clarity, we begin with a short manual describing how to read graphical results of hierarchical clustering called dendrograms. A dendrogram is a diagram displaying the hierarchical relations between observation, it is typically used to illustrate the results of hierarchical clustering analysis. To explain the interpretation of dendrogram, we will use a simple two-dimensional example. Consider six observations as in the left-hand panel of Figure 5.1 where a raw data points are plotted. Whereas in the right-hand panel of Figure 5.1, there is the resulting dendrogram of hierarchical clustering (with Euclidean distance and complete linkage) on example's observations. Each leaf (# stands for the number of leaf) of the dendrogram represents one of six observation of our example dataset.

In the initial stage ($height = 0$), all leaves are levelled, then as height increases the closest pair of points (i.e. points #5 and #6) fuse together and create a new branch consisting of leaves #5 and #6. In the next step, we can see that leaves #1 and #2 fuse together in the same way. As $height$ further increases branches themselves fuse, either with other branches or remaining leaves. Until, all observation are merged in the single cluster. James et al. (2013) point out that we should avoid one type of misinterpretation concerning the dendrograms. According to what have been described yet, the observation #3 and #7 should be quite similar to each other, since they are located near

[2] Source - https://www.r-project.org/

each other on the dendrogram. This conclusion is completely false as we easily can see from the left-hand panel of Figure 5.1. In fact, the dendrogram states that leaf #7 is no more similar to leaf #3 than it is to leaves #4, #5 and #6, since the dendrogram is constructed on dissimilarity matrix. Therefore, we are unable to draw conclusions about the similarity of two leaves based on their proximity along the horizontal axis. Hence, conclusions should rather be based on the proximity along the vertical axis.

Figure 5.1 - Dendrogram manual



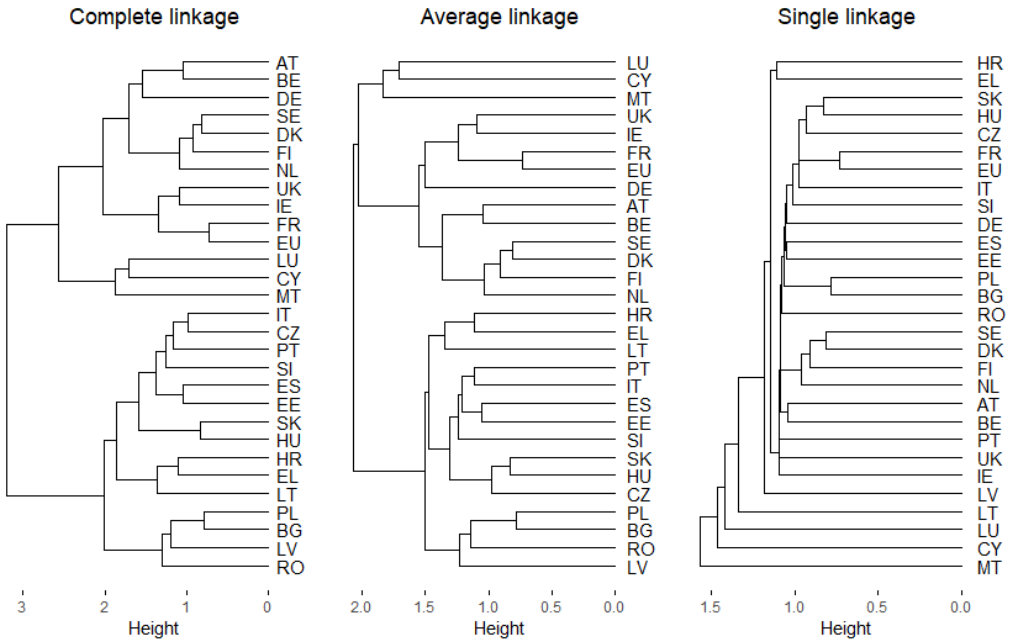
Source - authors own analysis

Furthermore, we are able to identify clusters from the dendrogram by cutting the tree. Imagine, we add a horizontal line to the dendrogram from Figure 5.1 at $heigh = 4$, such horizontal line would cross three vertical lines (easily labelled - the left branch, leaf #7 and the right branch). Generally, we could specify the height of cut ourselves or determine the number of clusters we want, then the algorithm will select cut at appropriate height, both of these approaches are introduced later on.

We begin with a display of dendrograms with different linkage methods for both Euclidean and Manhattan distance. Figure 5.2. and Figure 5.3. show resulting dendrograms with the Euclidean distance and the Manhattan respectively.

The dendrograms derived in this section are repeated in further analysis and figures, when we study other aspects, e.g. dendrogram with Manhattan distance and complete linkage is presented in Figure 5.3, Figure 5.4, Figure 5.5, Figure 5.6 and Figure 5.7.

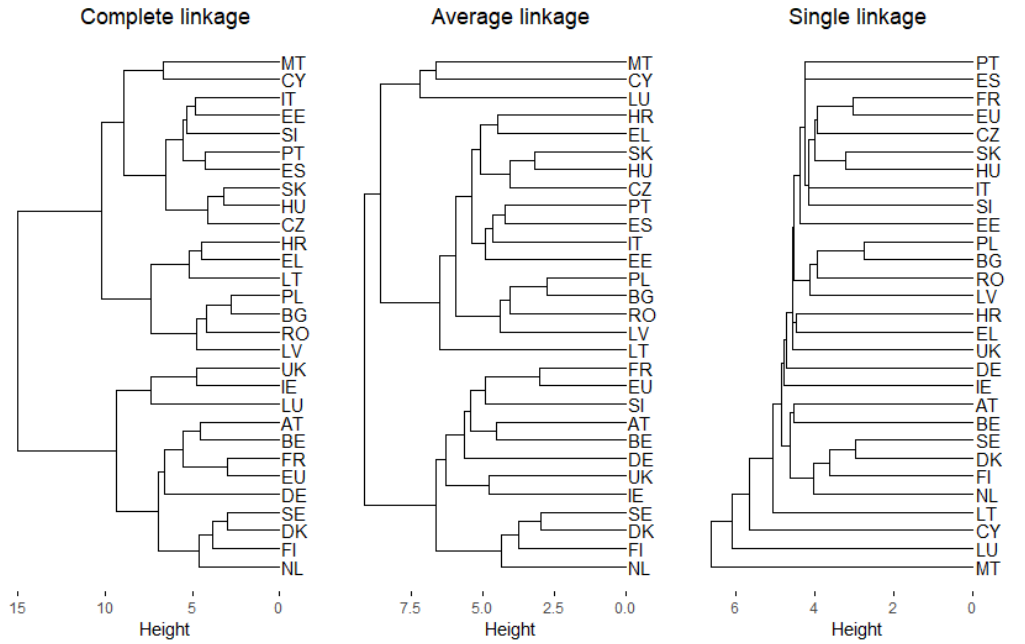
Figure 5.2 - Euclidean distance comparison



Source: authors own analysis based on the EIS data

Note: a manual how to read a dendrogram is provided in the beginning of this section

Figure 5.3 - Manhattan distance comparison



Source: authors own analysis based on the EIS data

Note: a manual how to read a dendrogram is provided in the beginning of this section

At first glance, we can see that the single linkage method is inappropriate for our purpose, i.e. constructing compact groups of EU Member States. However, such result was expected since single linkage tends to combine observations that are linked by a series of close intermediate observations, a phenomenon called chaining (also referred as a *friend-of-friend* procedure - e.g. in Mecke & Stoyan, 2008). Nevertheless, single linkage mainly provides a valuable overview of countries that significantly differ from others as these countries join the principal group in the latest stages. In other words, the single linkage method detects these countries as outliers. Table 5.1. summarizes the last four countries joining the principal cluster.

Table 5.1 - The last four countries entering the principal cluster

Euclidean distance	Lithuania, Luxembourg, Cyprus and Malta
Manhattan distance	Lithuania, Cyprus, Luxembourg and Malta

Source: authors own analysis based on the EIS data
Note: countries are recorder in order how they join the principal cluster

Since both distance measures detected the same countries, we can conclude that Lithuania, Cyprus, Luxembourg & Malta significantly deviate from the other EU Member States. Consequently, we can expect that these countries will deviate in further analyses, saying that we exploit the single linkage method and will further work with complete and average linkages.

Comparing complete linkage from Figure 5.2. and Figure 5.3, both distance measures unambiguously divide the first two groups with similar size. From there, the division by distance measure differs. The dendrogram with Euclidean distance further divides the tree into three clusters, but subsequent segmentation becomes indeterminable. On the other hand, dendrogram with Manhattan distance enables segmentation for at least three branching.

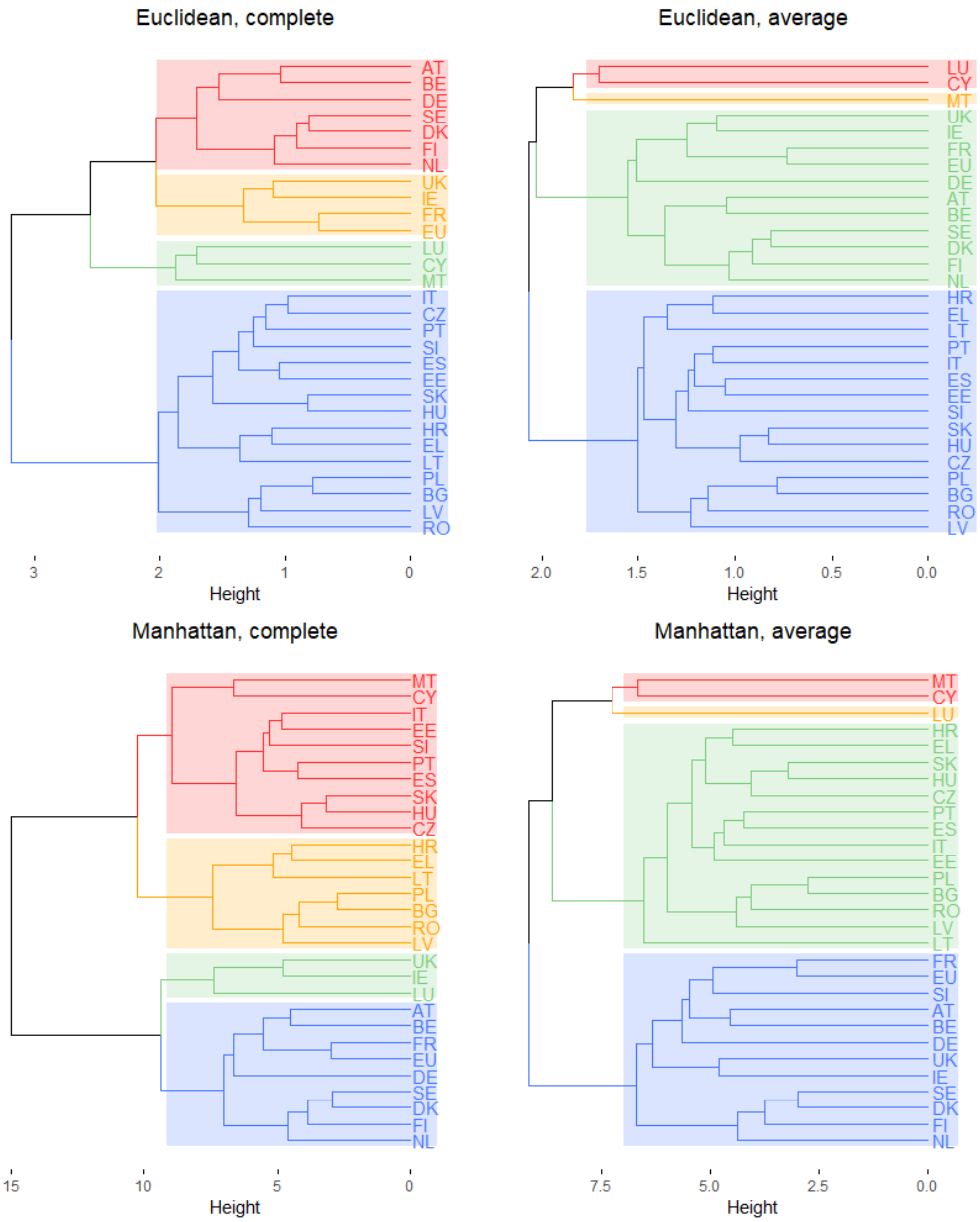
Average linkage in both figures (Figure 5.2. and Figure 5.3.) implies segmentation into three clusters, where we have two relatively big clusters and one small including only Malta, Cyprus & Luxemburg for both distance measures, small reminder - these

countries were earlier indicated as deviating by the single linkage method. We should bear in mind that average linkage is a compromise between two outmost methods. Thus, conclusions drawn from average linkage should be seen as the most significant. Unfortunately, the segmentation of average linkage with more than three clusters is messy.

Another aim of our analysis is to challenge the status quo number of clusters ($k = 4$) on the alternative scheme. First, we present results obtained from applying four cluster partition on dendrogram derived in the initial analysis (Figure 5.4). Then, we examine differences in a partition with an increasing number of clusters (from one to six). Lastly, we try to apply one proposed rule for finding an optimal number of clusters and discuss results.

European Innovation Scoreboard 2018 works with four innovation performance groups. Figure 5.4 shows a comparison of all derived methods when we would follow the original number of cluster $k = 4$. As concluded before, the partition is better illustrated by complete linkage when the number of clusters is higher than three, this is true for both distance measures. However, for Manhattan distance the number of countries in each cluster is more evenly distributed than in the dendrogram using Euclidean distance. Consequently, we conclude that the best partition to four groups is made by dendrogram using Manhattan distance and complete linkage that can be found in the bottom left corner of Figure 5.4. This dendrogram is used for further analysis.

Figure 5.4 - Dendrogram comparison for $k = 4$



Source: authors own analysis based on the EIS data

Note: a manual how to read a dendrogram is provided in the beginning of this section

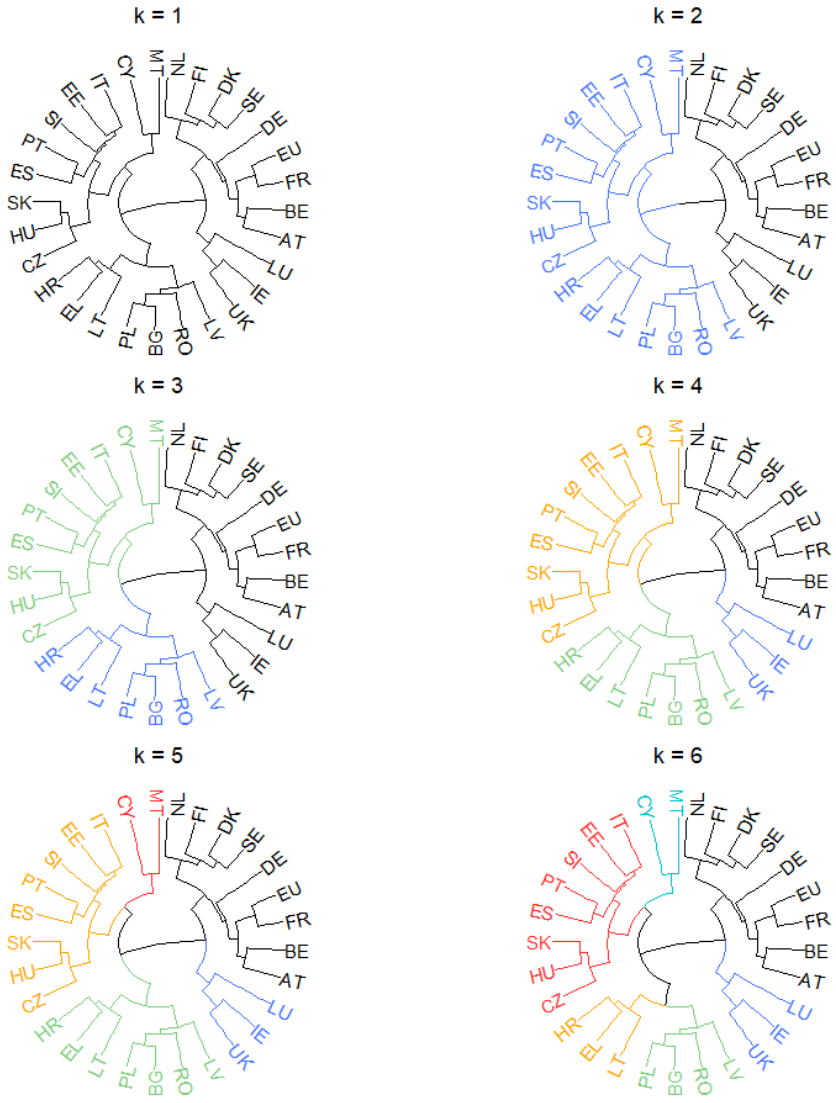
Figure 5.5 displays six circle dendrograms for an increasing number of clusters - k, starting with $k = 1$ and ending with $k = 6$ that is the highest number clusters where each cluster contains at least two EU Member States. Interestingly, the comparison has an apparent geographical pattern of countries across clusters. Since first division ($k = 2$) which could be described as splitting on the North-West part of the EU and the South-East part. Then in the next step, the South-East cluster splits to the Balkan-Baltic

countries (without Estonia) plus Poland in cluster that is labelled by green colour in dendrogram with $k = 3$ and the rest of the South-East countries including countries from, the South-Western part of the EU (Spain, Portugal, Italy), three quarters of the Visegrad Group (without Poland), the island states in the Mediterranean Sea (Malta and Cyprus) plus Slovenia and Estonia.

In each of two following steps one small group of cluster separate from its original cluster. Finally, in the last step ($k = 6$), where we will use cluster colours from the last dendrogram for clarity, we can see that the original "North-West" cluster (the black one) was split only once by separation of the British Isles and Luxembourg - the blue cluster - otherwise the North-West countries remains stable in one cluster. For that reason, these countries indicate a similar structure of their innovation performance. On the other hand, the "South-East" cluster split in four smaller clusters.

The red cluster remains the largest remaining cluster from the South-East branch, including Spain, Portugal, Italy, the Czech Republic, Slovakia, Hungary, Estonia and Slovenia. In contrast, the smallest cluster (turquoise colour) contains only Malta and Cyprus, both island states in the Mediterranean Sea, among them is strong geographical similarity, thus it makes sense that they have similar patterns in innovation performance. The light-green cluster contains of countries that are in the long term low-performing in innovations (*EIS 2018* and older versions), namely Poland, Romania, Bulgaria and Latvia. The last-torn golden cluster consists of Croatia, Lithuania and Greece. Their innovation performance is under average but better than countries from the green cluster. Moreover, we cannot observe any geographical pattern in this cluster. In conclusion, this analysis shows that segmentation of the EU Member States in innovation performance is reasonable, when the considered number of clusters lies between two to six. Then, it depends on what conclusions are drawn from the partition.

Figure 5.5 - Dendrogram comparison for increasing k



Source: authors own analysis based on the EIS data

Note: dendrogram using Manhattan distance and complete linkage is illustrated;
a manual how to read a dendrogram is provided in the beginning of this section

Generally speaking, hierarchical clustering is not a technique constructed for determining the optimal number of clusters. However, there exist stopping rules for cutting the tree that attempt to elect an optimal number of clusters. We will show one well-known method so-called Mojena's rule (Mojena, 1977) that is used also in recent literature (Szymańska & Zalewska, 2018). The Mojena's rule uses the following formula:

$$\hat{d}_{j+1} > mean(d) + k \cdot sd(d), \tag{5.1}$$

where $\hat{d}_{j+1}$ represents the value of criterion in stage $j+1$, $mean(d)$ and $sd(d)$ are, respectively, mean and standard deviation of the distance matrix of the SII. According to Milligan & Cooper (1985), we assume that k in (5.1) equals 1.25. Then, we compute the mean and the standard deviation of the SII dataset and obtain for Euclidean distance

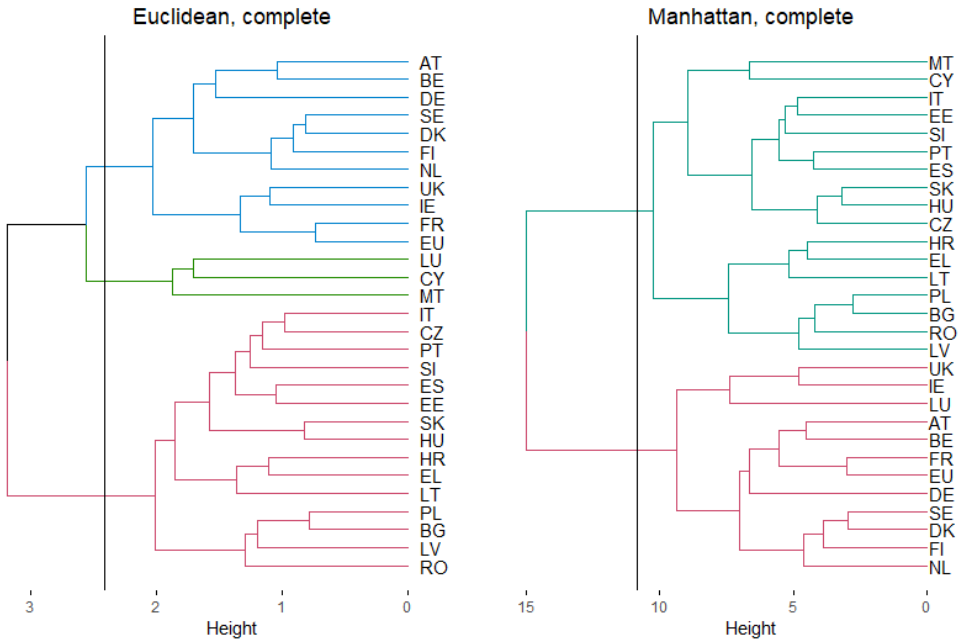
$$\hat{d}_{j+1}{}^{E} > 1.8 + 1.25 \cdot 0.5 \sim 2.4, \tag{5.2}$$

and analogically for Manhattan distance

$$\hat{d}_{j+1}{}^{M} > 7.77 + 1.25 \cdot 2.45 \sim 10.82. \tag{5.3}$$

When we apply Mojena's stopping rule on our data, results are not illustrative or produce any unique solution, as we can see in Figure 5.6. enclosed below.

Figure 5.6 - Mojena's rule



Source: authors own analysis based on the EIS data

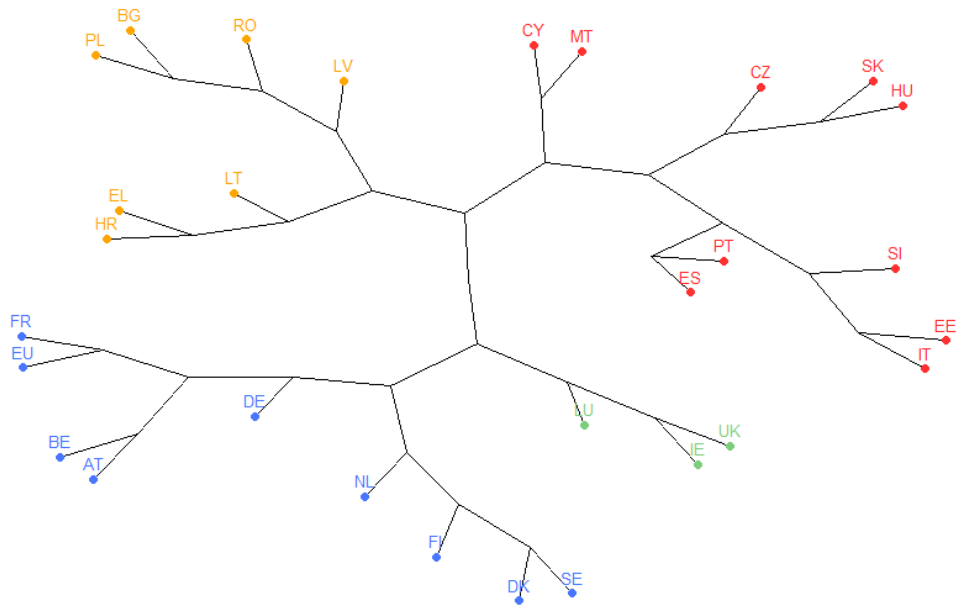
Note: a manual how to read a dendrogram is provided in the beginning of this section

To sum up, we set a goal to develop an alternative scheme of the partition of EU Member States using hierarchical clustering. This goal has been successfully achieved. In the initial stage, we presented and compared possible outcomes (under before established framework). Using a single linkage, we discovered four countries (Cyprus, Lithuania, Luxembourg, Malta) whose innovation performance differs significantly from the rest of the EU. The distinction of these countries was validated by both dissimilarity measures. This finding is not represented in the Summary Innovation Index.

In the next phase, we study the number of clusters. From analysis using the status quo number of clusters ($k = 4$), we concluded that the dendrogram using Manhattan distance and complete linkage provides the most suitable partition. Therefore, we further examined how the partition for this dendrogram differ for an increasing number of clusters. This analysis revealed that a number of clusters between three and seven is a feasible outcome. Furthermore, we applied Mojena's stopping rules on dendrograms, but it resulted with rather small number of clusters, which were too large and thus does not provide a reasonable outcome.

In conclusion, the dendrogram with Manhattan distance and complete linkage performed best in all our analyses. Therefore, we decided to include another graphical representation of it - a phylogenetic tree (Figure 5.7). The phylogenetic tree does not provide any further information about the partition of observations, but in our case, it provides an understandable graphics. We consider Figure 5.7 as the most interpretable output of our clustering analysis that summarizes individual groups showing not only the individual groups but also the closeness of countries within each group. Furthermore, we implemented partition into four clusters in Figure 5.7, to be able to compare final outcome to the existing EIS partition.

Figure 5.7 - Phylogenetic tree - Manhattan, complete, $k = 4$



Source: authors own analysis based on the EIS data
Note: a manual how to read a dendrogram is provided in the beginning of this section

The comparison of our alternative scheme to the status quo can be done by comparing Figure 5.7 to Table 2.1. Note that, using hierarchical clustering, we are unable to rank individual clusters among each other. Since hierarchical clustering does not know how to compare clusters. Therefore, we use neutral labelling using colours from Figure 5.7 to name the clusters of the alternative scheme.

At first glance, we can conclude one significant pattern. All countries, described as Innovation Leaders or Strong Innovators by the *EIS 2018,* are in a green or blue cluster

of our scheme. Analogically, countries from categories Moderate Innovators and Modest Innovators are all in a red or yellow cluster. Except one anomaly, Slovenia, that is a Strong Innovator in red cluster. Another conclusion can be drawn from the fact that the innovation performance of all countries from yellow and red cluster is below the European average. On the other hand, all countries from blue and green cluster have above average innovation performance. Consequently, we can conclude that the general pattern of both schemes is very similar.

In addition to the EIS, our scheme shows how countries are sorted among branches. As a result, for a specific country it can be easily recognized to which countries it is most similar. For example, the green cluster illustrates the diversity of these countries (Ireland, the United Kingdom and Luxembourg) from other groups. However, it is also shown that in the comparison with other clusters, these countries are relatively closer to the blue cluster. Moreover, many geographical patterns can be found at the two-in-branch level, e.g. Spain and Portugal, Malta and Cyprus or Slovakia and Hungary.

Furthermore, there exists a pattern that relates the year of country's entrance into the EU to its performance group. Consider following situation: when we would divide countries into two groups by the year of accession into the EU by the year 2000. All countries that have entered the EU after 2000 are either in the red or yellow cluster. Analogically, countries that have been in the EU before the year 2000, they are all in either blue or green clusters. The only exception being a group of four southern European countries (Portugal, Spain, Italy and Greece). According to this division, we can conclude that these four countries are lagging behind in the innovation performance.

To sum up, we stress out two conclusions from our alternative scheme. Firstly, it generally confirms the segmentation of the EU Member States by the innovation performance as suggested by the *EIS 2018.* As the main patterns of both schemes are very similar. Secondly, our scheme extends the EIS segmentation with insight mainly from sublevels of the branches, e.g. the similarity of the Nordic countries or the similarity of Mediterranean island countries. The main distinctive outcome is the differentiation of the British Isles from the rest of above average performing countries.

5.2 Regression

In this section, we attempt to widen the existing literature with estimation of an econometric model using the EIS dataset in order to test hypothesis of innovation's effects on productivity in the first half of the second decade of the 21st century.

A serious issue that is present in the EIS dataset is the high level of correlation among majority of variables included. When we were doing an explanatory analysis of our dataset, we constructed the correlation matrix that can be found in Appendix (Figure 5.7). Later on, we computed the variance-inflation factor. Those results can be found in Table 5.2, values that are above (or equal) our threshold set to 10 are displayed in red colour. The VIF analysis resulted in ten indicators above our threshold. The *EIS 2018* consists of ten innovation dimensions, and at least one VIF value over the threshold has been found in eight of them. Based on the correlation matrix and the VIF prove the existence of multicollinearity in our dataset.

The aim of regression analysis is that we want to estimate the impact of innovation on productivity. In our model, we use the EIS indicators as an independent variables and the real labour productivity per person (RLP.pp) as a dependent variable. We want to verify the existing findings (Griliches, 1979) that suggest a positive effect of innovation on productivity. Therefore, coefficients that are positive and significant confirm the previous research. Previously, we have established three panel data approaches, i.e. pooled OLS, first difference (FD) and fixed effects (FE). An overview of empirical results from these estimations is presented in the next section.

Table 5.2 - Variance-inflation factor

Variable	VIF value
doc_grad	4,1
ter_edu	4,3
lif_lea	16,3
sci_pub	20
TOP_pub	13,4
for_doc	9,8
bro_pen	4,9
opp_ent	10
PERD	9,9
VC	3,3
BERD	11,6
noRD_ex	2,5
IT_tr	4,4
SME_pro	24,1
SME_org	8,4
SME_inh	21,8
SME_col	5,5
pp_pub	19,1
pr_PERD	4,3
pat_app	15,6
tra_app	9,2
des_app	5,8
emp_act	16,3
emp_sec	2,8
tech_exp	3,1
ser_exp	6,8
sal_inn	2,3

Source: authors own analysis based on the EIS data

First, we present some characteristics that all estimation results have in common, then we summarize them individually. The majority of independent variables (i.e. the EIS indicators) turns up to be insignificant, when we recorded a maximum of five significant variables across all panel methods. Importantly, a robustness check is done for all the presented results. The maximum number of observations is 156, i.e. 26 countries (the EU Member States without Greece and Malta) across six time periods (2010-2015). However, with the FD method we lost the first year of observations. Unfortunately, significant variables differs among panel data methods, with the only sci_pub being significant for both FE and FD methods. All presented regression results are summarized in Table 5.3.

In fixed effects estimation, we have considered fixed effects estimation with "two-way" effect, i.e. regarding both individual and time dimensions. Fixed effects estimation indicates just four significant variables (sci_pub, BERD, SME_col and des_app), from which the latter two have positive sign. That means that these variables have a positive effect on labour productivity. What is more, the majority of the EIS indicators have a negative sign, thus suggest a negative impact on labour productivity. However, most of them are insignificant, whereas half of the positive is significant.

From Table 5.3 we can see that in the case of the FD estimation over a half of coefficients is positive. However, from all 27 indicators only there are only four significant variables (sci_pub, SME_org, pp_pub and pat_app), with the intercept being significant as well. If we wanted to answer our research question by the first-difference estimation, we would note that only SME_org and pp_pub confirm the positive impact on productivity. Regarding pooled OLS estimation, we can find ten positive coefficients in Table 5.3. Three of them are significant (doc_grad, ter_edu and ser_exp) and another significant variable being lif_lea as well as the intercept.

To select the most suitable model we incorporate an econometric test into our analysis. Intuitively, as also we have mentioned earlier, the fixed effect estimation should be the most appropriate method in our case. Pooled OLS, in contrast with fixed effects, assumes there are no unique effects either of individuals or across time. However, this is not true in our case. To prove this intuitive conclusion econometrically, we run a test to compare pooled OLS and fixed effects. Croissant & Millo (2008) recommend F-test with H_0: $no\ fixed\ effects$. When we carry out such test on our

model, we obtain p-value close to zero and thus we reject the null hypothesis. Consequently, fixed effects estimation should be preferred over pooled OLS, as we have intuitively suggested.

To provide a comparison between fixed effects and first difference, we execute Wooldridge's first-difference test for serial correlation in panel data and test two hypotheses. Firstly, we test serial correlation in the FD estimation with H_0: $no\ serial\ correlation\ in\ differenced\ errors$, when we receive a p-value equal to 0.002 resulting in rejecting the null hypothesis. Secondly, the test of serial correlation in fixed effect estimation with H_0: $no\ serial\ correlation\ in\ original\ errors$ returns p-value close to zero and therefore we can reject also this null hypothesis. Such a result (rejecting both hypotheses) suggests that the true lies somewhere in the middle and the robust covariance estimators should be used. However, we have had a priori stated usage of such approach.

Table 5.3 - RLP.pp regression results

Regression Results *Dependent variable:* RLP.pp									
Models:	**FE**	**FD**	**pooled OLS**	**lasso**	Models:	**FE**	**FD**	**pooled OLS**	**lasso**
First half of variables (1/2)					**Second half of variables (2/2)**				
Intercept		2.290***	98.395***	0.54	**SME_pro**	-0.101	-0.125	-0.084	.
(SE)		(0.595)	(4.197)			(0.179)	(0.110)	(0.108)	
doc_grad	-0.268	0.014	3.019***	.	**SME_org**	0.031	0.073*	-0.072	-0.03
	(1.152)	(0.642)	(1.138)			(0.082)	(0.043)	(0.101)	
ter_edu	-0.303	-0.162	0.219***	0.14	**SME_inh**	0.015	0.067	-0.047	.
	(0.821)	(0.453)	(0.077)			(0.125)	(0.085)	(0.111)	
lif_lea	0.171	0.030	-0.336*	.	**SME_col**	0.334*	0.078	-0.073	.
	(0.393)	(0.213)	(0.172)			(0.194)	(0.163)	(0.101)	
sci_pub	-0.015***	-0.013***	0.003	0.01	**pp_pub**	-0.019	0.046*	-0.036	.
	(0.005)	(0.005)	(0.003)			(0.046)	(0.023)	(0.041)	
TOP_pub	-0.313	-0.236	-0.301	.	**pr_PERD**	17.982	-8.718	-8.651	.
	(0.596)	(0.308)	(0.378)			(58.216)	(25.336)	(19.300)	
for_doc	-0.299	-0.145	-0.037	-0.20	**pat_app**	0.063	-1.032*	-0.296	.
	(0.204)	(0.103)	(0.063)			(0.855)	(0.576)	(0.315)	
bro_pen	-0.042	0.133	0.105	0.27	**tra_app**	-0.630	-0.230	-0.179	.
	(0.299)	(0.212)	(0.103)			(0.428)	(0.273)	(0.144)	
opp_ent	0.328	0.230	0.222	.	**des_app**	0.507**	0.156	0.174	.
	(0.317)	(0.220)	(0.388)			(0.209)	(0.142)	(0.246)	
PERD	-3.179	-2.079	0.494	.	**emp_act**	1.430	0.104	0.124	0.20
	(6.442)	(5.789)	(5.778)			(0.927)	(0.558)	(0.282)	
VC	-17.834	3.423	3.223	.	**emp_sec**	-0.263	-0.475	-0.359	.
	(13.882)	(5.085)	(10.658)			(0.432)	(0.498)	(0.219)	
BERD	-8.161*	-5.655	0.889	.	**tech_exp**	0.204	0.090	-0.022	.
	(4.377)	(4.282)	(1.362)			(0.133)	(0.077)	(0.037)	
noRD_ex	-0.087	-0.526	0.688	.	**ser_exp**	-0.227	-0.145	0.099**	.
	(1.147)	(0.800)	(0.980)			(0.157)	(0.116)	(0.039)	
IT_tr	-0.029	0.110	-0.056	.	**sal_inn**	-0.007	0.024	0.008	.
	(0.118)	(0.135)	(0.067)			(0.133)	(0.105)	(0.145)	
$N =$	156	130	156	130		156	130	156	130
R^2	0.35	0.20	0.47	-		0.35	0.20	0.47	-
Adj. R^2	-0.03	-0.01	0.36	-		-0.03	-0.01	0.36	-

Source: authors own analysis based on the EIS data

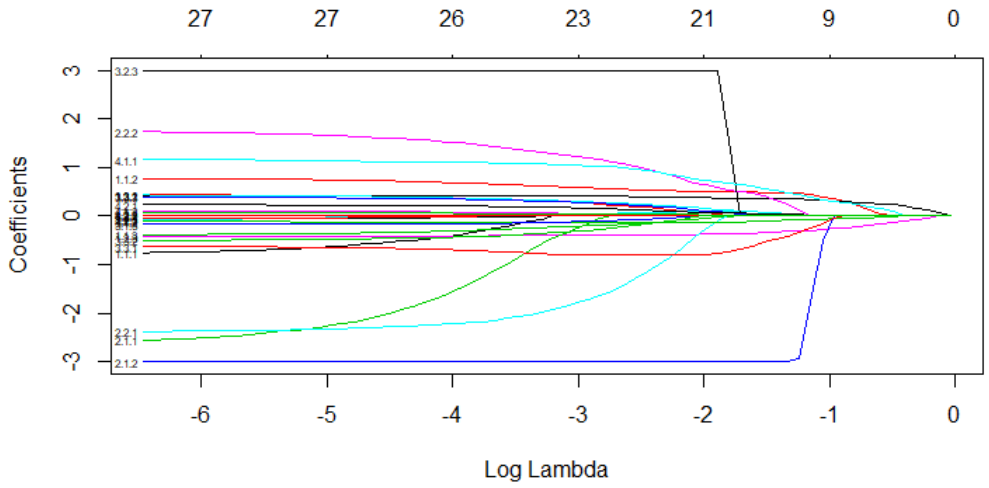
*Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$*

To sum up, all of the panel data methods fail to sufficiently answer our hypothesis about the positive effects of innovation on productivity. As a consequence, we decided to apply shrinkage methods of supervised learning for a variable selection. Furthermore, we have manually adjusted (i.e. time-demeaned) dataset to control the panel data structure, in order to be able to estimate our model.

When proceeding the lasso estimation, we can display so-called lasso path (i.e. graphical representation showing the entrance of variables into the model. Figure 5.8

illustrates a lasso path for our case. In *Figure 5.8,* y-axis represents the value of individual coefficients, the bottom x-axis is log-value of the λ-penalty (sometimes called L1 norm in case of lasso), whereas top x-axis labels a number of variables that are present in the model for given λ. Please pay attention to the fact that one must read these graphs from right to left. Then we can see that a few variables enter the model in the beginning. In order to make Figure 5.8 more illustrative, we set an upper and lower limits equal to 3 and -3 respectively, such values were selected in order to limit only one variable on each side with a view not to disorder the resulting model. Setting these limits have undetectable effect on the resulting estimation, therefore we use this limited version of model in further analysis.

Figure 5.8 - RLP.pp model lasso path



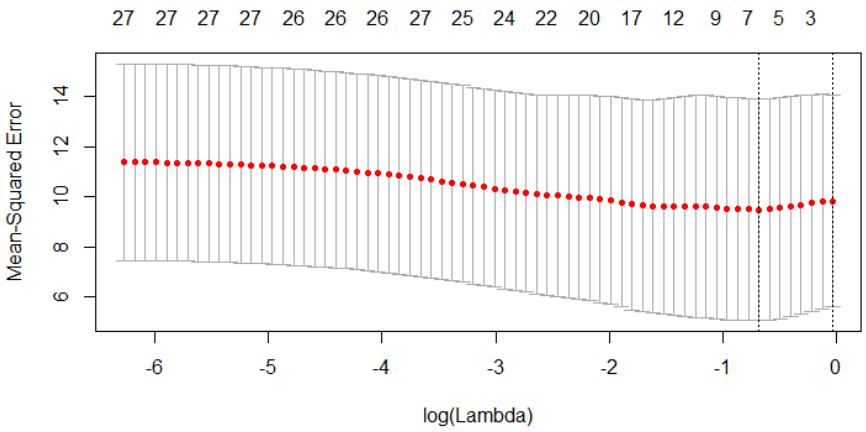
Source: authors own analysis based on the EIS data

To be able to select the optimal model, we expand our analysis with k-fold cross-validation. Cross-validation in *glmnet* package is done by *cv.glmnet()* function that returns a plot including the cross-validation curve (red dotted line), and upper and lower standard deviation curves along the λ sequence (error bars). Concerning the selection of suitable k, we observed several possibilities, hence all options returned very similar results, we thus decided to use default number of folds $k = 10$.

Figure 5.9. displays resulting cross-validation plot, which illustrates the cross-validation curve (red dotted line), every dot has upper and lower standard deviation displayed as a whisker. On top of that, the cross-validation plot provides two vertical-

dotted lines for two values of parameter λ. Firstly, `lambda.min` is the value of λ that gives minimum mean cross-validated error and is illustrated in Figure 5.9. by the left dotted line. Secondly, `lambda.1se` gives the most regularized model such that the error is within one standard error of the minimum (right dotted line).

Figure 5.9 - RLP.pp model cross-validation



Source: authors own analysis based on the EIS data

From the top x-axis we can easily see that `lambda.1se` model includes three or fewer variables, and `lambda.min` model includes between five to seven variables. In fact, using `lambda.1se` only the intercept is included, which is an unsuitable outcome. Hence, for the lasso estimation we use `lambda.min` model with penalty $\lambda = 0.503$ that includes seven variables. Results of the lasso estimation for `lambda.min` are reported also in Table 5.3 where dot (.) means a shrunk variable.

To conclude, the lasso estimation selects six EIS indicators (ter_edu, sci_pub, for_doc, bro_pen, SME_org and emp_act), from which only two have negative coefficients (for_doc and SME_org). Not only that the lasso provides an appropriate variable selection, the lasso model with penalty term `lambda.min`, when $\lambda = 0.503$, address our research question the best. However, given that our lasso findings are not based on a proper review of literature (due to the thesis's limited scope), the results from such analysis should, therefore, be interpreted with considerable caution. Thus, the lasso estimation could be seen as a possibility of improving the poor preliminary results.

6 Conclusion

In this thesis, we analyse the innovation indicators in the EU, we follow the *European Innovation Scoreboard 2018* that contains a comprehensive analysis of the innovation performance of EU Member States and attempts to the extend existing literature in two directions.

Firstly, we wanted to develop an alternative scheme of the partition of the EU Member States according to the innovation performance. We have achieved that goal using hierarchical clustering on the EIS dataset. From our analysis, we have selected one dendrogram as a final outcome. Then, we compared our results to the status quo. From the comparison of schemes, we concluded that the general patterns are very similar. Thus, the alternatively proposed scheme validates the existing partition of the EU Member States according to the innovation performance. The main distinctive outcome between schemes is the differentiation of the British Isles from the rest of above average performing countries in our scheme. In addition to the EIS, our scheme provides a valuable insight about the within-cluster similarities. We have discovered, e.g., the similarity of Finland, Sweden and Denmark and their relative distinction from France, although they belong to one cluster. Moreover, from analysis of year of accession into the EU, we have concluded that Italy, Spain, Portugal and Greece are lagging behind the other countries that have entered the EU before the year 2000.

Secondly, we attempt to test a hypothesis concerning effects of innovations on productivity, when we wanted to verify findings of previous researchers (e.g. Griliches, 1979) that suggest a positive effect of innovation on productivity. According to the data structure, we have provided an estimation of three methods - pooled OLS, fixed effects and first-difference. Unfortunately, the empirical findings were ambiguous and differed considerably between the applied methods. We found only a weak evidence to support our hypothesis, since the vast majority of our variables was insignificant and often even with a negative sign. Therefore, we introduced an additional method of penalized regression, the lasso, to perform a variable selection. The lasso estimation have selected six EIS indicators (ter_edu, sci_pub, for_doc, bro_pen, SME_org and emp_act), most of which had a positive coefficient (except for for_doc and SME_org). Thus, the lasso

estimation could be seen as a suggestion of a possible solution improving the poor preliminary results.

Last but not least, we have several propositions for further research. The first possible extension is to broaden the clustering analysis by introducing other clustering methods, e.g. k-means or Principal Component Analysis. Secondly, the scope of analysis could be broaden by studying the Global Innovation Index (*The Global Innovation Index 2018*) that provides detailed metrics about the innovation performance of 126 countries. Concerning the regression part, further research could use more penalized regression methods or some other methods to improve original regression results.

7 Bibliography

Baltagi, B. H. (2005). *Econometric analysis of panel data* (Vol. 3). Wiley New York.

Bilbao-Osorio, B., & Rodríguez-Pose, A. (2004). From R&D to innovation and economic growth in the EU. *Growth and Change*, *35*(4), 434–455.

Castellani, D., Piva, M., Schubert, T., & Vivarelli, M. (2016). *The Productivity Impact of R&D Investment: A Comparison between the EU and the US*. Retrieved from https://papers.ssrn.com/abstract=2786021

Commission Staff Working Paper: 2001 Innovation Scoreboard. (2001, September 14). Commission of the European Communities.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43.

Edquist, C., Zabala-Iturriagagoitia, J. M., Barbero, J., & Zofío, J. L. (2018). On the meaning of innovation performance: Is the synthetic indicator of the Innovation Union Scoreboard flawed? *Research Evaluation*, *27*(3), 196–211. https://doi.org/10.1093/reseval/rvy011

European Innovation Scoreboard 2018. (2018). European Commission.

European Innovation Scoreboard 2018: Methodology Report. (2018, June 15). European Commission.

Fortmann-Roe, S. (2012, June). *Understanding the Bias-Variance Tradeoff*. Retrieved from http://scott.fortmann-roe.com/docs/BiasVariance.html#fn:1

Griliches, Z. (1979). Issues in assessing the contribution of research and development to productivity growth. *Bell Journal of Economics*, *10*(1), 92–116.

Guellec, D., & Van Pottelsberghe de la Potterie, B. (2004). From R&D to productivity growth: Do the institutional settings and the source of funds of R&D matter? *Oxford Bulletin of Economics and Statistics*, *66*(3), 353–378.

Hall, B. H. (2011). *Innovation and productivity*. National bureau of economic research.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

Innovation in a knowledge-driven economy, Communication from the Commission to the Council and the European Parliament. (2000, September 20). Commission of the European Communities.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kuhlman, C., Ramamurthy, K. N., Sattigeri, P., Lozano, A. C., Cao, L., Reddy, C., … Varshney, K. R. (2017). How to foster innovation: A data-driven approach to measuring economic competitiveness. *IBM Journal of Research and Development*, *61*(6), 11:1-11:12. https://doi.org/10.1147/JRD.2017.2741820

Mecke, K. R., & Stoyan, D. (2008). *Statistical Physics and Spatial Statistics: The Art of Analyzing and Modeling Spatial Structures and Pattern Formation*. Springer.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179. https://doi.org/10.1007/BF02294245

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal*, *20*(4), 359–363. https://doi.org/10.1093/comjnl/20.4.359

Pełka, M. (2018). Analysis of Innovations in the European Union Via Ensemble Symbolic Density Clustering. *Econometrics*, *22*(3). Retrieved from https://content.sciendo.com/view/journals/eada/22/3/article-p84.xml

Sajeva, M., Gatelli, D., Tarantola, S., & Hollanders, H. (2005). *Methodology Report on European Innovation Scoreboard 2005*.

Schibany, A., & Streicher, G. (2008). The European Innovation Scoreboard: drowning by numbers? *Science and Public Policy*, *35*(10), 717–732.

Szymańska, A., & Zalewska, E. (2018). Towards the Goals of the Europe 2020 Strategy: Convergence or Divergence of the European Union Countries? *Comparative Economic Research*, *21*(1), 67–82. https://doi.org/10.2478/cer-2018-0004

The Global Innovation Index 2018: Energizing the World with Innovation. (n.d.). Cornell University, INSEAD, and WIPO (2018).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tools for Composite Indicators Building. (2005). European Commission.

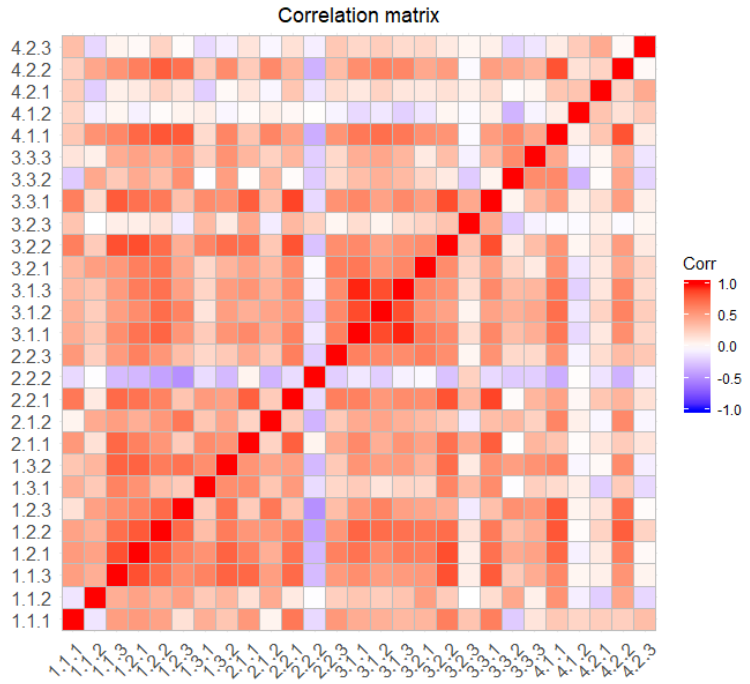
Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach, Fifth edition*. OH: South-Western.

8 Appendix

Table 8.1 - List of abbreviations used

EC	European Commission
EIS	European Innovation Scoreboard
EU	European Union
JRC	Joint Research Centre
lasso	Least absolute shrinkage and selection operator
R&D	Research & Development
SII	Summary Innovation Index
UNU-MERIT	United Nations University - Maastricht Economic and Social Research Institute on Innovation and Technology
WEF	World Economic Forum

Figure 8.1 - The correlation matrix



Source - authors own analysis based on the EIS data

Table 8.2 - List of labels and codes of the EU Member States

BE	Belgium
BG	Bulgaria
CZ	Czechia
DK	Denmark
DE	Germany
EE	Estonia
IE	Ireland
EL	Greece
ES	Spain
FR	France
HR	Croatia
IT	Italy
CY	Cyprus
LV	Latvia
LT	Lithuania
LU	Luxembourg
HU	Hungary
MT	Malta
NL	Netherlands
AT	Austria
PL	Poland
PT	Portugal
RO	Romania
SI	Slovenia
SK	Slovakia
FI	Finland
SE	Sweden
UK	United Kingdom

Source: Eurostat

Table 8.3 - The Summary Innovation Index score 2018

Country	Value
Austria	0.58
Belgium	0.59
Bulgaria	0.23
Croatia	0.26
Cyprus	0.39
Czech Republic	0.42
Denmark	0.67
Estonia	0.4
EU Average	0.5
Finland	0.65
France	0.55
Germany	0.6
Greece	0.33
Hungary	0.33
Ireland	0.58
Italy	0.37
Latvia	0.29
Lithuania	0.36
Luxembourg	0.61
Malta	0.4
Netherlands	0.65
Poland	0.27
Portugal	0.41
Romania	0.16
Slovakia	0.32
Slovenia	0.47
Spain	0.4
Sweden	0.71
United Kingdom	0.61

Source: European Innovation Scoreboard 2018