
The Evolution of Immigrant Groups in Luxembourg What Are the Different Pathways in the Labour Market?

Catarina Campos de Melo Sousa Silva

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:
Professor Paula Brito
Professor Pedro Campos

2023

Acknowledgments

I would like to start by thanking my parents for constantly supporting me, helping me to achieve my goals, allowing me to open my horizons and have new experiences both academically and personally and making me excellent proposals to relax when I am most stressed.

I would also like to thank D. Manuela, who throughout my journey has always been present and twice a week makes my study/dissertation writing moments sweeter.

Furthermore, I want to thank Professor Paula Brito, who introduced me to the Symbolic Data Analysis field, the basis of this dissertation and who, together with Professor Pedro Campos, gave me essential tips and guided me through the whole process. In addition, I want to thank Statec, especially Anne, Marco and Johann for making me feel so welcome and giving me great support throughout the internship.

Lastly, I would like to thank my family, friends and teachers who have been present in my academic journey and who have always helped me to improve and in one way or another taught me things that I will take with me for life.

Abbreviations

ARI	Adjusted Rand Index
EFTA	European Free Trade Association
EMOS	European Master in Official Statistics
EU	European Union
EU-LFS	European Union Labour Force Survey
EU-27	European Union as of the 1 st of February 2020 (- United Kingdom)
EU-28	European Union as of the 1 st of July 2013 (+ Croatia)
FOI	First-order individuals
GEG	Gender Employment Gap
GHG	Gender Hours Gap
GOEG	Gender Overall Earnings Gap
GPG	Gender Pay Gap
HINoV	Heuristic Identification of Noisy Variables
ILO	International Labour Organization
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
LFS	Labour Force Survey
MEC	Monitoring the Evolution of Clusters
NACE	Nomenclature statistique des activités économiques dans la Communauté européenne
NC	Neighbouring Countries (Belgium, France and Germany)
NSI	National Statistical Institute
NUTS	Nomenclature of territorial units for statistics
RI	Rand Index
SDA	Symbolic Data Analysis
SOI	Second-order individuals
Statec	National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg
UN	United Nations

Abstract

Luxembourg is distinguished by its demographic dynamism among European countries. From 2010 until 2020, around 80% of the population growth was due to migration. In general, the main reason why people immigrate to Luxembourg is labour.

Taking into account the phenomenon of immigration, the aim of this work is to profile different groups of immigrants in the labour market, through time, combining data from the Labour Force Survey (LFS) of Luxembourg with Symbolic Data Analysis (SDA) techniques.

Briefly, SDA is used by aggregating micro-data of Luxembourgish's LFS. The symbolic objects are formed based on the place of birth (Luxembourg, European Union (EU) country, Neighbouring country (NC), Portugal, non-EU country) and length of residence in Luxembourg, making a total of 21 objects for analysis. After the aggregation, the HINoV algorithm for variable selection is applied and then cluster analysis is used. This is done for each year from 2014 to 2022. Since time is an important factor a Cluster Monitoring technique is used.

The results show that by using just six variables it is possible to break down the 21 objects into groups with similar labour market profiles. Furthermore, it is also possible to verify that people who were born in a NC or in an EU country (excluding the NC, Portugal and Luxembourg) have similar profiles in the labour market as they are highly educated, work mainly in white-collar professions, are among the highest earners in society and are allowed to do telework. On the opposite side, even though some changes may be seen recently, the Portuguese people are low educated, mainly work in blue-collar professions, are among the lowest earners and are not able to do telework. Concerning the Luxembourgers, they have medium levels of education, mainly work in white-collar professions and do not do much telework. Regarding the non-EU people, it is quite challenging to profile them since the non-EU objects include people that were born in countries with very different states of development. Lastly, regardless of the place of birth, working in Business services is common among recently arrived people and for long-established people is common to work in Public administration, defence, education, health and social work. In addition, as time of residence passes, people start to leave the cities and move to places with a different degree of urbanization.

A major part of the present work was developed through an internship at the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (Statec), at the 'Marché du travail et Education' unit, under the European Master in Official Statistics (EMOS) programme.

Keywords: clustering, immigration, labour force survey, Luxembourg, symbolic data analysis, monitoring

Resumo

O Luxemburgo distingue-se entre os países Europeus pelo seu dinamismo demográfico. De 2010 até 2020, cerca de 80% do crescimento demográfico deveu-se à imigração. Em geral, a razão principal pela qual as pessoas imigram para o Luxemburgo é o mercado de trabalho.

Considerando o fenómeno de imigração, o principal objetivo deste trabalho é perfilar cada um dos diferentes grupos de imigrantes no mercado de trabalho, ao longo do tempo, combinando dados do Inquérito ao Emprego Luxemburguês com técnicas de Análise Simbólica de dados (SDA).

Resumidamente, SDA é utilizada, agregando dados do Inquérito ao Emprego Luxemburguês. Os objetos simbólicos são formados tendo em conta o local de nascimento (Luxemburgo, país da União Europeia (UE), país vizinho, Portugal, país fora da UE) e a duração de residência no Luxemburgo, perfazendo um total de 21 objetos para análise. Após a agregação, o algoritmo HINoV é utilizado para fazer seleção de variáveis e seguidamente técnicas de *clustering* são aplicadas. Este processo é feito para cada ano entre 2014 e 2022. Uma vez que o tempo é um fator importante, a evolução dos *clusters* é monitorizada.

Os resultados mostram que com apenas seis variáveis é possível dividir os 21 objetos em grupos cujos perfis no mercado de trabalho são semelhantes. Além disso, é possível verificar que as pessoas nascidas nos países vizinhos ou num país da EU (excluindo os países vizinhos, Portugal e Luxemburgo) têm perfis semelhantes no mercado de trabalho. Estes têm níveis de educação elevados, trabalham essencialmente em profissões de colarinho branco, estão entre as pessoas que mais ganham na sociedade e têm bastante facilidade em fazer teletrabalho. No lado oposto, e apesar de recentemente se verificarem algumas mudanças, os Portugueses têm níveis baixos de educação, trabalham principalmente em profissões de colarinho azul, estão entre as pessoas com salários mais baixos e não realizam teletrabalho. Relativamente aos Luxemburgueses, estes têm níveis médios de educação, trabalham essencialmente em profissões de colarinho branco e não fazem muito teletrabalho. No que toca às pessoas de países fora da UE, é difícil identificar algum tipo de perfil uma vez que os objetos não-UE incluem pessoas que nasceram em países com variadíssimos estados de desenvolvimento. Por fim, independentemente do local de nascença, trabalhar no setor de Serviços empresariais é muito comum entre os recém-chegados e para os imigrantes estabelecidos há muito tempo é comum trabalhar na Administração pública, defesa, educação, saúde e serviços sociais. Adicionalmente, à medida que o tempo de residência passa, as pessoas começam a sair das cidades e movem-se para locais com outro grau de urbanização.

Uma grande parte do presente trabalho foi desenvolvida através de um estágio no Instituto Nacional de Estatística e Estudos Económicos do Grão-Ducado do Luxemburgo (Statec), na unidade ‘Marché du travail et Education’, no âmbito do programa European Master in Official Statistics (EMOS).

Palavras-Chave: análise de dados simbólicos, *clustering*, imigração, inquérito ao emprego, Luxemburgo, monitorização

Contents

1	Introduction	1
1.1	Motivation and Problem Description	1
1.2	Objectives	3
1.3	Official Statistics	3
1.4	Dissertation Structure	4
2	Migration and the Luxembourgish Labour Force	5
2.1	Europe Migration Flows	5
2.2	Luxembourg	7
2.2.1	Immigration Flows	7
2.2.2	Labour Market Conditions	8
2.3	Labour Force Survey	11
2.3.1	Sampling Strategies	14
3	Methodology	16
3.1	Symbolic Data Analysis	16
3.1.1	Types of Variables	17
3.1.2	Clustering	18
3.1.3	Determination of the Number of Clusters	23
3.1.4	Variable Selection	25
3.2	Cluster Monitoring	27
4	Data and Software	33
4.1	Software	33
4.2	Variables Description and Pre-Processing	33
4.2.1	Missing Values	40
4.3	Sample Description	41
4.4	Descriptive Analysis	41
5	Models	47
5.1	From Microdata to Symbolic Data	47
5.2	Algorithm Selection	48
5.3	Variable Selection	51

6	Analysis of Results	53
6.1	Cluster Description	53
6.2	Cluster Monitoring	67
7	Conclusions	74
7.1	Main conclusions	74
7.2	Final Considerations and Future Work	75
	Appendix	77
A		78
B		78
C		83
D		98
	Bibliography	130

List of Tables

3.1	Migration reasons by country of birth	17
3.2	Notation for comparing two partitions Adapted from Hubert and Arabie (1985)	26
4.1	Sample Characterization 2014-2022	41
4.2	Final Employees Sample Characterization 2014-2022	42
4.3	Population Characterization 2014-2021	43
5.1	Validity Indices, using all the 16 variables - 2022	49
5.2	Number of years each variable appeared among the most relevant variables using the HINoV algorithm	52
A.1	Exogenous transitions formal definitions Adapted from Oliveira (2010)	77
A.2	Endogenous transitions notation Adapted from Oliveira (2010)	77
D.1	Average weight of each category of a variable for the objects belonging to a cluster - 2014	98
D.2	Average weight of each category of a variable for the objects belonging to a cluster - 2015	99
D.3	Average weight of each category of a variable for the objects belonging to a cluster - 2016	100
D.4	Average weight of each category of a variable for the objects belonging to a cluster - 2017	101
D.5	Average weight of each category of a variable for the objects belonging to a cluster - 2018	102
D.6	Average weight of each category of a variable for the objects belonging to a cluster - 2019	103
D.7	Average weight of each category of a variable for the objects belonging to a cluster - 2020	104
D.8	Average weight of each category of a variable for the objects belonging to a cluster - 2021	105
D.9	Average weight of each category of a variable for the objects belonging to a cluster - 2022	106
D.10	Cluster Description - 2014	116
D.11	Cluster Description - 2015	117
D.12	Cluster Description - 2016	118

D.13 Cluster Description - 2017	119
D.14 Cluster Description - 2018	120
D.15 Cluster Description - 2019	121
D.16 Cluster Description - 2020	122
D.17 Cluster Description - 2021	123
D.18 Cluster Description - 2022	124
D.19 Test for equality of proportions of people doing telework in 2019 and 2020 . .	126
D.20 Test for equality of proportions of people doing telework in 2020 and 2022 . .	127
D.21 Clusters Composition 2014-2022	129

List of Figures

3.1	Cluster representation	19
3.2	MEC Bipartite Graph representation	28
3.3	Representation of some endogenous and exogenous transitions	31
3.4	Bipartite Graph applying the extension proposed for the MEC framework	32
4.1	Total Population Evolution 2014-2022	42
4.2	Immigrant groups total population evolution 2014-2022	42
4.3	Migration Reasons 2021	44
4.4	Unemployment Rate 2014-2022	45
4.5	Employment Rate 2014-2022	45
4.6	Age Density Plot per Group 2022	46
5.1	Extract of a symbolic data table for the year 2022 produced by the RSDA package	48
5.2	Validity Indices HCLUST, using all the 16 variables - 2022	50
6.1	Dendrogram 2014	54
6.2	Dendrogram 2015	54
6.3	Dendrogram 2016	55
6.4	Dendrogram 2017	55
6.5	Dendrogram 2018	56
6.6	Dendrogram 2019	57
6.7	Dendrogram 2020	57
6.8	Dendrogram 2021	58
6.9	Dendrogram 2022	58
6.10	Education attainment of recently arrived Portuguese immigrants (1.PT) 2014-2022	60
6.11	Percentage of Luxembourgers whose highest level of education attained is sec- ondary education 2014-2022	61
6.12	Education attainment level of people born in a non-EU country and living in Luxembourg for between 9 and 15 years (3.OUTEU) - 2021.	61
6.13	Percentage of recently arrived Portuguese immigrants (1.PT) working in white- collar high-skilled professions 2014-2022	62
6.14	Percentage of people living in each of the different types of urbanisation for the different NC groups - 2018	63

6.15	Degree of Urbanization across the among the recently arrived and longer established immigrants - 2022	64
6.16	Telework variation European Union community 2019 - 2022	65
6.17	Telework variation Neighbouring countries community 2019 - 2022	65
6.18	Telework variation Portuguese community 2019 - 2022	66
6.19	Telework variation Outside the European Union community 2019 - 2022	66
6.20	Telework variation Luxembourgish natives 2019 - 2022	67
6.21	Cluster Monitoring Graph 2014-2022	68
6.22	ISCO4D across $C_3(2020)$ objects 2020-2021	72
6.23	TELEARB across $C_4(2021)$ objects 2021-2022	73
B.1	Association between migration reasons and place of birth - χ^2 test results	78
B.2	Age Density Plot per Group 2014	78
B.3	Age Density Plot per Group 2015	79
B.4	Age Density Plot per Group 2016	79
B.5	Age Density Plot per Group 2017	80
B.6	Age Density Plot per Group 2018	80
B.7	Age Density Plot per Group 2019	81
B.8	Age Density Plot per Group 2020	81
B.9	Age Density Plot per Group 2021	82
C.1	TOPRI plot 2014 (all variables)	83
C.2	TOPRI plot 2015 (all variables)	84
C.3	TOPRI plot 2016 (all variables)	84
C.4	TOPRI plot 2017 (all variables)	85
C.5	TOPRI plot 2018 (all variables)	85
C.6	TOPRI plot 2019 (all variables)	86
C.7	TOPRI plot 2020 (all variables)	86
C.8	TOPRI plot 2021 (all variables)	87
C.9	TOPRI plot 2022 (all variables)	87
C.10	Final symbolic data table - 2014	89
C.11	Final symbolic data table - 2015	90
C.12	Final symbolic data table - 2016	91
C.13	Final symbolic data table - 2017	92
C.14	Final symbolic data table - 2018	93
C.15	Final symbolic data table - 2019	94
C.16	Final symbolic data table - 2020	95
C.17	Final symbolic data table - 2021	96
C.18	Final symbolic data table - 2022	97
D.1	Validity Indices - 2014	107
D.2	Validity Indices - 2015	108
D.3	Validity Indices - 2016	109
D.4	Validity Indices - 2017	110

D.5	Validity Indices - 2018	111
D.6	Validity Indices - 2019	112
D.7	Validity Indices - 2020	113
D.8	Validity Indices - 2021	114
D.9	Validity Indices - 2022	115
D.10	TOPRI plot 2020	125
D.11	TOPRI plot 2021	125
D.12	TOPRI plot 2022	126

Chapter 1

Introduction

In this chapter, we make an overview of the theme that is developed throughout the dissertation. First, the main motivations that led to the choice of the topic are described, followed by the enumeration of the objectives that this study intends to achieve and a brief contextualization of Official Statistics. It ends with the general structure of the work.

1.1 Motivation and Problem Description

Regardless of the reasons, whether voluntary or involuntary and the methods, whether regular or irregular, in the global context, a migrant is defined as a person who is outside the territory of the state of which he/she is a national or citizen and who has lived in a foreign country for more than one year. Despite this, the term ‘migrant’ lacks a specific legal definition. The United Nations (UN) recommendations for population censuses and data on international migration states that a change in the country of one’s ordinary residence must last at least a year in order to be considered ‘migration’ (Eurostat, 2022e).

Immigration to Luxembourg is not a recent phenomenon, it dates back to the late 19th century when the nation first began its industrialisation. The country’s migration is not only, but also related to the transformation of Luxembourg’s economy. In recent years, the percentage of people that were born in a foreign country surpasses 50%. Throughout all these years several waves of immigrants, detailed later on, have been observed. There are different reasons why people immigrate to Luxembourg such as the proximity of neighbouring countries, the existence of a migratory tradition, and family reunification, but the main factor is labour (Peltier & Klein, 2020).

Several works have been published, especially by the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (Statec), regarding the labour market conditions in the country, for example, Ametepe et al. (2022); Division sociale (2023); Reiff (2019); Rota and Larue (2021); Schockmel, Hartung, Reiff, and Neumayr (2022); Statec (2016). However, most of them are about the population as a whole. Moreover, when different immigrant groups are considered, publications generally only take one year into account and not a period of time which allows to have an idea of the evolution of the characteristics. Regarding the evolution of immigration, there is one major work published by Statec, Zahlen (2012), which analyses

the evolution from the 60s until 2011, however, it does not include the evolution of the labour market characteristics across groups.

According to Portail de l'emploi (2023), the Luxembourgish labour market is highly internationalised. This is easy to understand given that a significant percentage of the population is foreign (see later on) and that employment is the main reason why people immigrate to the country (Peltier & Klein, 2020). Considering this, understanding if there are differences in the job market between immigrant groups and natives is crucial. By comparing labour market characteristics, it is possible to identify disparities or gaps in employment opportunities and earnings. This information can help policymakers understand the factors contributing to these disparities, such as discrimination, differences in skills or education levels or barriers in accessing certain sectors. Furthermore, it may help policymakers identifying specific areas where immigrants may require additional support or integration programs, for example, language training, skills development, and recognition of foreign qualifications. Therefore, it is pretty relevant to carry out a study that makes a comparative analysis of the labour market characteristics across different immigrant groups and the natives. It is also important to take the time factor into account to see how things are changing over the years. This may help understanding if the adopted policies are having any effect.

To develop the mentioned study, this work uses data from the Labour Force Survey (LFS) (Eurostat, 2021c) together with Symbolic Data Analysis (SDA) (Billard & Diday, 2003, 2007; Bock & Diday, 2000; Brito, 2014; Diday & Noirhomme-Fraiture, 2008; Noirhomme-Fraiture & Brito, 2011) techniques. The LFS is a sample survey covering the population living in private households. In the present work, the LFS of Luxembourg is used. Therefore this work covers the private households in Luxembourg. The survey provides quarterly and yearly results on people of working age (15 years and older) who are participating in the labour force as well as on those who are outside it (Eurostat, 2022b). It is important to note that besides the national LFS, concerning each individual country, the European Union Labour Force Survey (EU-LFS) is a cross-sectional and longitudinal household sample survey, collated by Eurostat from data from the national LFSs provided by the EU member states, three European Free Trade Association (EFTA) countries and four EU-candidate countries. Therefore, the EU-LFS is a way of harmonising and comparing labour market data between the various countries of the European Union. Additionally, SDA is an extension of traditional data analysis where the resulting concepts are structured and entail inherent variability.

Briefly, SDA is used by aggregating the micro-data of the LFS. The symbolic objects are based on the duration of stay in Luxembourg in completed years and on the place of birth (Luxembourg, European Union (EU) country, Neighbouring country (NC), Portugal, non-EU country), making a total of 21 objects for the analysis. After the aggregation, the Heuristic identification of noisy variables (HINoV) algorithm is applied to select the most relevant variables. Then Cluster Analysis using those variables is applied to the aggregated data to form homogeneous groups of immigrants. The clustering algorithm was applied every year from 2014 to 2022 to a set of 21 symbolic objects that were described by the distributions of six categorical modal variables. These 21 objects correspond to groups in the population and not to individuals. Time is an important factor to see if something is changing. Therefore, the cluster's evolution through time is monitored using the Monitoring the Evolution of Clusters (MEC) approach (Oliveira & Gama, 2010a).

This technique not only allows understanding in which years the greatest changes occurred but is also a starting point to try to understand why they occurred.

A major advantage of combining the above techniques with LFS data is that this study can be easily replicated in other EU and EFTA countries, allowing for a comparison of results. Furthermore, it makes it possible to keep monitoring in the future.

Lastly, it is important to note that the greatest contribution of this work lies in the use of a methodology that is not yet widely used in Official Statistics. Although in Portugal there are already some works that combine SDA with Official Statistics, namely Campos (2007); Grilo (2012), there is no work that combines data from the Luxembourgish LFS with SDA. In fact, this methodology was a novelty in the Statec unit where this work was developed. However, it has an extremely important application in the field of Official Statistical data since, by law and for confidentiality reasons, the National Statistics Institutes (NSI) are forbidden to share individual information, thus analyzing aggregated data allows for freely sharing the results. Additionally, the aggregation of the data also allows different independent surveys made on the same population to be combined, if the same variables are used to form groups. This means that it is possible to combine the LFS with another survey that may be relevant to extend the analysis done in this work. Moreover, the use of the LFS makes it easy to replicate the work in other countries, for example, in countries where immigration is also a large phenomenon and to compare the results.

1.2 Objectives

One of the first main points of this work is to understand what differentiates the most the different immigrant groups in the labour market. 'Is it the age?', 'Is it the level of education?', 'Is it the sector of activity?', 'Is it the wages?', 'Is it the ability to telework?',... In addition, the main objective of the work is, based on those results, to split the population into homogeneous groups and characterize their profiles in the labour market. Another important goal is to portray the clusters' changes and understand if those profiles have changed over time.

Besides, some other more specific questions are also addressed, for example: 'Based on the place of birth, what is the age structure of the people who live in Luxembourg?', 'Is there any association between the reasons for immigrating and the place of birth?', 'How do employment and unemployment rates vary across the different sub-populations?', 'Since covid, how did telework evolve across groups? Was telework just a temporary situation?'

1.3 Official Statistics

A major part of the proposed work was developed through an internship at Statec, at the 'Marché du travail et Education' unit, under the European Master in Official Statistics (EMOS) programme.

Official statistics are statistics produced within the national statistical systems. Statistical organisations and divisions that together gather, process, and disseminate official statistics on behalf of the national government are referred to as national statistical systems. Official statistics are often produced within a legal framework and in accordance with fundamental principles

that guarantee minimum professional standards, such as independence and objectivity. For the European Union, the European Regulation (EC) No 223/2009 serves as the legislative framework and the European Statistics Code of Practice is the collection of guiding principles (Eurostat, 2017).

Statec is the Luxembourgish National Institute of Statistics and Economic Studies. More concretely, it is a department reporting to the Ministry of Economy. It establishes its own work program in accordance with national and European statistical legislation. Its goal is to offer high-quality public information services to citizens as well as to public and private decision-makers. The purpose of Statec is to produce studies, analyses, and statistics that present a complete, trustworthy, and impartial picture of Luxembourg society (Statec, 2022b).

1.4 Dissertation Structure

This work is organized into seven chapters. The first is an introduction where the motivation, the problem studied, and the main objectives are described. In the second chapter, which is divided into three sections, a contextualization is made and some of the works related to the subject are presented. In chapter three, the methods used to make the proposed analysis are described. The fourth chapter is concerned with the software, the data pre-processing techniques and a descriptive analysis. The fifth chapter follows explaining the selection process of the models. In chapter six one may find all the results obtained through the analysis. The work ends with a section dedicated to the main conclusions of the work, final considerations and future work.

Chapter 2

Migration and the Luxembourgish Labour Force

2.1 Europe Migration Flows

Migration movements into and within Europe have changed significantly throughout time in terms of scale and pattern. The size and structure of the overall population are both impacted in a short and long time by these migration flows (European Commission and Eurostat, 2011). It is common to attribute migration to Europe to conflict, poverty, population growth, and climate change in African and Middle Eastern nations (de Haas & Vezzoli, 2018).

Migration from non-EU nations must be distinguished from migration inside the EU. The freedom to live and work in another member state is guaranteed to EU citizens. They are also exempt from various limitations, such as those relating to the length of stay in a country that may be applied to people who are not citizens of an EU country. Figures on migrant integration (from 2018) suggest that 2.4 million immigrants from non-EU countries entered the EU, while 1.4 million people who had previously lived in an EU member state moved to another member state. The country with the highest number of immigrants in absolute terms was Germany (893.9 thousand), followed by Spain (643.7 thousand). Malta had the greatest rate of immigration (56 immigrants per 1000 people), followed by Luxembourg, (41 immigrants per 1000 people) (Commission, Eurostat, Kraszewska, Juchno, & Todorova, 2020).

The position of Europe on the world migration map has drastically changed since the end of World War II. While, in the past, Europeans have been colonising, conquering, and settling in other nations, migration was partially reversed in the second half of the twentieth century (Castles, Haas, & Miller, 2014). Decolonization, demographic shifts, fast economic growth, and the establishment of the EU as a free trade and immigration zone have contributed to several European nations currently matching or surpassing the immigration rates of the USA and other traditional immigration countries.

In the transition of European migration, there are four key periods that may be identified. These phases are linked to various stages of economic development, shifts in economic ideology, adjustments to the structure of the labour market demand in Europe, changes in the population, and the twinning processes of European integration (de Haas & Vezzoli, 2018).

Labour migration, which was thought to satisfy the demand for labour force during the economic boom after the end of the Second World War, was prevalent during the first phase (1945–1970s). Migrants were seen as a flexible, rotating and temporary workforce (Hartung, 2010). Large-scale businesses focused their economic strategy during this time on post-war investment and production development in the already highly industrialised European nations (Castles et al., 2014). Germany, France, and the United Kingdom among others saw a rise in labour immigration in the 1950s and 1960s. Several other nations, including Ireland, Spain, and Italy, had a high rate of emigration during the same period. Migration patterns in this period from the West Indies and the Indian subcontinent to the United Kingdom and from Algeria and Morocco to France are examples of how migration has frequently mirrored historical or linguistic links between countries (European Commission and Eurostat, 2011). This period, more concretely since the 1960s, coincides with the increasing role of women in labour migration, which was until then a male-dominated phenomenon (Castles et al., 2014). The 1973 Oil crisis, which caused a recruiting freeze, marked the end of this period (de Haas & Vezzoli, 2018).

The second phase, which lasted from the middle of the 1970s to the middle of the 1990s, was characterised by economic downturns and was marked by an increase in the prevalence of neoliberal economic policies. These included policies like ‘flexibilizing’ labour markets (which increased the opportunity for temporary recruitment of migrant workers), suspending state subsidies for only certain industries like mining, less progressive taxation,... The relatively high migration levels even during a period of recessions can be attributed to reasons such as family reunions and new marriages between members of the ‘second generation’ and spouses from the origin communities (de Haas & Vezzoli, 2018). Asylum (in the 1980s) was another main entry channel for migrants (Hartung, 2010). In the mid-1980s, economic growth resumed. While labour demand was maintained in industries like agriculture and construction, the fast rise of the service sectors attracted an increasing number of low and high-skilled migrant workers (de Haas & Vezzoli, 2018).

The collapse of the Berlin Wall (1989) fueled migratory flows from the mid-1990s until 2008 (third phase). This marked the Soviet Union’s collapse (de Haas & Vezzoli, 2018), which triggered large-scale migration from Yugoslavia and the former Soviet republics (Hartung, 2010). The immediate consequence of the collapse of the Berlin Wall was the temporary influx of refugees seeking shelter in Western Europe. As a result, Central and Eastern Europe saw an unexpected emergence of a new ‘labour frontier’ during this period. This phase was characterised by economic growth and the emergence of new migration destinations like Ireland, Italy, and Spain. These new destinations attracted migrants from Eastern Europe, North and West Africa, and Latin America which was a significant shift from the past when migration from Turkey and the Maghreb comprised the majority of non-EU migration to the EU (de Haas & Vezzoli, 2018).

Finally, the global economic crisis that occurred in 2008 and interrupted the economic expansion marked the beginning of the fourth phase (from 2008 to the present). The crisis had a particularly negative impact on several nations, like Greece, Spain, Portugal, and Ireland, which, as a result, saw a rise in emigration. More specifically, this resulted in a rise in the number of Spaniards, Portuguese, and Italians moving to northern Europe. Between 2007 and 2012, there was a decline in migration to the EU. Starting in 2012, and particularly since 2015, both intra-EU and extra-EU migration have rebounded, mostly as a result of the North-West European region’s resumed economic growth and declining unemployment. In 2015, a significant influx of Syrian

and other refugees entering Europe via Turkey and Greece started (de Haas & Vezzoli, 2018).

According to de Haas and Vezzoli (2018), the last phase lasted, at least, until 2018. However since then the pandemic, and more recently the war in Ukraine, have certainly contributed to some changes in the migration flows in Europe. Especially an increased inflow of Ukrainians looking for asylum since February 2022.

Regarding the primary causes of migration, most nations experience a variety of migration types at once rather than just one. Some of those types are labour migration, family reunion migration, refugee movement migration, or permanent settlement migration (Castles et al., 2014). Nevertheless, in EU, in 2021, almost half of the foreign-born people migrated for family reasons (47.4% of total foreign-born people aged 15-74). Family reasons were followed by reasons related to employment. In addition, 9.6% migrated for international protection or asylum and 6.1% for education reasons. A very small share of foreign-born people migrated to retire in the host country (0.5% of the total). However, reasons for migrating change significantly according to the country of birth, age and gender (Eurostat, 2022d).

2.2 Luxembourg

Luxembourg, a country that stands out for its demographic dynamism, is one of the smallest countries in Europe. Projections show that Luxembourg's population will continually grow and will reach 1 066 377 inhabitants in 2080 (Statec, 2017). Over the last ten years, more than 80% of the population growth is due to net migration (difference between immigration and emigration) (Ametepe et al., 2022).

2.2.1 Immigration Flows

The country's economic progress is often reflected in the migratory balance. Strong economic expansion is associated with very high net migration, as seen in the first half of the 1960s, the early 1970s as well as the period from the middle 80s onwards. On the other hand, a brief fall in the net migration rate is a sign of economic slowdown or slow economic growth (Zahlen, 2012). Historically, immigration in Luxembourg has been dominated by inflows of people coming from other European countries (Beine & Souy, 2016).

The beginning of immigration to Luxembourg may be dated to the industrialization of the country in the late 19th century when the mining and steel industries emerged as the primary economic drivers of the country. Italian and German workers were hired as a result of the rise in demand for both low- and high-skilled labour. Their social rank varied greatly, despite the fact that they were hired to achieve the same economic goal. Italian employees occupied the bottom position in the industry, while German workers occupied the highest ones (Gerstnerova, 2014). From 1961 until 1981, the Italian community was the most numerous one (Zahlen, 2012).

Things changed a little after World War II. German and Italian immigration decreased, and many workers went back to their home countries. An influx of Portuguese immigrants began in the 60s. This most recent cohort of immigrants began their careers in low-skilled jobs in the manufacturing, catering, and construction industries. The population of Portuguese immigrants

increased steadily throughout the second half of the 20th century as a result of Portuguese families' reunion and Portugal's admission to the EU in 1986 (Gutfleisch, 2022).

In the decades that followed the 1970s steel crisis (a recession in the global steel market from 1973 to 1975 that was further exacerbated by the 1979 oil crisis), the rise of the services sector, particularly the financial sector, led to rapid economic expansion and an increase in the demand for highly educated workers. The domestic labour force was unable to satisfy this demand. High-skilled workers from the nearby nations of Belgium, France and Germany were mostly employed to fill such posts (Gutfleisch, 2022).

The net migration rate in Luxembourg has been rising since 2004, with a significant decline in 2009, during the peak of the country's financial and economic crisis. From 2010 onwards and particularly in 2011, the flow of entries was on the rise again (Thill & Zahlen, 2012). 2011 was a year of significant growth, with the net migration rate in Luxembourg hitting a record-high of 21.6 per thousand (Zahlen, 2012).

The Luxembourg government declared a state of emergency on March 18, 2020, lasting until June 24, 2020, which significantly affected immigration to Luxembourg. Movement restrictions severely hampered immigration flows in those three months before they started to rise again and peaked in September (Ametepe et al., 2022).

The number of Ukrainians who are officially residents of Luxembourg quadrupled between January and June 2022. The majority of this growth occurred between March and April when the war first broke out (Ametepe et al., 2022).

In Luxembourg, the Portuguese people are usually distinguished by lower socioeconomic level than locals (in terms of income, labour market positions, and education) (Hildebrand, Pi Alperin, & Van Kerm, 2017), whereas the Germans and the French are generally highly skilled and frequently work in more attractive sectors. Given that they know at least one of Luxembourg's three official languages (German, French, Luxembourgish), the two latter groups' language barriers are less severe than those of the Portuguese group (Gutfleisch, 2022). The Italian group is similar to the Portuguese one but it has a much longer history of presence in the Luxembourgish society (Kankaraš & Moors, 2012).

Work is the main reason for the immigration phenomenon to Luxembourg, but there are other factors that encourage people to move, such as the proximity of neighbouring countries, family or neighbourhood networks, knowledge of the country's economic situation, the existence of a particular immigration tradition, like the one between Luxembourg and Portugal and family reunification (Peltier & Klein, 2020).

Despite the high immigration flows to Luxembourg, the country is not the primary destination choice of forced migrants, but when they apply for asylum the two main reasons are family/friends already there or 'by accident' (Glorius & Nienaber, 2022).

2.2.2 Labour Market Conditions

Due to the significant number of immigrants who move to Luxembourg and the cross-border employees (foreign workers residing in neighbouring countries - Belgium, France or Germany - but who are employed and work in Luxembourg), the Luxembourgish labour market is strongly internationalised. In fact, Luxembourg is the EU nation that relies on foreign labour the most to

support its national economy (Choe & Van Kerm, 2018).

This section provides an overview of the current Luxembourgish labour market as it explores some of the work that has been done concerning some labour market indicators as well as the work done on the differences of working conditions between immigrants and natives.

The Luxembourgish labour market's internationalisation varies greatly across sectors of activity. Succinctly, Luxembourgers are heavily represented in the public administration, health and social work sectors, as well as in the previously public-owned sectors of energy production and distribution. Immigrants are mostly employed in the construction, household services, hotel and catering industries. Cross-border employees typically work in manufacturing, financial services, and administrative and support service sectors (Thomas, 2020).

Comparing the Grand Duchy to other EU countries, its employment structure is atypical. While 22% of workers are employed in the financial, insurance, and scientific and technological sectors - where average wages are relatively high - these sectors only account for 8%, 9%, and 11% of workers in Belgium, Germany, and France, respectively. On the other hand, the construction sector - where average wages are relatively low - employs 11% of employees, which is significantly more than the EU average (5%). There are other sectors such as trade, industry, health, and education, which play a weaker role in Luxembourg compared to other EU countries and the EU as a whole. For example, trade employs 10% of the workforce in Luxembourg, while it represents 22% of the jobs in the EU (Reiff, 2021b).

The Luxembourgish labour market has always been one of the most dynamic in Europe, growing by an average of +3.4% per year between 2016 and 2019, as opposed to +1.3% at European level. Unlike almost all European countries (at EU level the employment contracted by 1.5%), employment, in Luxembourg, has increased in 2020 (+2% compared to 2019). Instead of coming from a particular branch of the economy, the increase in employment is the result of a relatively widespread phenomenon throughout the economy. The annual variation in employment was negative only in two out of ten branches¹, agriculture and industry, whereas in the EU as a whole, employment has fallen in seven branches. The strongest increases in Luxembourg occurred in the following branches: public administration, education, health and social work, real estate and construction (Rota & Larue, 2021).

The unemployment rate (ILO definition²) in Luxembourg decreased to 5.3% in 2021, which was below the EU-27 average (7.0%). Unemployment in the Grand Duchy is greatly determined by the level of education, being 8.5% among people with a basic level, a percentage that drops to less than half for those with a tertiary level qualification (4%) (Ametepe et al., 2022).

Luxembourg has the highest average gross annual salary in the EU (€64 932), which is 182% more than the EU average. However when the purchasing power of a EUR is considered, the discrepancy narrows and the average gross annual salary in Luxembourg is 145% of the EU average. These significant differences in the salaries between Luxembourg and other countries

¹ Arts, entertainment and recreation and other service activities; Public administration, defence, education, health and social work; Business services; Real estate activities; Financial and insurance activities; Information and communication; Trade, transport and hospitality; Construction; Industry; Agriculture.

² Unemployment according to the ILO definition is measured by the Statec Labour Force Survey. The number of unemployed corresponds to the number of people who declare that they are unemployed, available to take up a job within two weeks and actively seeking work.

only occur in certain branches such as education, public administration, health and social work and financial and insurance activities; in other sectors, the gap is smaller, and Luxembourg even has lower average wages than some of its neighbours, for example in administrative and support services activities, hotels and catering, trade, construction and industry (Reiff, 2021b).

Furthermore, the average gross annual salary of an employee can vary greatly between sectors within the country. For instance, in 2018, the annual average wage for an employee in the hotel and catering industry was €33 796, whereas the average salary for an employee in the financial and insurance industry was almost three times as much, €99 250. Although these discrepancies may appear significant, they may be explained by the employee's degree of education, their profession, their sector of employment, whether they work full- or part-time, ... (Reiff, 2021a). The disparity in annual wages between natives and immigrants is mostly because residents and immigrants have different characteristics, such as different levels of education or types of occupations. If the different job characteristics of foreign and native workers are taken into account, one may notice a significant reduction in the gap between the counterfactual and observed wage distributions, albeit not an eradication. In addition, the top immigrant wages converge towards top native wages (Van Kerm, Yu, & Choe, 2016).

The Gender Pay Gap³ (GPG), in Luxembourg, has been decreasing for the last 20 years (Schockmel et al., 2022). According to the latest figures, in 2021 at -0.2% the pay gap between men and women became negative (i.e. in favour of women). With this near-equality in pay, Luxembourg is the exception in the EU which still has a GPG of 12.7%. Nevertheless, the gaps in Luxembourg are still largely in favour of men in certain branches of activity, such as financial activities (23%), real estate activities (23%), specialised, scientific and technical activities (22%) and commerce (18%) (Division sociale, 2023).

Overall and in the sectors of most concern, Luxembourg has a smaller percentage of inhabitants with temporary contracts than its neighbours (particularly in the hotel and catering sector), which helps minimising the number of vulnerable positions (Rota & Larue, 2021).

In Reiff (2019) a comparison of the hours worked in Luxembourg and in the Neighbouring countries is done. The average number of hours worked by a full-time employee in Luxembourg is 1 701 hours annually, which is higher than the comparable figures for the neighbouring nations of Germany (1 677), France (1 544), and Belgium (1 495).

When it comes to countries where employees have the most control over their working hours, Luxembourg, Denmark, and the Netherlands are tied in third place. In 2019, employers, organisations, or clients, primarily determined the working hours for 49% of workers, compared to 61% at the EU level. Albeit, in Luxembourg, one in every three people (37%) had to adjust their working hours at least once a week due to unforeseen demands, which is much higher than the EU average (21%) (Eurostat, 2020d).

Regarding the percentage of part-time workers, although it has increased (from 13.2% in 2008 to 18.3% in 2016), it is still lower than the average for the European Union (26.0%). A comparison between 2011 and 2021 shows that the gap between foreigners and Luxembourgers has widened considerably: part-time work has decreased among foreigners (from 15.6% to 13.6%) and increased among Luxembourgers (from 20.4% to 23.2%), which is the opposite of

³GPG = $\frac{\text{average gross hourly wage of men} - \text{average gross hourly wage of women}}{\text{average gross hourly wage of men}}$ (in %)

what typically occurs in other countries, where immigrants tend to have part-time jobs more frequently than natives (Ametepe et al., 2022).

An amount of 32.3% of the employed people in Luxembourg worked atypical hours (night work, evening work, Saturday or Sunday work and shift work) in one way or another in 2021. This was a slight increase compared to the previous year, nonetheless, it was less than the European average (34.2%) (Ametepe et al., 2022).

Nowadays, in the Grand-Duchy, there is great heterogeneity between the different immigrant groups regarding the level of education, distinct from the situation in many other European countries. A large proportion of immigrant workers who arrived in some EU countries in recent decades have lower levels of education than the native-born. It has been shown, for residents aged 15-74 years, that non-Luxembourg nationals are generally more likely to have a higher level of education than Luxembourgers. However, this is not the case for Portuguese nationals, who have significantly fewer university degrees. Bearing in mind the immigrants' qualifications it is essential to understand to what extent is the human capital of immigrants underused in Luxembourg. The answer is that one in every ten immigrants works in a less qualified job than before migration. This proportion is slightly higher for women. Thus, while immigration may enable women to work more often, men benefit from more qualified jobs than in the country they previously lived (Ametepe et al., 2022).

When searching for a job, 28.1% of the immigrants found it before arriving in Luxembourg. Three out of four people did not find any obstacles in finding a position corresponding to their diploma. The other quarter found that the main obstacle was the language barrier. Luxembourg comes in third, in comparison to other EU countries, as a country where the highest percentage of people (34.6%) from a non-EU country reported not having any problem finding a suitable job (Eurostat, 2022d).

According to the 2020 EU-LFS data, the EU over-qualification rate for non-EU citizens was 41.4%, whereas it was 32.3% for citizens of other EU Member States. The over-qualification rate for nationals, however, remained at 20.8%. With an over-qualification of 3.2% for nationals, 3.5% for citizens of other EU Member States, and 10.8% for non-EU citizens, Luxembourg had the lowest over-qualification rates (Eurostat, 2021d).

Overall, Luxembourg's labour market conditions are better than those of the other European member states, both for native-born people and immigrants, making it a very attractive country to work.

2.3 Labour Force Survey

The EU-LFS is the largest sample survey of private-households population in Europe and is the primary European source for information on the labour market (Eurostat, 2020a). It provides quarterly and yearly results on people of working age (15 years and older) who are participating in the labour force as well as on those who are outside it (Eurostat, 2022b). While questions regarding a person's position in the labour market are restricted to those who are at least 15 years old, demographic data are gathered for all age groups of the population (Eurostat, 2022a). The EU-LFS is a significant source of information on the state and trends of the labour market

in the European Union thanks to its statistics on employment, unemployment, and inactivity. It is important to make a distinction between the national LFSs and the EU-LFS. The former is the survey that is done in each individual country and, as described above, labour market and demographic data are also collected. The EU-LFS is a cross-sectional and longitudinal household sample survey, collated by Eurostat from the labour force survey data of 35 participating countries that are detailed later on. Therefore, the EU-LFS is a way of harmonising and comparing labour market data between the various countries. It is important to note that there are several overlaps between the two, as the LFSs largely contribute to the EU-LFS. For the purpose of this work, the national LFS of Luxembourg is used and it enables a fuller understanding of trends of the labour market in the country. The majority of variables used follow the guidelines from the EU-LFS, nevertheless, there are a couple of variables used that are specifically collected in Luxembourg and are not mandatory for the EU-LFS. However, with minor changes, it is possible to replicate the work in other countries so that the results can be compared. For example, if only variables that are harmonized by Eurostat are used.

The main statistical goal of the LFS and the EU-LFS is to divide the population of working age into three exhaustive mutually exclusive groups: the employed, the unemployed which together make up the labour force (active population) and the people who are not in the labour force (Eurostat, 2022a). Different variables are gathered based on an individual's employment status. The following is an explanation of the three concepts of labour status.

- A person who is employed is one who, during the reference week, fell into one of the following groups and is between the ages of 15 and 89 (measured in completed years at the end of the reference week): those who participated in at least one hour of labour for pay or profit, those who had a business or a job during the reference week but were temporarily off the clock, or those who produce agricultural products with the primary goal of selling or exchanging them for other items (Eurostat, 2021b). The employment rate is computed as follows:

$$\text{Employment Rate} = \frac{\text{Employed Population}}{\text{Working Age Population}} \times 100 \quad (2.1)$$

The employment rate is usually computed for different age classes.

- Anyone between the ages of 15 and 74 who is not employed during the reference week in accordance with the definition of employment and who is available for paid employment or self-employment before the end of the two weeks that follow the reference week, is considered to be unemployed. This person must also be actively looking for work, i.e. had taken actions in the four weeks preceding the reference week to look for paid work or self-employment, or had found a position to begin in at most three months from the end of the reference week (Eurostat, 2022c). The unemployment rate is calculated in the following way:

$$\text{Unemployment Rate} = \frac{\text{Unemployed Population}}{\text{Active Population}} \times 100 \quad (2.2)$$

Like the employment rate, the unemployment rate is also usually computed for different age classes.

- A person is outside the labour force if he or she is neither employed nor unemployed. The set of people outside the labour can include, for example, pre-school children, school children, students, pensioners, and housewives or -men given that they are neither employed nor available for employment nor either searching for a job (Eurostat, 2020b).

Traditionally, in Europe, LFSs were undertaken once a year, more concretely, in spring. Spring was chosen as a ‘representative’ time of year because seasonal influences had no impact on the number of employed persons. This approach, however, did not enable yearly averages or the tracking of short-term labour market dynamics. Given that, the survey transitioned from an annual to a quarterly format and the data started to be collected every week of the year. Although some did it sooner and others later (the last happened in 2015), the majority of member states updated their surveys to the continuous form between 1998 and 2004, with Luxembourg making the shift in 2003 (Eurostat, 2021a). Nonetheless, Statec only started to provide quarterly and annual results in 2007 (European Commission and Eurostat, 2015).

Detailed information on the surveyed variables of the EU-LFS is provided by the explanatory notes that can be found in Eurostat (2023). For each variable, it informs the readers about its periodicity, purpose and definition. This document has had some changes throughout time and the present work picks up four of those changes, something that carefully needed to be taken into account. More concretely those changes happened in 2014, 2016, 2017 and the biggest one in 2021.

Currently, 34 countries participate in the EU-LFS, supplying Eurostat with information from national labour force surveys. These include all the EU member states, Iceland, Norway, and Switzerland (EFTA countries) as well as Montenegro, North Macedonia, Serbia, and Turkey (candidate countries). The United Kingdom participated up to the third quarter of 2020. Four major methods are used to gather data for the EU-LFS: in-person meetings, phone interviews, online interviews, and self-administered questionnaires. The EU-LFS employs the same concepts and terminology, adheres to ILO regulations, uses common classifications (NACE, ISCO, ISCED, NUTS), and records the same set of variables in each country to make sure that the statistical findings are comparable between nations and through time (Eurostat, 2022a).

In Luxembourg, since 2015, participation in the LFS is compulsory (European Commission and Eurostat, 2015, 2016b) and the interviews are conducted by INFAS (*Institut für Angewandte Sozialwissenschaft GmbH*). Telephone or online interviews are currently the two possible ways to participate in the survey (Statec, 2022a).

The EU-LFS is highly regarded since, as mentioned, it uses internationally accepted concepts and definitions and provides consistent and comparable labour market indicators over long periods. It is a valuable tool for measuring employment and unemployment, as it provides a large amount of detailed information that can be used to inform policy decisions. Furthermore, it captures a significant amount of socio-demographic information. Since it includes information on nationality, country of birth and date of arrival, it allows countries and the EU to compile statistics on migrant workers who are employed. This allows to understand how the stock of

foreign and foreign-born workers varies over time as well as their working conditions (Cangiano, 2012).

Nevertheless, there are some limitations that affect the quality of the estimates and the analysis of the results obtained when the data source is the LFS, especially when migration is analysed. According to European Commission and Eurostat (2011), the LFS is likely to underrepresent the migrant population. For example, migrants move around more than the population of long-term residents, making them less likely to meet any criteria for inclusion in the sample, such as continuous residence at the current address. Other limitations include typical survey limitations such as respondents might not provide truthful, accurate responses, particularly if the respondents are migrants. The language barrier may lead to inaccurate or incomplete answers. Additionally, if a respondent's residence or employment status does not fully comply with immigration laws, they might not feel comfortable answering questions that place them in a bad manner, limiting the quality of the analysis.

2.3.1 Sampling Strategies

According to national specificities, NSI employs the design that best suits its requirements and quality standards.

Like other surveys, the EU-LFS uses a sample of the population to produce estimates that are representative of the entire population. Given that not every unit of the target population is surveyed, the EU-LFS estimates are susceptible to the typical errors associated with sample techniques, such as sampling errors. Nevertheless, those errors can be quantified using measures that may be based on the variance (Eurostat, 2020c).

Before going to the Luxembourgish case, it is important to understand some concepts related to the sampling strategies used by the countries that participate in the EU-LFS.

Probability Sampling

The distinction between probability and non-probability sampling must first be understood. Non-probability sampling is a subjective (non-random) method of choosing units from a population. It is a quick, easy, and cost-effective method for obtaining information. However, it is important to make the dangerous assumption that the sample is representative of the population in order to draw inferences about the population. It is typically challenging to verify whether this assumption is realistic. On the other hand, probability sampling is the process of choosing a sample from a population when the choice is based on the randomization principle. Because units are selected at random and the selection probabilities for each unit can be calculated, accurate estimates can be produced (lowering the risk of over- or under-representing the population and guaranteeing that the results are representative of the population) and statistical inferences about the population can be drawn. For the majority of statistical programmes, probability sampling is the ideal option as it enables the quantification of estimation error, which supports the drawing of accurate conclusions about the whole population (Eurostat, 2020c).

All the countries participating in the EU-LFS use probability sampling as its foundation.

Simple Random Sampling

Every person in the population has an equal probability of being chosen in a simple random sampling. The systematic random sampling approach is another option, and it relates to the scenario when the sample is chosen from an ordered sampling frame (Eurostat, 2020c).

Stratified Sampling

In stratified sampling, the population is partitioned into non-overlapping groups known as strata and then a sample is selected from each stratum. Sampling units are independently sampled from each stratum. To make sure that particular population subgroups are represented in the sample, stratification is used. The goal of each stratum is to represent a uniform segment of the population (Eurostat, 2020c).

Cluster Sampling

Cluster sampling divides the population into groups that are internally heterogeneous yet externally homogeneous. Each cluster essentially serves as a tiny representation of the overall population. After splitting the population, one or a few clusters are selected at random and a census is carried out within each one. A more sophisticated type of cluster sampling known as multistage cluster sampling involves two or more phases of sample selection and uses progressively smaller clusters at each stage (Eurostat, 2020c).

LFS Sampling Strategy Luxembourg

The present work covers all the years from 2014 until 2022. In 2014 the sampling units were the households and they were selected by the procedure of random digit dialling. Nevertheless, from 2015 onwards the procedure is different. Simple random sampling is the chosen sampling strategy and people are sampled from the population register. Moreover, a rotational scheme has been applied since then. Each individual stays in the sample for five quarters. Even though the sampling unit is the individual, all the individuals included in the sampling unit's household have to answer some questions. Once the individuals have been selected, weights are calculated so that the sample is representative of the population. In a simplified way, at Statec, strata are defined based on sex, age class, household size and nationality (national/foreign). To compute the weights, the total population in each stratum is divided by the number of interviewed people in the same stratum and all the people in each of these 288 strata have the same weight (European Commission and Eurostat, 2015, 2016b, 2018b, 2018c, 2019, 2021c, 2022).

To sum up, it is important to bear in mind that in 2014 the methodology adopted by Statec was a little different from what it has been since 2015. The survey was not mandatory, the units were the households and not the individuals, and people were not sampled from the population register, actually, households to be contacted were selected by the procedure of random digit dialling (set of all telephone numbers).

Chapter 3

Methodology

3.1 Symbolic Data Analysis

Symbolic Data Analysis (SDA) (Billard & Diday, 2003, 2007; Bock & Diday, 2000; Brito, 2014; Diday & Noirhomme-Fraiture, 2008; Noirhomme-Fraiture & Brito, 2011), initiated by Diday in the late 1980s, is concerned with the analysis of data that exhibit inherent variability. It provides a framework to represent and analyse groups gathered on the basis of some given criteria. Since the resulting concepts are structured and entail inherent variability, they can only be characterised by a more sophisticated type of data (symbolic data) when compared to the standard observations. The ability to summarise each class or group into lists, intervals, histograms, etc. in order to preserve more internal information, that is, the ability to account for the intrinsic variability of the groups is a tremendous advantage. The standard approach to representing aggregate data is to calculate summary indicators for groups - for example, means, medians, or modes - thus ignoring a lot of information, and limiting comparison between different groups. Another advantage is that by using data compression, SDA gets around the usual problem of the standard approach: dealing with enormous volumes of data, which is computationally costly and makes it difficult to understand the data's general nature. Furthermore, SDA has an extremely important application in the field of Official Statistics since, by law and for confidentiality reasons, the NSIs are forbidden to share individual information, thus analyzing aggregated data allows for freely sharing the results. Additionally, the aggregation of the data also allows different independent surveys made on the same population to be combined, if the same variables are used to form groups. This is not always possible when dealing with micro-data because when considering different surveys the individuals may not always be the same.

In SDA two types of individuals can be distinguished: First-order individuals (FOI) and Second-order individuals (SOI). A FOI represents a physical entity of the real world (such as a resident of Luxembourg), while an SOI is a statistical unit of a higher level representing a concept (for example place of birth) (Diday, 2000).

Symbolic data is represented as an extension of the classical tabular model. In the classical model, each row corresponds to an individual who is described by variables represented in the columns and that can only take one value. Symbolic data allows each variable to have more than one value for each unit, in particular, it allows objects to be described by distributions (Di-

day, 2000). New variables, described in Subsection 3.1.1, were introduced for this new type of representation.

To sum up, the purpose of SDA is to analyse, visualise, classify, and reduce the information contained in a symbolic data matrix.

3.1.1 Types of Variables

As previously mentioned, SDA created a need for new types of variables. There are still numerical (or quantitative) and categorical (or qualitative) variables, like in classical statistics. In SDA, within the numerical variables, there are: single-valued (real or integer) variables which take a single value of an underlying domain for each unit, multi-valued variables whose values are finite subsets of the domain and interval-valued variables whose values are intervals of $\mathbb{R}$. A variable is referred to as a histogram-valued variable if a distribution over a set of sub-intervals is given. Regarding the categorical variables, they may also be single-valued (ordinal or nominal) if for each entity it takes one category from a given finite category set $O = \{m_1, \dots, m_c\}$ or multi-valued if its values are finite subsets of O . Additionally, a categorical modal variable Y with a finite domain $O = \{m_1, \dots, m_c\}$ is a multi-valued variable when for each element a category set is given and, for each category m_i , a frequency or probability indicating how frequent or likely that category is for this element is also provided (Brito, 2015).

In summary, there are three main new types of variables that should be considered: multi-valued, histogram-valued (that includes the particular case of interval-valued variables) and categorical modal variables.

Distributional variables (histogram-valued and categorical modal) are better when compared to the other types of variables as they allow for keeping more information on the internal variation.

Considering the software limitations, for the analysis proposed in Chapter 1, only categorical modal variables are used in the work. Otherwise, the continuous variables would be aggregated into histogram ones and the categorical variables into categorical modal variables.

To exemplify the categorical modal variables, Table 3.1 was created. The table contains the migration reasons of people born in three different countries. From country A, 55% of people migrated for reason 1, 15% for reason 2,... From country B, 45% of people migrated for reason 1, 10% for reason 2,... Thus MIGREAS is a categorical modal variable.

Country of Birth	MIGREAS
A	$\{1(0.55), 2(0.15), 3(0.20), 4(0.05), 5(0.01), 6(0.04)\}$
B	$\{1(0.45), 2(0.10), 3(0.25), 4(0.10), 5(0.02), 6(0.08)\}$
C	$\{1(0.50), 2(0.05), 3(0.15), 4(0.15), 5(0.07), 6(0.08)\}$

Table 3.1: Migration reasons by country of birth

3.1.2 Clustering

Clustering may be defined as ‘the formal study of methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity’ (Jain, 2010). The fact that the class to which each of the units belongs is not previously defined and the fact that its objective is to find the underlying structure of the data by discovering patterns and exploring the hidden relationships makes clustering an unsupervised learning technique (Oliveira, 2010).

Clustering aims at classifying the objects of a given set Ω into homogeneous groups/clusters $C_1, C_2, \dots$ (Chavent & Bock, 2000).

The most commonly used clustering strategies are the non-hierarchical methods, which directly determine a partition of the set to be classified, and the hierarchical methods, which build nested class structures (Rizo Rodríguez & de Carvalho, 2022).

When a non-hierarchical clustering algorithm starts from an initial partition and moves one element at a time from one class to another using a criterion to compare partitions, it is a transfer method. On the other hand, there are the centroid methods in which cluster centers are determined and then the Ω elements are assigned to the cluster whose center is at minimum distance; once all elements are assigned, a new center in each cluster is determined. The stopping criterion is the non-changing of the partitions in two successive steps of the algorithm. Formally, the goal of non-hierarchical clustering is to determine partitions $P = \{C_1, \dots, C_k\}$ on Ω , that is, families of classes which do not intersect and that jointly cover the whole set Ω :

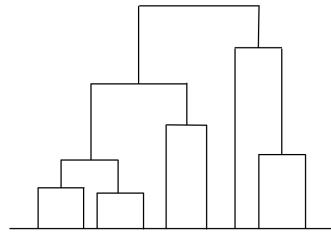
$$\boxed{\begin{aligned} P = \{C_h \subseteq \Omega, h = 1, \dots, k\} : \\ C_h \cap C_j = \emptyset, i \neq j, \\ \bigcup_{l=1}^k C_l = \Omega \end{aligned}}$$

In hierarchical methods, two different types of algorithms can be distinguished: the agglomerative and the divisive. In the former, each object in Ω is itself a cluster of cardinal one, and the algorithm merges units into larger classes in an ascending way (bottom-up approach), according to a linkage criterion. A linkage criterion is a function that computes the distance between two sets of data points as a function of the pairwise distance between observations in each set. For example, when using single linkage as the linkage criterion the algorithm computes the distance between clusters as the minimum distance between any two points in the clusters. Furthermore, one can say that hierarchical divisive clustering is an inverse of the agglomerative process as it begins with all units in one cluster and progressively subdivides each cluster into smaller clusters (top-down approach) (Chavent, 1998). This last type of hierarchical algorithms may be monothetic, when only one variable at a time is used to form the clusters, or polythetic when all variables are used to cluster the objects at each stage. Furthermore, the divisive technique is not frequently used because of its heavy computational load (Yim & Ramdeen, 2015). One of the main advantages of hierarchical algorithms is that there is no need to define the number of clusters *a priori*. Nonetheless, unlike the non-hierarchical algorithms, it is only suitable for relatively small problems because, in case of large datasets, the algorithm can be computationally complex.

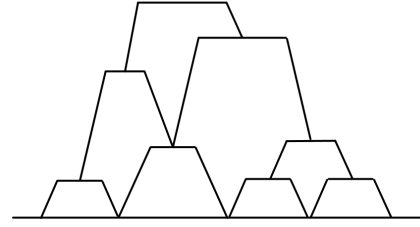
In addition, a pyramidal model for clustering has been developed. This model is a generalization of hierarchies as it allows for non-disjoint classes at each level. In a pyramidal classification,

an element can belong to at most two classes. In short, by using a pyramid one can obtain a set of overlapping clusters, while by using hierarchies one gets a set of partitions (Brito, 2002). In Figure 3.1 one may see the difference between a dendrogram and an indexed pyramid which are the representations of hierarchical and pyramidal approaches, respectively. Formally one can define hierarchies and pyramids as follows (retrieved from (Brito, 1998)):

Hierarchy H on Ω: Family of non-empty subsets of Ω such that: $\forall a \in \Omega, \{a\} \in H$ $\Omega \in H$ $\forall h, h' \in H, h \cap h' = \emptyset \text{ or } h \subseteq h' \text{ or } h' \subseteq h$	Pyramid P on Ω: Family of non-empty subsets of Ω such that: $\forall a \in \Omega, \{a\} \in P$ $\Omega \in P$ $\forall p, p' \in P, p \cap p' = \emptyset \text{ or } p \cap p' \in P$ There exists an order θ such that each element of P is an interval of θ
---	---



(a) Dendrogram



(b) Indexed Pyramid

Figure 3.1: Cluster representation

Clustering methods can be based on dissimilarities or they can be conceptual. On the one hand, there are clustering algorithms in which data are grouped based on numerical dissimilarity measures and the formed clusters can only be described or interpreted *a posteriori*, these are the methods based on dissimilarities. On the other hand, in conceptual clustering algorithms, the clusters are formed only if they can be described by a concept, which makes it easier to explain and understand data (Boubacar & Niu, 2014).

A wide range of methods, based on dissimilarities between the units to be clustered, have been proposed for clustering symbolic data. Since SDA uses new types of variables to represent units, different methods have been proposed to measure dissimilarity in this context. It is important to note that the fact that distance measures satisfy the triangular inequality makes them a special case of dissimilarities. Briefly, some of the dissimilarity measures that have been proposed for interval-valued variables include the Gowda and Diday's dissimilarity measure (Gowda & Diday, 1991), three measures proposed in Ichino and Yaguchi (1994), five measures proposed by de Carvalho (de Carvalho, n.d., 1994) and the Euclidean, City-Block, Hausdorff and Malahanobis distances. Regarding histogram-valued variables Wasserstein, Mallows and Euclidean distances can be chosen. For categorical modal data, Euclidean, City-Block and de Carvalho distances can be selected (Brito, 2015; Hardy & Lallemand, 2004). Additionally, in Malerba, Esposito, and

Monopoli (2002) a two-step process to compute distances between symbolic objects described by probabilistic variables (categorical modal and histogram variables) is explained. The first step is to compute the distance between objects using component-wise distance measures such as: Kullback-Leibler divergence (J), χ^2 -divergence (CHI), Renyi's divergence (REN), Chernoff's distance (CHER), and modified Minkowski metrics (LP). Then, the dissimilarity between the objects can be obtained by aggregating the component-wise distance measures through the generalized and weighted Minkowski's metrics. Given the development of the present work, it was concluded that using the two-step process would be the most appropriate choice (see Section 5.2). More concretely, using the χ^2 divergence distance and a L_2 type aggregation function. Given two discrete probability distributions S and V , the χ^2 divergence is computed as in Equation 3.1.

$$m_\chi^2(S, V) = \sum_{m_j \in O} \frac{|s(m_j) - v(m_j)|^2}{s(m_j)} \quad (3.1)$$

where s and v are two rows of a symbolic data table and $O = \{m_1, \dots, m_c\}$ is the domain of a categorical modal variable Y .

The aggregation function is applied as shown in Equation 3.2.

$$d_2(x_a, x_b) = \sqrt{\sum_{r=1}^p (c_r m_\chi^2(S, V))^2} \quad (3.2)$$

where, $\forall r \in 1, \dots, p$, $c_r > 0$ are weights with $\sum_{r=1 \dots p} c_r = 1$, p is the number of variables and x_a and x_b are symbolic objects.

Lastly, for mixed symbolic objects, the Manhattan distance or the Euclidean distance can be used (Ichino & Yaguchi, 1994).

Clustering Algorithms

Many different hierarchical and non-hierarchical approaches have been developed for SDA. Similar to dissimilarity measures, various clustering algorithms have been proposed considering the type of variables used.

For interval data there is an extensive list of methods, in particular of partitional methods. Adapted methods of k-means and dynamical clustering using Minkowski-type, Mahalanobis or Hausdorff distances were developed by Csernel and de Carvalho (1999), de Souza and de Carvalho (2004a), de Souza, de Carvalho, and Silva (2004), de Souza, de Carvalho, and Tenorio (2004), de Carvalho, Brito, and Bock (2006), Chavent, de Carvalho, Lechevallier, and Verde (2006). In de Souza and de Carvalho (2004b), de Souza and de Carvalho (2004a), de Carvalho, de Souza, Chavent, and Lechevallier (2006), de Carvalho and Lechevallier (2009), de Carvalho and Tenório (2010), extensions of the k-means algorithm by using adaptative distances, are proposed. Furthermore, adaptations of classical fuzzy c-means algorithm have been made, the first one is proposed in El-Sonbaty and Ismail (1998), followed by other approaches, for example, in D'Urso and Giordani (2006) and Jeng, Chuang, Tseng, and Juan (2010) and, more recently, in de Carvalho and Simões (2017) and Bao, Peng, He, and Wang (2018). In addition, the Fuzzy c-Ordered-Medoids algorithm, a more robust fuzzy clustering algorithm was analysed in D'Urso and Leski

(2016). A model-based approach using previously proposed Gaussian models is presented in Brito, Duarte Silva, and Dias (2015). Clustering methods based on Kohonen’s self-organizing maps are proposed in Pacifico and de Carvalho (2011), Dantas and de Carvalho (2011), Hajjar and Hamdan (2011), and Yang, Hung, and Chen (2012). More recently, in Jeng, Chen, Chang, and Chuang (2019) an interval possibilistic fuzzy c-means is proposed. Furthermore, Rodríguez and de Carvalho propose a method, based on adaptive Euclidean distance, that takes into account the relevance of variables (see (Rizo Rodríguez & de Carvalho, 2019)); moreover the authors, also suggest an algorithm with adaptative Euclidean and City-Block distances, which is described in detail in Rizo Rodríguez and de Carvalho (2022). Hierarchical approaches for interval-valued variables can be found in Rasson, Pirçon, Lallemand, and Adans (2008), and Hardy and Kasoro (2009). Another hierarchical clustering method that uses the width of range Euclidean distance is proposed in Galdino and Maciel (2019). Two different clustering methods based on Ward’s method and a new dendrogram representation are proposed in Ogasawara and Kon (2021). In addition, in Liu (2016) a k-regression-based clustering for interval data is described.

There are not many clustering methods developed for distributional data compared to those developed for interval data. Nevertheless, in Kim (2009) a polythetic clustering algorithm is developed. A k-means clustering method for categorical modal variables and an extension of Ward’s method can be found in Korenjak-Černe, Batagelj, and Japelj Pavešić (2011). In Kim and Billard (2012) a divisive clustering algorithm for categorical modal data is presented. Additionally, in Verde and Irpino (2007), Verde and Irpino (2008) and in Irpino, Verde, and de Carvalho (2014) a description of methods for partitional clustering for histogram-valued data is done. A fuzzy c-means clustering algorithm based on L_2 -Wasserstein distance for histogram-valued data was proposed in de Carvalho, Irpino, and Verde (2015) and is improved in Irpino, Verde, and de Carvalho (2017). In Billard and Kim (2017) the construction of hierarchies by using a divisive polythetic algorithm based on dissimilarity measures is explored for histogram-valued variables. In Kim and Billard (2018) a double monothetic clustering method that considers both the location and the internal variation of histograms is presented. A convex clustering method that is claimed to have the best performance compared to its competitors has been proposed for histogram-valued data in Park, Choi, Delcher, Wang, and Yoon (2019). Additionally, in Kejžar, Korenjak-Černe, and Batagelj (2021), the theoretical foundations for compatible leaders and agglomerative clustering techniques with alternative dissimilarities for categorical modal-valued variables are outlined. Note that it should be possible to apply all methods developed for histogram-valued variables to interval-valued data.

The clustering algorithms presented above are only suitable for a particular type of variable, but some methods have been developed for multiple data types. A divisive hierarchical clustering method that generates ‘monothetic’ clusters proposed in Chavent (1998), which was later extended in Brito and Chavent (2012) to data described by interval and/or histogram-valued variables. In Brito (1991) a pyramidal model-based clustering bottom-up algorithm for symbolic data is proposed. In Brito (1994) a method of symbolic pyramidal clustering using a set of multi-valued data is presented. The pyramidal model-based clustering was later extended to categorical modal data in Brito (1998). Additionally, a common framework for the generalisation of numerical, ordinal and categorical modal variables was proposed in Brito and Polaillon (2011). This method has been applied to hierarchical or pyramidal clustering in Brito (2015). In Xu (2016) a

new hierarchical clustering approach that can be applied to qualitative multi-valued or interval-valued data is suggested. Both methods can be applied to only one type of variable at a time. Lastly, Brito and Ichino developed clustering methods based on quantile representations of data. Using quantile-vector representations as a common representation enables a unified analysis of a dataset that has variables of different types. In Ichino and Brito (2022), quantile-based methods to analyse data are described.

Although in this work the variables used are only categorical modal, some reasons led to the choice of algorithms that allow for the use of several types of variables simultaneously. First, the algorithms that are currently most developed in the software chosen for the work (R - R Core Team (2022)) are those ones that allow for multiple data types. Moreover, the introduction of histogram-valued variables is being developed in some of the existing packages, which can be very useful for the analysis of continuous variables in the future. Furthermore, if, in the future, variables are added to the LFS that are more suitable for analysis in ways other than as categorical modal, no major changes will need to be made. The most important clustering algorithms for this work, given their applicability and state of development in R, are described in more detail below.

It is important to first emphasise that it is possible to adapt classical hierarchical methods to symbolic data if one correctly defines a dissimilarity matrix between the units to be clustered. It follows a brief description of the different types of linkages used in hierarchical clustering.

- Single linkage: the dissimilarity between two clusters is the minimum dissimilarity between members of the two clusters;
- Complete linkage: the dissimilarity between two clusters is the maximum dissimilarity between members of the two clusters;
- Centroid linkage: the dissimilarity between two clusters is the dissimilarity between the corresponding centroids;
- Average linkage: the dissimilarity between two clusters is the mean of all distances between members of the two clusters;
- Ward's linkage: the dissimilarity between two clusters is the increase in dispersion (measured by the sum of squares of distances to the centroid) when the clusters are merged;
- Median linkage: the dissimilarity between two clusters is defined recursively so that at each level the distance of the most recently formed cluster to any other cluster is the average of the median linkages of its two component clusters to that cluster;
- McQuitty linkage: the dissimilarity between two clusters is defined recursively, where at each step of the algorithm the dissimilarity of the most recently merged clusters A and B from any other cluster C is the average between the distances from A to C and from B to C.

Furthermore, two main approaches to dynamic clustering can be distinguished: SCLUST (symbolic clustering algorithm) and DCLUST (symbolic clustering algorithm on distance tables) (de Carvalho, Lechevallier, & Verde, 2008).

SCLUST, a module for the SODAS software that was later developed in an R package, is a dynamical clustering-like method (Diday & Noirhomme-Fraiture, 2008). It determines iteratively a series of partitions that improve at each step the underlying clustering criterion. Starting with a partition on a predetermined number of clusters, an assignment step (based on minimum distances) and a representation step are carried out alternately until convergence is established (or a given number of iterations is achieved). This approach locally optimises a given criterion, based on the assignment-distance function, that assesses how well cluster members fit prototypes (means, axes, probability laws, etc.) (Brito, 2015). The purpose of the method is to minimize the within-clusters inertia and thus to maximize the between-clusters inertia of the partition (Partners ASSO Developers, 2004). The prototypes as well as the dissimilarities for the assigning phase are chosen based on the variable's type. Concerning the prototypes, for categorical modal-valued variables, average weights are computed for each category in each class, while for interval-valued variables, the prototype of each class is defined as the interval that minimizes the Hausdorff distance to all the elements belonging to the class (Brito, 2015). The algorithm can be initialised with a random partition or a set of random prototype centers (Campos, 2007). How the algorithm works (adapted from Partners ASSO Developers (2004)) is described below.

Initialization: Let $P^{(0)} = (C_1^0, \dots, C_k^0)$ be the initial random partition of Ω in k classes.

Representation step t: $L^{(t)} = g(P^{(t)})$

For $h = 1, \dots, k$, compute a prototype $g^{(h)}$ representing cluster $C_h \in P^{(t)}$

Allocation step t: $P^{(t+1)} = f(L^{(t)})$

For all i , each object $x_i \in \Omega$ is assigned to the class C_h , iff $d(x_i, g^{(h)})$ is minimum:

$$C_h^{(t+1)} = \{x_i \in \Omega \mid d(x_i, g^{(m),t}) \leq d(x_i, g^{(j),t}) \forall j \neq m\}$$

Stopping rule or stability: If $P^{(t+1)} = P^{(t)}$ then STOP else GO TO previous step.

The DCLUST algorithm is also a module of SODAS software that was later implemented in R. It groups the symbolic objects in homogeneous clusters on the basis of a dissimilarity matrix. It clusters interactively a large set of symbolic objects into a pre-fixed number of classes. Its aim is to minimize a criterion that is based on the sum of dissimilarities between individuals belonging to the same cluster (Partners ASSO Developers, 2004). The next algorithm scheme, adapted from Partners ASSO Developers (2004), can be used to describe how DCLUST works.

Initialization: Let $P^{(0)} = (C_1^0, \dots, C_k^0)$ be the initial random partition of Ω in k classes.

Allocation step t: For all i , each object $x_i \in \Omega$ is assigned to the class C_h , iff $\sum_{x \in C_h^{(t)}} d(x_i, x)$ is minimum:

$$C_h^{(t+1)} = \{x_i \in \Omega \mid \sum_{x \in C_h^{(t)}} d(x_i, x) \leq \sum_{x \in C_m^{(t)}} d(x_i, x) \forall m \neq j\}$$

Stopping rule or stability: if $P^{(t+1)} = P^{(t)}$ then STOP else GO TO previous step.

3.1.3 Determination of the Number of Clusters

Determining the appropriate number of clusters is quite important when dealing with clustering algorithms. On the one hand, non-hierarchical algorithms usually require the user to pre-

define the number of clusters. On the other hand, in the hierarchical approach, the user needs to choose the number of clusters in the final solution (Hardy, 2007).

In Dudek (2007) some suggestions for using well-known indices for symbolic data are given and indices created specifically for symbolic data are described. Additionally, after presenting some simulation results, recommendations for selecting the most suitable indices for popular algorithms are given. The authors use three cluster algorithms: the Ward Hierarchical clustering method, the Partitioning Around Medoids technique and a Dynamical clustering for symbolic objects. By using those algorithms, eight indices are compared: the Silhouette index (S), the Baker and Hubert index (G2), the Hubert and Levine index (G3), the Calinski and Harabasz index (F), the Hartigan index (H), the Krzanowski and Lai index (KL), the Inertia for symbolic objects (SI) and the Homogeneity based quality index (SH). At the end, considering the results achieved by the three different algorithms, the indices G2 and G3 are selected as the best overall, so they are used for this work. If the order of results from best to worst is followed, the SH and SI indices are the next ones, however, considering what is currently available in the software packages it is not possible to directly implement these indices. Thus, Silhouette (that appears in the 5th position) was the third index to be selected for this work. Three indices were selected to make it easier to make decisions when the two indices already selected disagree.

The index G2, an adaptation of Goodman and Kruskal's Gamma statistic for clustering, is defined in the following way:

$$G2 = \frac{v(+) - v(-)}{v(+) + v(-)} \quad (3.3)$$

where $v(+)$ is the number of times a distance between two points that belong to the same cluster is strictly smaller than the distance between two points not belonging to the same cluster, while $v(-)$ is the number of times the distance between two points is in the opposite situation. For better results, the objective is that this index is close to one, maximum values are the better ones (Milligan & Cooper, 1985).

Expression 3.4, retrieved from Gordon (1999), shows how to compute the index G3.

$$G3 = \frac{D(k) - D_{min}}{D_{max} - D_{min}} \quad (3.4)$$

where $D(k)$ is the sum of all within-cluster dissimilarities in a partition of the objects into k clusters. If the partition has a total of r such dissimilarities, D_{min} (D_{max}) is the sum of the r smallest (largest) dissimilarities. Since the goal in a clustering algorithm is that the distances within each cluster are as small as possible, this index must have values as small as possible.

According to Rousseeuw (1987), the Silhouette index, S, is computed as follows:

For each point x_i the silhouette width of the point is computed.

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (3.5)$$

where $a(i)$ is the mean distance of point x_i to the other points of the cluster it belongs to and $b(i)$ is the smallest mean distance of x_i to the points of each of the other clusters. A value closer to

1, indicates that the point is affected to the right cluster, on the other hand, a value closer to -1, suggests that the point should belong to another cluster.

Then, the cluster mean Silhouette for a given class C_h is computed.

$$S_h = \frac{1}{n_h} \sum_{x_i \in C_h} S(i) \quad (3.6)$$

Lastly, the global Silhouette index, the mean of the mean silhouettes through all clusters is computed.

$$S = \frac{1}{H} \sum S_h \quad (3.7)$$

Given the clustering main objective, that is, finding a configuration that maximizes the overall similarity within each cluster and minimizes the similarity between clusters, the Silhouette values should be large, indicating that the clusters are well separated and internally homogeneous.

3.1.4 Variable Selection

To enhance the performance of the models, a subset of relevant variables may be chosen through a variable selection procedure. This choice is important to not use ‘noisy’ variables with minimal clustering information which can lead to inaccurate results. Besides, selecting the variables is important to improve the efficiency of the clustering process as it reduces the dimensionality of the dataset and it also enhances interpretability. In some situations selecting a smaller set of variables can make the clustering results easier to understand and interpret.

In Carmone Jr, Kara, and Maxwell (1999), the authors proposed an approach called the Heuristic Identification of Noisy Variables (HINoV), which can be used to identify possible noisy variables using the k-means algorithm. In Walesiak and Dudek (2008), an extended version of the algorithm for symbolic data, which is available in an R package (Dudek, Pelka, Wilk, & Walesiak, 2022), is proposed.

An index of agreement between pairs of cluster solutions is the basis of HINoV algorithm. More concretely, based on Milligan and Cooper (1986), it uses the Adjusted Rand Index (ARI), a similarity measure between two partitions. This index is an adaptation of the Rand Index (RI) as it accounts for the fact that some agreement between two clusters can occur due to chance (Hubert & Arabie, 1985). Before explaining how the measure is computed, it is important to introduce Table 3.2.

		Partition V					
		Class	v_1	v_2	...	v_C	Sums
Partition U	u_1		n_{11}	n_{12}	...	n_{1C}	$n_{1\cdot}$
	u_2		n_{21}	n_{22}		n_{2C}	$n_{2\cdot}$
	$\cdot$		$\cdot$	$\cdot$		$\cdot$	$\cdot$
	$\cdot$		$\cdot$	$\cdot$		$\cdot$	$\cdot$
	$\cdot$		$\cdot$	$\cdot$		$\cdot$	$\cdot$
	u_R		n_{R1}	n_{R2}	...	n_{RC}	$n_{R\cdot}$
	Sums		$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot C}$	$n_{\dots} = n$

Table 3.2: Notation for comparing two partitions
Adapted from Hubert and Arabie (1985)

The above contingency table contains the information on class overlap between two partitions U and V . n_{ij} denotes the number of objects that are common to classes u_i and v_j . $n_{i\cdot}$ and $n_{\cdot j}$ refer to the number of objects in classes u_i (row i) and v_j (column j), respectively. The ARI is computed in the following way:

1. Let n be the number of observations in the dataset.
2. Let P_1 and P_2 be two different partitions.
3. Let a be the number of pairs of observations that are in the same cluster both in P_1 and P_2 .
4. Let b be the number of pairs of samples that are in different clusters in P_1 and P_2 .
5. Compute the Rand Index (RI):

$$RI = \frac{a + b}{\binom{n}{2}} \quad (3.8)$$

6. Compute E , the expected value of the RI. It gives the expected number of pairs of objects that are placed in the same class in different partitions.

$$E = \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}} \quad (3.9)$$

where $n_{i\cdot}$ is the number of distinct pairs that can be constructed within rows and $\sum_i \binom{n_{i\cdot}}{2}$ is the number of distinct pairs that can be created from columns (see Table 3.2).

7. Compute the ARI:

$$ARI = \frac{RI - E}{\max(RI) - E} \quad (3.10)$$

According to Hubert and Arabie (1985), there is more than one possible way to determine the upper bound on the raw index ($\max(RI)$). To perform the HINoV algorithm, the R function HINoV.SDA from the symbolicDA package uses the following expression:

$$\max(RI) = \frac{1}{2} \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] \quad (3.11)$$

Thus, the final formula to compute the ARI is shown in Equation 3.12.

$$ARI = \frac{\frac{a+b}{\binom{n}{2}} - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}}{\frac{1}{2} \left[\sum_i \binom{n_{i\cdot}}{2} + \sum_j \binom{n_{\cdot j}}{2} \right] - \frac{\sum_i \binom{n_{i\cdot}}{2} \sum_j \binom{n_{\cdot j}}{2}}{\binom{n}{2}}} \quad (3.12)$$

The way the HINoV algorithm works is next described (adapted from Carmone Jr et al. (1999)).

1. Perform the selected cluster algorithm with all the p variables. This yields a preliminary estimate of the number of clusters, k .
2. Perform the clustering algorithm requesting k clusters, one variable at a time, for each of the p variables.
3. Compute the ARI for partitions formed from all distinct pairs of the p variables.
4. Construct a $p \times p$ matrix, PARIM, with the values of the Adjusted Rand indices.
5. Sum the values for each row of PARIM which give the total pairwise Adjusted Rand Index (TOPRI) over the partitions for each variable.
6. Rank TOPRI values and plot a diagram.

The size of the TOPRI values indicates the contribution of a variable to the cluster structure. The TOPRI values for noisy variables are expected to be significantly smaller than those for the non-noisy variables. The scree diagram allows for the identification of sharp changes in TOPRI values. The relatively low-valued TOPRI variables (corresponding to noisy variables) are eliminated from further analysis and the relatively high-valued TOPRI variables correspond to the ones that have the most relevant clustering information.

3.2 Cluster Monitoring

Monitoring the dynamics of cluster structures contributes to the sedimentation and/or formation of sustained knowledge about the phenomena underlying the data. For example, in the sociological area, this type of study can be used to understand how certain social groups evolve over time, which allows for deeper knowledge about societies and trends (Oliveira, 2010). Particularly, in this work, the aim is to detect changes in static data environments (snapshot comparison).

The MONIC framework (Spiliopoulou, Ntoutsis, Theodoridis, & Schult, 2006), later extended in Ntoutsis, Spiliopoulou, and Theodoridis (2009), and the MEC framework (Oliveira & Gama, 2010a) are two known approaches for the detection of cluster changes through time.

The MEC framework was developed on the foundations of the MONIC approach and extends it in two ways: it proposes a visualization tool in the form of a bipartite graph and it uses a different similarity function. Although in Oliveira and Gama (2010a) the visualization tool is referred to as a bipartite graph, there is room to improve this concept in the future. Nevertheless, for simplicity reasons, the original concept (the bipartite graph concept as it is in the original work) is used throughout this work. Additionally, in Oliveira and Gama (2010a) a standardization of the transition nomenclature found in eight different schemes is proposed. Another difference between the two frameworks lies in the fact that MEC is a more complete framework as it is not limited to clusters represented by enumeration, but also considers clusters represented by summary statistics (explained below) (Oliveira & Gama, 2010b).

The MEC framework aims at solving the problem of monitoring cluster transitions over time, through the identification of temporal relations between these structures. Moreover, in Oliveira (2010) two strategies to represent clusters are proposed: representation by enumeration and representation by comprehension. Furthermore, two methods to detect the transitions experienced by clusters, each adapted to one of the representation schemes, are proposed. When clusters are represented by enumeration (which is what this dissertation focuses on), the method to detect transitions is based on the use of bipartite graphs and conditional probabilities. For this work, it is suggested that oriented edges are considered (see Figure 3.2). Besides, when clusters are in the comprehension representation, the method to detect transitions is based on computing and evaluating the degree of similarity between them.

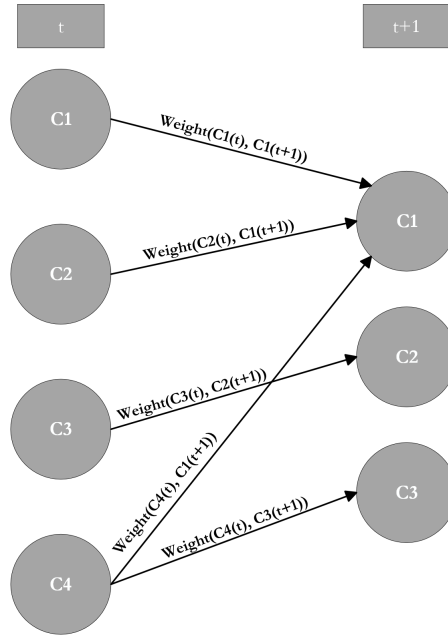


Figure 3.2: MEC Bipartite Graph representation

In the enumeration representation, a cluster is characterized by its members, that is, by the observations that have been assigned to it by a given clustering algorithm. Formally, it is defined as follows:

Let $\vec{\mathbf{x}}_i$, be the i -th observation ($i = 1, \dots, n$), denoted as a vector of numerical attributes in a p -dimensional space ($j = 1, \dots, p$), $\vec{\mathbf{x}}_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})$. A possible temporal representation of clusters can be denoted as follows:

$$C_h(t) = \{\vec{\mathbf{x}}_1, \dots, \vec{\mathbf{x}}_{n_i}\}$$

where n_i represents the number of observations assigned to cluster $C_h(t)$, $h = (1, \dots, k)$ and $t = (1, \dots, T)$.

In Oliveira (2010) the MEC framework for clusters in enumeration representation is defined in the following way:

Given the partitions P_i, P_j ($i < j$), obtained at t_i, t_j , a graph $G = (U, V, E)$ can be constructed, where U represents the first subset of vertices (clusters at time t_i), V represents the second subset of vertices (clusters at time t_j), and E denotes a set of weighted edges between any pair of clusters belonging to P_i and P_j . Formally, the weight assigned to the edge connecting clusters $C_m(t_i)$ and $C_u(t_j)$ ($m = (1, \dots, k_{t_i})$ and $u = (1, \dots, k_{t_j})$), where k_{t_i} and k_{t_j} are the number of clusters returned by a given clustering algorithm at time points t_i and t_j , respectively, are estimated in accordance with the conditional probability:

$$weight(C_m(t_i), C_u(t_j)) = P(X \subset C_u(t_j) | X \subset C_m(t_i)) = \frac{|C_m(t_i) \cap C_u(t_j)|}{|C_m(t_i)|} \quad (3.13)$$

where X is the set of observations assigned to cluster $C_m(t_i)$ and $P(X \subset C_u(t_j) | X \subset C_m(t_i))$ represents the probability of X being included in cluster C_u at t_j given that X forms the cluster C_m obtained at a previous timestamp t_i .

Changes that take place in the overall clustering are referred to as exogenous transitions. The key concept in detecting and evaluating these transitions consists in discovering the exact correspondences between the clusters at time instant t_i and the clusters at a later time instant t_{i+1} , in case they still exist. When the clusters are represented in enumeration form, the correspondence is done in terms of objects.

As previously mentioned, in Oliveira (2010) a new taxonomic scheme, based on eight others that previously existed, for the typification and classification of cluster exogenous transitions, is proposed. Before describing it, the following two concepts need to be introduced.

- Survival threshold (τ): The minimum linkage weight between two clusters at different moments so that they can be considered an approximate correspondence of each other.
- Splitting threshold (ρ): Minimum linkage weight between clusters at different temporal moments useful in clarifying which situations should or should not be considered as a cluster split.

In this work and according to what is considered in Oliveira (2010), the following taxonomy to detect changes is taken into consideration.

- **Birth** of a cluster occurs when the weights associated with all the edges connected to this cluster are lower than the survival threshold.
- **Death** of a cluster occurs when the links with the clusters of the following instant present weights lower than the survival threshold, with the exception of the cases foreseen in the split clusters definition.
- When a cluster splits into at least two pieces at the following instant and the weights of each of the individual edges are equal to or greater than the splitting threshold and less than the survival threshold and their sum is equal to or greater than the survival threshold, the cluster is said to have **split**.
- **Merge** occurs when the weights of the links from at least two clusters at one time instant to one given cluster at the subsequent time instant are equal to or greater than the survival threshold.
- **Survival** happens when a link between two clusters that are separated in time has a weight equal to or greater than the survival threshold and there are no other edges with a weight greater or equal to the survival threshold in each one of the clusters.

The formal definition of the exogenous transitions can be found in Appendix A, Table A.1.

Additionally, if the exogenous transition is survival, the MEC framework also allows identifying endogenous transitions which are changes related to the content of each cluster alone. For this purpose, two types of transitions are considered: size transitions and compactness transitions. The former predicts the Expansion or Shrinkage of the cluster, depending on whether the cardinality of the surviving cluster increases or decreases. Compactness transitions include Compression or Dispersal of the cluster depending on the increase or decrease in the density of the surviving cluster. In Oliveira (2010), the endogenous transitions are only formally described for clusters represented by comprehension. Nevertheless, for the present work, the concepts were easily adapted to clusters represented by enumeration when using a hierarchical algorithm. In order to detect compactness transitions, the height of the dendrograms (which were created using the same linkage) was observed. More specifically, if the height of a specific class in the dendrogram increases from one year to the next, it means that there is greater dispersion in that class. On the other hand, if the height decreases, it means that there was a compression. It may also be the case in which no endogenous transition is detected. The notation of these transitions can be found in Appendix A, Table A.2.

Figure 3.3 aims at illustrating some of the transitions already introduced. In the figure, three consecutive time instants are represented, t , $t+1$ and $t+2$. From t to $t+1$, it is possible to observe a birth (yellow cluster in $t+1$), a merge (blue and beige clusters of t), a split (red cluster in t) and the survival of the green cluster. Moreover, from $t+1$ to $t+2$, all the clusters survived except the green that died. Regarding endogenous transitions, it is possible to verify that the yellow cluster has expanded and become more dispersed, meaning that its cardinality has increased and density

has decreased. The red and brown clusters saw an increase in cardinality. On the other hand, the blue cluster has shrunk, due to a slight reduction in the number of observations that compose it.

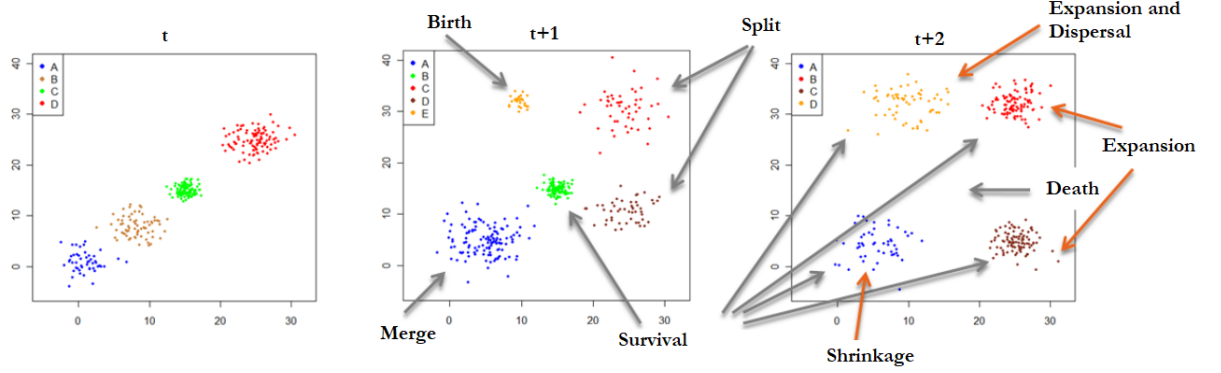


Figure 3.3: Representation of some endogenous and exogenous transitions

(Adapted from Oliveira (2010))

Briefly, the steps for the MEC framework application for clusters represented by enumeration are the following:

- Establishing the survival and split thresholds. In this work, and according to what was considered most reasonable in Oliveira (2010), $\tau = 0.5$ and $\rho = 0.2$.
- Application of the MEC algorithm.
- Bipartite graph design.

The MEC approach is easy to apply and delivers findings that are simple to interpret; nevertheless, there is a major drawback: it imposes that the set of objects remains constant across time. Nonetheless, since Symbolic Data Analysis is used and the objects created are very unlikely to disappear this drawback has no direct implications for the work carried out. Nevertheless, in the very unlikely situation that any of the objects disappear from one year to the next, some extensions to the method are proposed. Those extensions include adding extra arrows to the different graph nodes. In case objects appear in a cluster at t_{i+1} that were not present at t there is no modification in the algorithm, just in the graph representation: an unweighted entering arrow is drawn. When objects are present at t and disappear at t_{i+1} , one may imagine that there is an extra cluster to which the missing objects have gone to. The weight of the link between each cluster and the new extra cluster is calculated by the expression 3.13. Regarding the graph representation, an exiting arrow is drawn. These changes ensure that the sum of the weights of edges and arrows leaving a cluster is equal to 1.

Take the following example as an illustration of the changes. Consider that at t , a clustering algorithm came up with four clusters, namely, Cluster 1: A, B, C, D; Cluster 2: E, F, G; Cluster 3: H, I, J; Cluster 4: K, L, M, N, and at t_{i+1} , the algorithm splits the objects into 3 groups,

specifically: Cluster 1: A, B, C, E, F, G, M, N; Cluster 2: H, I, J, O; Cluster 3: K, L. One may conclude that D is an object that does not appear at t_{i+1} , so an exiting arrow, with a weight equal to 0.25 (computed using expression 3.13) is drawn in the graph. Furthermore, O is an object that does not appear at t , but appears at t_{i+1} so an entering arrow to Cluster 2 is drawn. Figure 3.4 serves to exemplify the transitions just described.

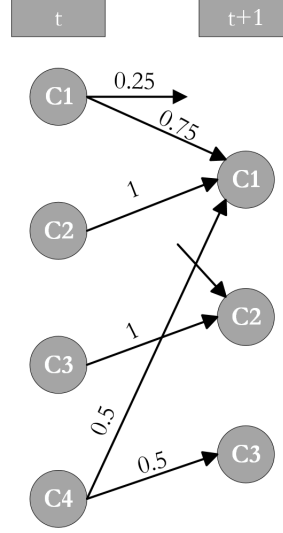


Figure 3.4: Bipartite Graph applying the extension proposed for the MEC framework

In our context, the MEC framework presented above is key for understanding which immigrants (considering the place of birth and the length of residence in Luxembourg) have similar profiles in the labour market and how trends are evolving. In addition, this methodology allows one to see whether there were many or few movements of objects from one year to the next. These movements may be triggered by some situation at country or world level (e.g. creation of immigration laws, pandemic, ...), but they may also be the consequence of changes in the methodology, in this case, in the LFS methodology. Thus, after obtaining the results of the transitions, it is up to the researcher to try to understand which kind of situations may have triggered the movements.

Chapter 4

Data and Software

4.1 Software

As already stated, R (R Core Team, 2022) was the software selected to do the analysis proposed.

Although some changes in the source codes were required, the `symbolicDA` package (Dudek et al., 2022) serves as a base for a major part of the work. Other important packages include `VIM` (Kowarik & Templ, 2016) to handle the missing values, `RSDA` (Rodríguez, Calderon, & Zuniga, 2014) mainly used for the creation of a symbolic data table, and `fpc` (Hennig, 2023) used for the validity indices.

4.2 Variables Description and Pre-Processing

The data used for the work is the annual Luxembourgish LFS from 2014 to 2022. The years prior to 2014 are not covered due to some methodological changes. Including those years would require more time to process the corresponding data, so that it would be in concordance with recent data. As mentioned in Section 2.3, it consists of demographic data and data regarding a person's position in the labour market.

As previously stated, the major aim of this work is to study the labour market profile of the immigrant population. Therefore, the cohort of people selected are the employees since for all nine years they represent around 90% of the employed population. Furthermore, there is a wider range of variables, including the earning deciles, that are asked to this specific group of workers and not to all the workers.

Before describing the variables selection process and providing a description of each one, a remark needs to be made. Even though the meaning is the same, some Eurostat variables' names changed from 2021 onwards. In this work the name that is used is the most recent one.

Regarding the choice of the variables, almost all Eurostat's variables (variables from the EU-LFS) that are described in the EU Labour Force Survey Explanatory Notes (European Commission and Eurostat, 2013, 2016a, 2018a, 2021a) and are asked to the defined target population (the employees), from 2014 to 2022, make part of the analysis. Nevertheless, a few are excluded *a*

priori for different reasons. Some are not very interesting for the Luxembourgish case (COUNTRYW¹ and REGIONW²), others are similar to variables that were selected (HOMEWORK³, HWWISH⁴, HWUSUAL⁵, EMPSTAT⁶), some variables are not of interest for the analysis, more concretely the ones that are related to absences, and finally, there are two that were excluded since they are binary variables and the classes are really unbalanced (TEMPAGCY⁷ and TEMP⁸)⁹. In addition to the questionnaire that is done by Eurostat, Statec asks some extra questions and collects some more variables and two of them were also selected. Therefore those two variables are not harmonized which does not imply that the work is not easily replicated in other countries where LFS is used.

Even though in the LFS data the majority of variables are categorical, there are also continuous variables. For this work, the latter were discretized. Discretization is an important data pre-processing technique used in knowledge discovery processes. The goal of discretization is to transform continuous variables into discrete ones which may lead to better algorithm performance (Hacibeyoglu & Ibrahim, 2018). In addition, taking into account what is currently available in R, if the discretization was not done, the only option would be to transform the continuous variables into interval ones. In this way, some information could be misleading. For example, in the presence of an outlier, the interval would give a wrong idea of what is happening in a certain group. The interval would have a large range, just because of one element that belongs to that group. For the present work, the discretization is made in such a way that, for each year, the categories of the new variable have roughly the same frequency.

Based on European Commission and Eurostat (2013, 2016a, 2018a, 2021a), the list below gives a brief description of the variables that are used throughout the analysis.

- **IDENT**: Unique identifier (25 character alphanumeric).
- **TWANNUAL**: weight that is assigned to each individual. It is computed as described in Subsubsection 2.3.1.
- **INTWAVE**: Sequence number of the survey wave (for this work the interest is in waves 1 and 5).
- **HHSEQNUM**: 2 digit (01-98) Sequence number allocated to each member of the household (when the sample is the individual, 01 is assigned to the originally selected person).

¹Country of place of work for the main job

²Region of place of work for main job

³Working at home for the main job

⁴Number of hours that the person would like to work in total in a week

⁵Number of hours per week usually worked in main job

⁶Being in employment

⁷If a person has a contract with a temporary employment agency for the main job

⁸Permanency of main job

⁹As an example, without replacing the missing values, for the 2022 sample, we had:

TEMP=1 (93.38%) and TEMP=2 (6.31%)

TEMPAGCY=1 (1.86%) and TEMPAGCY=2 (97.80%)

- **HHLINK:** Relationship to the reference person in the household.
 - 01: Reference person (what this work mainly focus on)
 - 02: Partner of reference person
 - 03: Son/daughter of reference person
 - 04: Son/daughter-in-law of reference person
 - 05: Grandchild of reference person
 - 06: Parent of reference person
 - 07: Parent-in-law of reference person
 - 08: Grandparent of reference person
 - 09: Brother/sister of reference person
 - 10: Other relative
 - 11: Other non-relative

- **SEX:** Biological and physiological characteristics that define a person to be either male or female.
 - 1: Male
 - 2: Female

- **AGE:** Age in completed years. Considering the distribution for each year, the original variable was grouped according to the following:
 - 1: $15 \leq \text{AGE} \leq 29$
 - 2: $29 < \text{AGE} \leq 35$
 - 3: $35 < \text{AGE} \leq 40$
 - 4: $40 < \text{AGE} \leq 45$
 - 5: $45 < \text{AGE} \leq 50$
 - 6: $50 < \text{AGE} \leq 55$
 - 7: $\text{AGE} > 55$

- **COUNTRYB:** Country of birth. The original Eurostat variable represents for each observation the country where the person was born. Since there are countries that are very poorly represented, that is, in the sample there are only one or two people who were born in that country, the variable is grouped as presented below. Some work already done (see Section 2.2) on immigrants in Luxembourg contributed to the definition of the categories.
 - LU: People born in Luxembourg (the natives)
 - PT: People born in Portugal
 - OUTEU: People born in countries outside the EU
 - EU: People born in countries that belong to the EU
 - NC: People born in the neighbouring countries (France, Germany and Belgium)

Some remarks need to be made regarding this variable. Until 2019, people born in the United Kingdom fall in the EU category, nevertheless since the Brexit, they fall within the

OUTEU category. From now on, for simplicity reasons, even though Portugal, Luxembourg and the Neighbouring countries belong to the EU, when EU countries are mentioned, these five countries are excluded. This variable is crucial for the whole analysis that is presented since the main focus is to analyse and compare the profile of immigrants in the labour market and also to compare it with the profile of natives. The immigrants are distinguished by their place of birth, in this case, EU countries, countries outside the EU, Portugal or Neighbouring countries.

- **CITIZENSHIP:** Country of main citizenship (if a person has Luxembourgish Nationality it is automatically the main one). The same thing occurs with this variable as it did with the previous one, meaning that, there are nationalities that are very poorly represented in the sample. However, since this variable is only used to handle the missing values (explained later in Subsection 4.2.1), it is used in its original version meaning that for each observation it represents the main nationality of the person.
- **DEGURBA:** The degree of urbanisation of the area where the person has his/her usual residence.
 - 1: Cities
 - 2: Towns and Suburbs
 - 3: Rural areas

- **YEARESID:** Duration of stay in the country of residence in completed years. Considering the variable's distribution for each year, the original variable is grouped according to the following:

- 1: $0 \leq \text{YEARESID} \leq 4$
- 2: $4 < \text{YEARESID} \leq 8$
- 3: $8 < \text{YEARESID} \leq 15$
- 4: $15 < \text{YEARESID} \leq 24$
- 5: $\text{YEARESID} > 24$
- 6: People born in Luxembourg

- **ILOSTAT:** Determines the internationally comparable ILO labour status according to International Labour Organization (2013) (the 19th ICLS resolution differentiating between employed persons, unemployed persons and those outside the labour force).

- 1: Employed
- 2: Unemployed
- 3: Outside the labour force

Prior to 2021, this variable was not part of Eurostat's variables, however, it can be constructed using other variables as it is explained in European Commission and Eurostat (2021a).

The variables described until now are asked to everybody in the original sample. Nonetheless, the next ones (except HATLEVEL and MIGREAS) are asked to all people whose

age is between 15 and 89. More concretely, they are either asked to all people who are employed or to the more specific case of people whose status in the main job is employee.

- **NUMJOB:** Number of jobs. Before 2021 it only had two categories, after 2021 it has three. Thus, the following categories were used in all years.
 - 1: Only one job
 - 2: More than one job
- **WISHMORE:** Wish to work more than the current number of usual hours.
 - 1: No
 - 2: Yes
- **TELEARB** (Statec variable): Ability to do remote work.
 - 1: Yes
 - 2: No
- **STAPRO:** Status in employment in main job.
 - 1: Self-employed person with employees
 - 2: Self-employed person without employees
 - 3: Employee (what this work focus on)
 - 4: Family worker (unpaid)
- **NACE3D:** Economic activity of the local unit for the main job. The original variable is the *Nomenclature statistique des activités économiques dans la Communauté européenne* (NACE) code at 3-digit level, which is a very refined division. Therefore, based in Rota and Larue (2021), this variable is grouped according to the division classification that can be found in European Comission (2008), resulting in:
 - 1: Agriculture
 - 2: Industry
 - 3: Construction
 - 4: Trade, transport and hospitality
 - 5: Information and communication
 - 6: Financial and insurance activities
 - 7: Real state activities
 - 8: Business services
 - 9: Public administration, defense, education, health and social work
 - 10: Arts, entertainment and recreation and other service activities
- **ISCO4D:** Occupation in the main job classified according to the categories set out by the ISCO-08 classification up to a 4-digit level (European Comission, 2009). As what happens with NACE3D variables, the original variable has a very refined division. Therefore, based on some Eurostat and Statec works, for example, Eurostat (2018) and Ametepe et al. (2022), the original categories are grouped in the following way:

- 1: White-collar high-skilled professions (includes legislators, senior officials and managers, professionals, and technicians and associate professionals)
- 2: White-collar low-skilled professions (includes clerks, and service workers and shop and market sales workers)
- 3: Blue-collar high-skilled professions (includes skilled agricultural and fishery workers, and craft and related trades workers)
- 4: Blue-collar low-skilled professions (includes plant and machine operators and assemblers, and elementary occupations)

More detailed information about this kind of professions may be found in European Commission (2009). With this grouping, the armed forces are excluded from the analysis. Nevertheless, they represent only a residual share of Luxembourg workers.

- **FTPT**: Full- or part-time main job (self-defined).
 - 1: Full-time job
 - 2: Part-time job
- **LOOKOJ**: Looking for another job.
 - 1: Person is not looking for another job
 - 2: Person is looking for another job
- **HATLEVEL**: Highest International Standard Classification of Education (ISCED) Educational attainment level (highest level of education successfully completed). For the current work, the variable is grouped as follows:
 - 1: Until High School (included)
 - 2: Any level between High School and Bachelor (included)
 - 3: Master or PhD

The variable categories are different before and after 2021, so to harmonize it, before grouping them, European Commission and Eurostat (2021b) is used.

- **YSTARTWK**: How many years ago did the person start working for the current employer. The discretization of this variable is done in the following way:
 - 1: $YSTARTWK < 5$
 - 2: $5 \leq YSTARTWK \leq 15$
 - 3: $YSTARTWK > 15$
- **SIZEFIRM**: Size of the local unit for the main job. The codes for this variable changed in 2021, to have the same categories throughout all years the next codification is used:
 - 1: Small-Medium Firm (less than 50 workers)
 - 2: Large Firm (50 workers or more)
 - 3: Do not know

Category 3 includes not only the people that do not know the size of the firm in which they work as well as the people that did not answer.

- **EARNINGDEC:** Monthly pay from the main job. Until 2021 there was the variable IN-CDECIL which corresponded to the monthly (take-home) pay (the pay after the deduction of income tax and National Insurance Contributions). Nonetheless, since 2021 the variable that is collected is the INCGROSS which corresponds to the person's gross monthly pay (the pay before the deduction of income tax and National Insurance Contributions after the deduction of employers' social insurance contributions) from the main job. To make these two variables comparable across time, the EARNINGDEC variable was created for this work. To create it, the deciles were used and the next codification was followed.

- 1: Person earns a value corresponding to the first decile or less
- 2: Person earns a value between the first and the second (included) deciles
- 3: Person earns a value between the second and the third (included) deciles
- 4: Person earns a value between the third and the fourth (included) deciles
- 5: Person earns a value between the fourth and the fifth (included) deciles
- 6: Person earns a value between the fifth and the sixth (included) deciles
- 7: Person earns a value between the sixth and the seventh (included) deciles
- 8: Person earns a value between the seventh and the eighth (included) deciles
- 9: Person earns a value between the eighth and the ninth (included) deciles
- 10: Person earns a value between the ninth and the tenth (included) deciles

Additionally, code 1 means that a person is in the lowest 10% of earners, code 2 means the next 10%, and so on, with code 10 indicating that a person is in the highest 10% of earners. For simplicity reasons, throughout this work people whose value in variable EARNINGDEC is between one and three are referred to as lower-income earners. If the value of the variable is between four and seven they are called middle-income earners. Lastly, people whose value in variable EARNINGDEC is eight or more are called upper-income earners.

- **SUPVISOR:** Supervisory responsibilities in main job.

- 1: Yes
- 2: No

- **EARNSATISF** (Statec variable): Satisfaction with the earnings.

- 1: Satisfied with the earnings
- 2: Not satisfied with the earnings

- **MIGREAS:** Main reason for migrating. This variable was only introduced in 2021 and it is collected every two years. For this work, it is coded in the following way:

- 1: Employment (with or without job found before migrating)
- 2: Family Reasons
- 3: Education or Training
- 4: Retirement
- 5: International Protection or Asylum
- 6: Other

4.2.1 Missing Values

In the nine years of analysis, variables always have missing values. The percentage of missing values is higher in some years than in others, however, the EARNINGDEC and NACE3D variables are the ones that always have the highest percentages.

Missing data occurs frequently in different kinds of studies. Nevertheless, they lead to some difficulties, pointed out in Barnard and Meng (1999), such as complications in handling and analysing data. These complications lie in the fact that most methods and algorithms (for example clustering algorithms) have no internal way to handle missing values.

Among the various possibilities for dealing with the problem of missing values, the most commonly used, according to Gama, Carvalho, Faceli, Lorena, and Oliveira (2017), are:

- Delete the objects with missing values. For the present work, this technique is not very appropriate since if all objects with missing values were deleted, very few objects would remain.
- Define and fill manually. This approach is not feasible considering the total number of missing values that exist in the data.
- Use some method or heuristic to set the values for the missing attributes automatically. This automatic definition has followed three different approaches:
 - Create a new value for the attribute that indicates that the attribute had an unknown value. However, the algorithm used in the analysis may assume that this new category is an important concept.
 - Use the mean, mode or median. Using any of them is very likely to create inconsistencies.
 - Use an inducer to estimate the attribute value. This is the most popular approach and it has the advantage that it uses information from the other variables to infer the missing value.

To handle the majority of missing values and taking into account the strengths and weaknesses of the techniques described above, an inductor to estimate missing values is the approach selected, more precisely, the 3-Nearest Neighbour (3NN) algorithm. It works as follows:

1. For each missing value, find the three nearest neighbours based on the values of the other variables in the dataset. The distance measure used in this work is a variation of the Gower distance (Gower, 1971; Kowarik & Templ, 2016).
2. Calculate the value to impute for the missing value based on the values of its 3-nearest neighbours. Since all the variables are categorical, the method to choose the value to be imputed is the mode among the selected neighbours. If all the values from three neighbours are different, one is randomly selected.

Nonetheless, there are three exceptions. For the variable COUNTRYB, the CITIZENSHIP variable is used as a proxy. Regarding variable SIZEFIRM, as mentioned previously, there was already an unknown category, so the missing values are classified in that category. Lastly, if the

COUNTRYB variable is equal to LU, the missing values in the YEARESID variable are replaced by the number 6.

It is important to note that, even though the selected technique uses the information from the other variables and that it is the most commonly used approach, some inconsistencies can also be produced. In this work, special attention was given to the observations where, after the imputation of the missing values, the country of birth (COUNTRYB variable) is different from LU and YEARESID is equal to six. Taking into account that these situations are not very recurrent, those observations were eliminated.

4.3 Sample Description

This section aims at providing a short description of each year's sample. The size of each year's sample, as well as the number of employees in the original sample, is shown in Table 4.1.

Year	Sample Size	Number of Employees
2014	13831	2781
2015	27703	5546
2016	31225	3444
2017	29616	3920
2018	27936	3977
2019	30868	4488
2020	30283	4595
2021	23228	4645
2022	19634	3882

Table 4.1: Sample Characterization 2014-2022

The sample of employees used for the cluster analysis ended up being slightly smaller than the original sample since some observations were removed due to inconsistencies. In Table 4.2, the employees' final sample size as well as the number of individuals in each COUNTRYB group are presented.

4.4 Descriptive Analysis

In this section, a characterization of the Luxembourgish society is made as well as a brief description of the composition of the five groups that are the main focus of this work.

Since 2014 the population of Luxembourg has been on the rise (Figure 4.1). It is important to note that using the data from the LFS, the total population is being underestimated since the survey only considers private households. This means that, for example, elderly people in retirement homes and inmates in correctional facilities are excluded.

Taking a closer look at the changes that occurred in the four immigrant groups (see Figure 4.2), one may conclude that there have been some ups and downs.

Year	EU	LU	NC	OUTEU	PT	TOTAL
2014	288	1458	580	195	257	2778
2015	605	2760	908	492	761	5526
2016	362	1782	547	303	440	3434
2017	433	2031	633	337	479	3913
2018	460	1967	684	352	500	3963
2019	570	2198	763	433	511	4475
2020	516	2289	747	519	517	4588
2021	561	2237	758	554	521	4631
2022	436	1831	641	494	409	3811

Table 4.2: Final Employees Sample Characterization 2014-2022

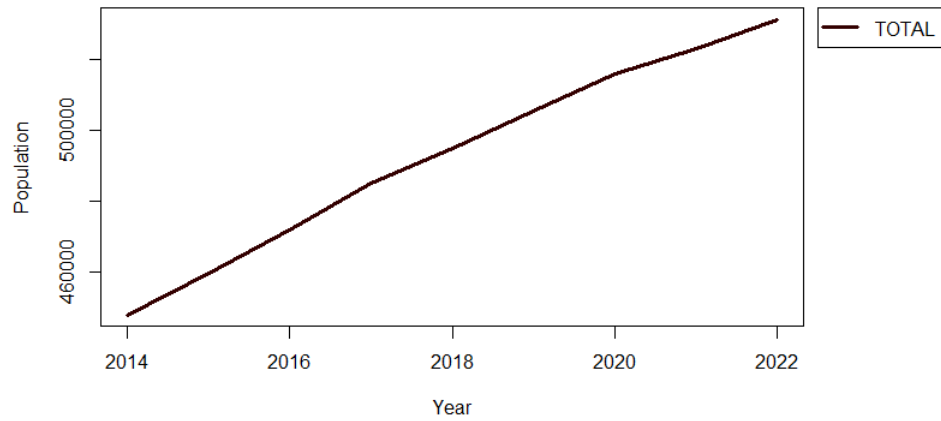


Figure 4.1: Total Population Evolution 2014-2022

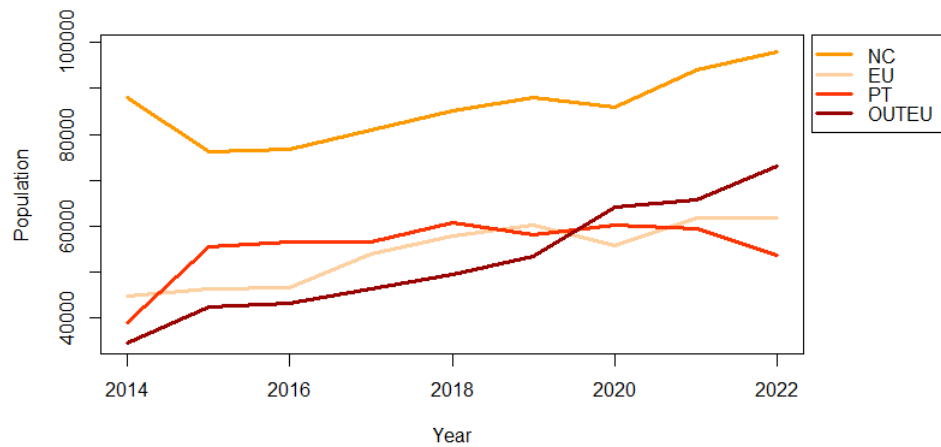


Figure 4.2: Immigrant groups total population evolution 2014-2022

At the moment, it is estimated that the Portuguese community is declining. Moreover, the OUTEU community has been on the rise since 2014 with some years having more marked increases than others. From 2014 to 2015, the marked increase in the OUTEU community coincided with the European Migrant crisis also known as the Syrian Refugee crisis. This led to an increase in the number of asylum seekers in Luxembourg (see Statec (2016)). From 2019 to 2020, the increase coincides with the Brexit, which led to the movement of UK-born people from the EU group (where in the plot one may see a decline) to the OUTEU group. Additionally, from 2021 to 2022, the increase in the number of people born in a non-EU country coincided with the war in Ukraine which prompted the arrival of Ukrainian refugees in Luxembourg. Since 2015, there has been an increase in Luxembourg residents who were born in a neighbouring nation, but in 2020, there was a fall from which they recovered right after in 2021.

By observing Table 4.3, it is possible to conclude that it is estimated that the percentage of people that lives in Luxembourg that were actually born in the country is declining and, nowadays it constitutes less than half of the population of the country. Nevertheless, the percentage of people whose main citizenship is Luxembourgish has remained more or less constant through time. As previously seen, Luxembourg's total population has been growing, and the fact that the proportion of people who were born abroad has also been growing leads one to believe that Luxembourg's immigration phenomenon is still growing. Additionally, among the people living in Luxembourg, it is possible to distinguish several distinct countries of birth. Although declining, the Portuguese community has almost always been the largest one in the analysed period.

Year	Luxembourg Born	Luxembourg National	Number of distinct countries of birth	Largest Immigrant community
2014	53.80%	54.90%	118	French (9.99%)
2015	51.18 %	54.30%	133	Portuguese (12.07%)
2016	51.80%	53.40%	119	Portuguese (12.00 %)
2017	50.23%	52.60%	120	Portuguese (12.68 %)
2018	48.09%	52.30%	125	Portuguese (12.29 %)
2019	47.50%	52.70%	142	Portuguese (11.50 %)
2020	46.90%	52.70%	139	Portuguese (11.69 %)
2021	45.91%	52.70%	137	Portuguese (11.39 %)
2022	45.87%	52.70%	142	Portuguese (10.11 %)

Table 4.3: Population Characterization 2014-2021

As seen in Section 2.2, the main reason for immigration to Luxembourg is work. In Figure 4.3, one may see that in 2021 that was still true as in three out of the four immigrant groups, more than half of the people migrated for employment reasons. Nonetheless, for people born in non-EU countries that is not the case. By performing a χ^2 test with the following hypothesis:

- H_0 : the migration reasons are independent of the place of birth.
- H_1 : there is an association between the reasons for migrating and the place of birth.

followed by an adjusted residuals analysis, for 5% significance level, it was possible to conclude that there is a strong association between being from Portugal and migrating for employment reasons and an even stronger association between being from a country outside the EU and migrating for International Protection or Asylum (see Figure B.1 in Appendix B).

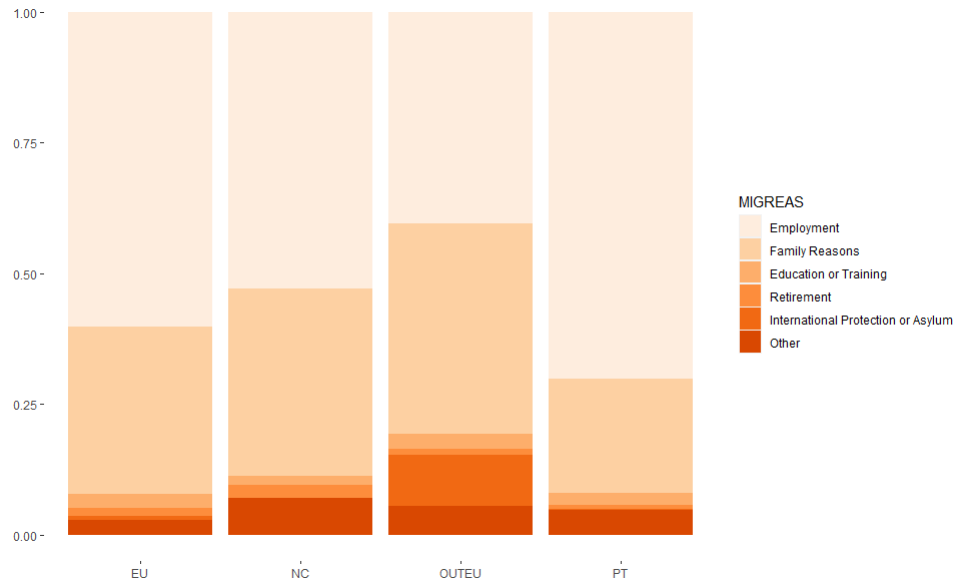


Figure 4.3: Migration Reasons 2021

Since work is a very strong reason for immigration to Luxembourg, something of interest is to analyse the evolution of employment and unemployment over time and across groups. The Unemployment and Employment rates from 2014 to 2022 are plotted in Figures 4.4 and 4.5, respectively. For the calculation of the rates, the Equations 2.2 and 2.1 are used, and the age class 15 to 74 years is selected.

From Figure 4.4, one may see that it is estimated that the unemployment rate is much larger for the people who were born in a non-EU country in comparison to the other groups and the total. In 2020 (the first year of the pandemic) it reached almost 16%. On the other hand, it is possible to conclude that the unemployment rate for the natives and for people from the NC has almost always been lower than the total unemployment rate. Regarding the Portuguese community, in the past, the percentage of unemployed people was larger or very close to the total, however in the last couple of years it declined, and in 2022 it was the lowest unemployment percentage among the five groups. Among the EU people, the unemployment rate has a similar behaviour to what happens in the Portuguese community, in the past it was larger or very close to the total, and since 2020 it is below the total.

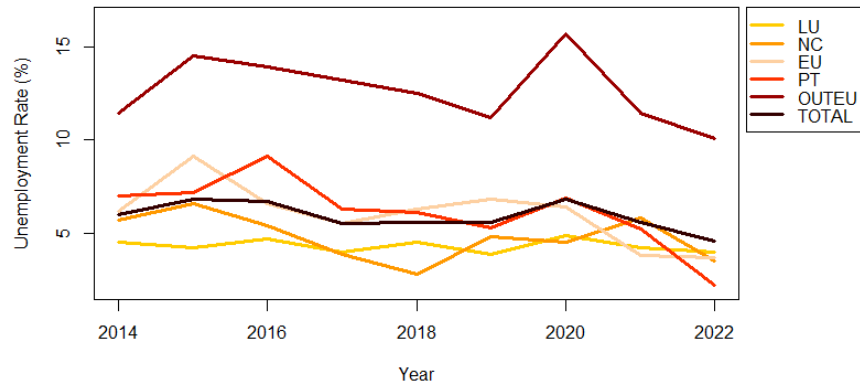


Figure 4.4: Unemployment Rate 2014-2022

Regarding the Employment rate, the lowest levels are found among the Luxembourgish natives. This fact may be explained by the way the rate is computed, that is, the employed population divided by people of working age which includes people since 15 years old. By the plot in Figure 4.6 (in the previous years a similar thing happens, the plots can be found in Appendix B, Figures B.2 to B.9) one may see that at younger ages (from 15 until a little bit after 25), the densest group is the Luxembourgish. All these people, given their age, may still be studying and are already contributing to the working-age population (and so a large denominator). However, most of them may not have entered the labour force yet (and so a smaller numerator). In addition, for the ages 25 to around 70, foreign people predominate and when they arrive in the country, they are most likely to go there to work and not study. Moreover, it is also possible to verify that closer to the retirement age the people who were not born in Luxembourg start leaving.

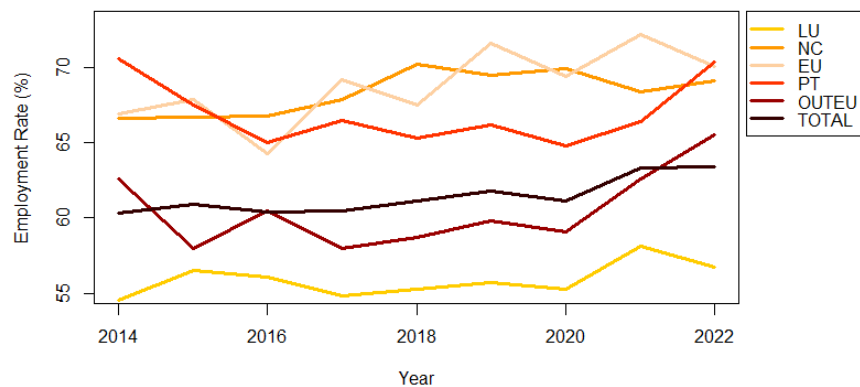


Figure 4.5: Employment Rate 2014-2022

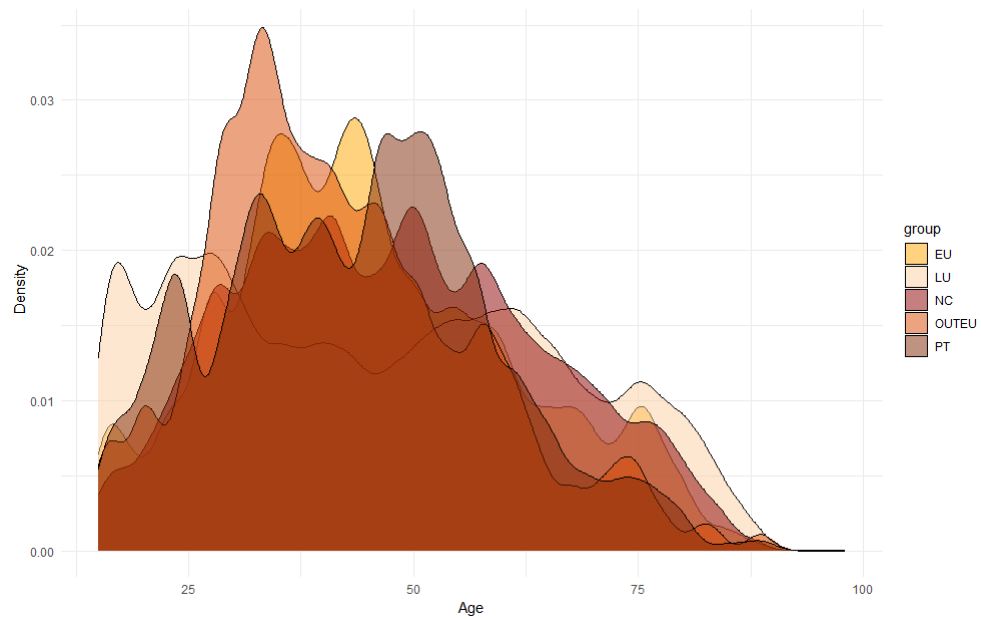


Figure 4.6: Age Density Plot per Group 2022

Chapter 5

Models

In this chapter the aggregation of the micro-data into symbolic data as well as the process of selection of the clustering algorithm are detailed.

5.1 From Microdata to Symbolic Data

After pre-processing all micro-data, two classical variables, YEARESID and COUNTRYB, were selected to aggregate data and create the second-order objects.

The variable COUNTRYB was selected since it allows researchers and policymakers to compare the experiences of the different groups and identify patterns and disparities in the labour market outcomes. This is very important for the development of policies and programs that target the specific needs and challenges of the different groups. To see if the profile of immigrants varies with how long they have been living in the country, in other words, to compare the immigrant profile of recent immigrants with the profile of the more established ones, the YEARESID variable was selected. By crossing these two variables, 21 symbolic objects were created.

For the next procedures, some variables that are detailed in Section 4.2 were no longer relevant, for different reasons. Those reasons include the fact that some of them are just relevant to choose the target population, for example, HHLINK and STAPRO and the fact that some of them are not asked every year, for example, MIGREAS. Consequently, the 21 symbolic objects are primarily described by 16 variables: SEX, AGE, DEGURBA, NUMJOB, WISHMORE, TELEARB, NACE3D, ISCO4D, FTPT, LOOKOJ, HATLEVEL, YSTARTWK, SIZEFIRM, EARNSATISF, EARNINGDEC and SUPERVISOR.

For every year, a symbolic data table was created. In Figure 5.1 it is possible to observe an extract of a symbolic data table from 2022. Briefly, using the extract as an example to explain how to interpret the symbolic data tables, five classical variables were selected to construct it. First, the variables YEARESID and COUNTRYB were selected to construct the different classes (symbolic objects). The classes correspond to the cartesian product of the categories of the variables selected. In the first column, the numbers 1 to 6 are the codes for the variable YEARESID, while the EU, NC, OUTEU, PT and LU are the categories of the variable COUNTRYB. For example, the first symbolic object 1.EU represent the people who were born in an EU country and have been living in Luxembourg for four years or less. In total 21 symbolic objects were created. The

other three variables: DEGURBA, TELEARB and NACE3D are categorical variables and are aggregated into categorical modal variables. As previously mentioned in Section 4.2, DEGURBA has three categories, TELEARB two and NACE3D ten. Picking up the first line as an example (the object 1.EU) and analysing the results in the table, one may say that in 2022, 55.3% of the people who were born in an EU country and have been living in Luxembourg for four years or less lived in Cities, while 25% lived in Towns and Suburbs and 19.7% in Rural Areas. Furthermore, it is also possible to verify that 67.1 % did telework. Lastly, the most common activity sectors for this group were 4 (Trade, transport and hospitality), 6 (Financial and insurance activities) and 10 (Arts, entertainment and recreation and other service activities). About 65.9% of the people in group 1.EU worked in one of these sectors in 2022. A similar analysis can be done for the other symbolic objects presented in the table.

	\$H	DEGURBA	1	2	3	\$H	TELEARB	1	2	\$H	NACE3D	1	10	2	3	4	5	6	7	8	9
1. EU	\$H	3	0.553	0.250	0.197	\$H	2	0.671	0.329	\$H	10	0.000	0.224	0.026	0.000	0.224	0.053	0.211	0.013	0.184	0.066
1. NC	\$H	3	0.360	0.297	0.342	\$H	2	0.559	0.441	\$H	10	0.000	0.054	0.018	0.018	0.162	0.072	0.279	0.000	0.216	0.180
1. OUTEU	\$H	3	0.505	0.351	0.144	\$H	2	0.629	0.371	\$H	10	0.000	0.093	0.021	0.010	0.237	0.113	0.227	0.000	0.247	0.052
1. PT	\$H	3	0.258	0.548	0.194	\$H	2	0.194	0.806	\$H	10	0.000	0.065	0.032	0.129	0.258	0.065	0.000	0.032	0.194	0.226
2. EU	\$H	3	0.582	0.227	0.191	\$H	2	0.755	0.245	\$H	10	0.000	0.227	0.027	0.000	0.200	0.109	0.245	0.000	0.145	0.045
2. NC	\$H	3	0.403	0.371	0.226	\$H	2	0.556	0.444	\$H	10	0.000	0.137	0.040	0.024	0.129	0.048	0.250	0.024	0.161	0.185
2. OUTEU	\$H	3	0.325	0.423	0.252	\$H	2	0.447	0.553	\$H	10	0.000	0.098	0.033	0.016	0.236	0.114	0.163	0.000	0.220	0.122
2. PT	\$H	3	0.188	0.562	0.250	\$H	2	0.250	0.750	\$H	10	0.000	0.062	0.031	0.219	0.156	0.062	0.062	0.000	0.281	0.125
3. EU	\$H	3	0.430	0.289	0.281	\$H	2	0.659	0.341	\$H	10	0.000	0.496	0.037	0.037	0.104	0.074	0.156	0.000	0.074	0.022
3. NC	\$H	3	0.297	0.338	0.366	\$H	2	0.538	0.462	\$H	10	0.000	0.159	0.048	0.021	0.124	0.076	0.269	0.014	0.117	0.172
3. OUTEU	\$H	3	0.194	0.510	0.296	\$H	2	0.276	0.724	\$H	10	0.000	0.173	0.020	0.082	0.184	0.051	0.143	0.010	0.122	0.214
3. PT	\$H	3	0.028	0.697	0.275	\$H	2	0.073	0.927	\$H	10	0.000	0.101	0.083	0.248	0.193	0.028	0.018	0.000	0.156	0.174
4. EU	\$H	3	0.382	0.324	0.294	\$H	2	0.647	0.353	\$H	10	0.000	0.515	0.000	0.000	0.103	0.029	0.176	0.000	0.088	0.088
4. NC	\$H	3	0.333	0.300	0.367	\$H	2	0.575	0.425	\$H	10	0.000	0.192	0.033	0.025	0.133	0.075	0.242	0.000	0.117	0.183
4. OUTEU	\$H	3	0.143	0.520	0.337	\$H	2	0.245	0.755	\$H	10	0.010	0.122	0.031	0.102	0.133	0.020	0.143	0.000	0.153	0.286
4. PT	\$H	3	0.047	0.581	0.372	\$H	2	0.047	0.953	\$H	10	0.000	0.081	0.035	0.233	0.314	0.000	0.023	0.000	0.186	0.128
5. EU	\$H	3	0.340	0.298	0.362	\$H	2	0.532	0.468	\$H	10	0.000	0.340	0.043	0.000	0.149	0.106	0.106	0.000	0.043	0.213
5. NC	\$H	3	0.106	0.440	0.454	\$H	2	0.333	0.667	\$H	10	0.000	0.113	0.071	0.035	0.156	0.035	0.170	0.021	0.078	0.319
5. OUTEU	\$H	3	0.192	0.449	0.359	\$H	2	0.231	0.769	\$H	10	0.000	0.090	0.090	0.051	0.269	0.051	0.077	0.000	0.103	0.269
5. PT	\$H	3	0.066	0.497	0.437	\$H	2	0.093	0.907	\$H	10	0.013	0.113	0.053	0.192	0.219	0.040	0.033	0.007	0.079	0.252
6. LU	\$H	3	0.084	0.398	0.518	\$H	2	0.237	0.763	\$H	10	0.006	0.040	0.053	0.025	0.128	0.042	0.066	0.005	0.052	0.583

Figure 5.1: Extract of a symbolic data table for the year 2022 produced by the RSDA package

This kind of tables simultaneously with the clustering results allowed making some interesting findings presented in Chapters 6 and 7.

5.2 Algorithm Selection

As previously mentioned, the main objective of the work is to split the population of Luxembourg into homogeneous groups, considering labour market characteristics, and describing their profiles. Given that clustering can be defined as ‘the formal study of methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity’ (Jain, 2010), it made sense to choose cluster analysis to explore the data and achieve the proposed main goal.

As mentioned in Subsubsection 3.1.2, given the current state of development of packages for Symbolic Data in R, there are three clustering algorithms that could be applied in this work. To select the most appropriate one, the use of the validity indices (Silhouette, G2 and G3) was crucial. It is important to note that the worse the validity indices results are, the less separate/less homogeneous the clusters are, and, therefore, the worse is the algorithm.

First, the three clustering algorithms Hierarchical (HCLUST) using Ward’s linkage, SCLUST and DCLUST were applied for all nine years. At this stage of the analysis and until something

else is stated, the χ^2 divergence distance is used. Based on a compromise between the results of the three indices, the optimum number of clusters for the different algorithms was selected. What happened for seven out of the nine years, was that the algorithm that allows obtaining a better compromise between the indices is HCLUST.

Based on Figure 5.2 and Table 5.1a), the choice was to select 4 as the optimum number of clusters and so Silhouette = 0.49, G2 = 0.80 and G3 = 0.04. In Tables 5.1b) and 5.1c), one may see the results for the same indices, for the same year but using different algorithms. In 2022, when using DCLUST and SCLUST, to achieve a compromise between the three indices, they are worst than when using the HCLUST. This means that the HCLUST allows obtaining a greater separation between the different clusters and greater homogeneity within each one. Something similar happened for the majority of the other years except in 2014, where DCLUST produced the best values for the indices and in 2017, where SCLUST was the best.

(a) HCLUST

Number of Clusters \ Validation Index	2	3	4	5	6
Silhouette	0.60	0.49	0.49	0.47	0.44
G2	0.62	0.78	0.80	0.85	0.91
G3	0.15	0.05	0.04	0.02	0.01

(b) DCLUST

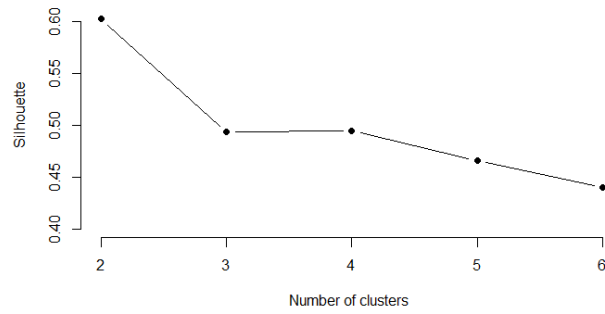
Number of Clusters \ Validation Index	2	3	4	5	6
Silhouette	0.53	0.52	0.27	0.37	0.25
G2	0.60	0.77	0.66	0.64	0.64
G3	0.18	0.05	0.07	0.06	0.07

(c) SCLUST

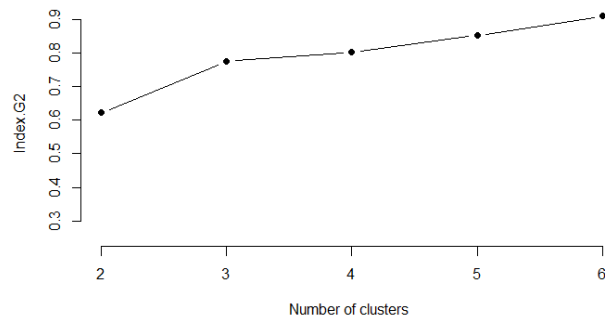
Number of Clusters \ Validation Index	2	3	4	5	6
Silhouette	0.50	0.34	0.38	0.21	0.22
G2	0.60	0.55	0.64	0.56	0.53
G3	0.20	0.15	0.08	0.10	0.10

Table 5.1: Validity Indices, using all the 16 variables - 2022

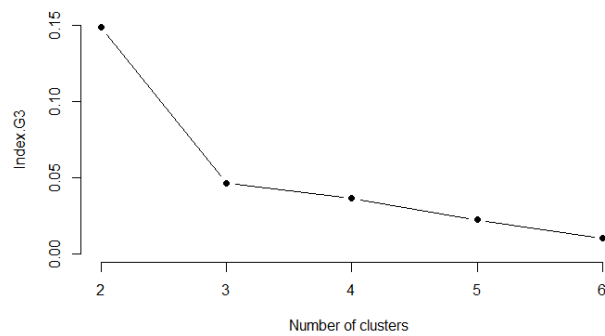
Considering all the mentioned above, HCLUST using Ward's linkage seems to be the best choice among the three. Nevertheless, within the hierarchical algorithms, there are other linkages that can be selected. Therefore, other linkages were analysed to see if there was one better than the linkage used until now. The objective is to select a hierarchical algorithm that is robust to



(a) Silhouette



(b) Index G2



(c) Index G3

Figure 5.2: Validity Indices HCLUST, using all the 16 variables - 2022

weights, does not produce inversions, and has good values in the validity indices. Moreover, an advantage of choosing a hierarchical algorithm instead of a partitional one is that it is not sensitive to initialization conditions, one should generally get the same results.

First of all, the Median and Centroid linkages may produce inversions (Murtagh & Contreras, 2017). Then, according to Ackerman, Ben-David, Brânzei, and Loker (2012), Ward’s method is sensitive to weights (affected by weights on all data sets), and the Average linkage is weight considering (respond to weights on some configurations of the data but not on others). The remaining options included Single linkage - which, even though it is weight robust (ignores weights) (Ackerman et al., 2012) when applied to the LFS data, produces very poor results on the validity indices - Mcquitty and Complete linkage. The last two produced very similar validity indices when applied to the LFS data. Nonetheless, the Complete linkage has an advantage, it has been proven in Ackerman et al. (2012) that it is robust to weights, making it the preferable choice. Additionally, it was also verified that the indices’ results when Complete linkage is used are generally better than when the SCLUST and DCLUST algorithms are used.

Having the algorithm selected, the final step was to see if there was another dissimilarity measure that would produce better results than the one used so far. For that, the year 2018 was randomly picked up and the Hierarchical algorithm using complete linkage was performed combining the different distances available in the R package symbolicDA (J, CHI, REN, CHER and LP) with the different aggregation functions available (type L_1 and L_2). The evaluation of the clustering results was once again made by using the validity indices and the following conclusions were drawn: for all distances, using an aggregation function of type L_1 or L_2 was irrelevant as the results were quite similar. Among all distances, the χ^2 divergence was undoubtedly the distance that produced better results. Consequently, the combination χ^2 divergence distance and an aggregation function of type L_2 was selected. For the purpose of this work, all variables were given equal weight.

5.3 Variable Selection

After selecting the algorithm it was important to select a relevant set of variables to not use variables with minimal cluster information that can lead to biased results. To do that, the following procedure, for each year, was followed:

- Compute the three validation indices to select the optimum number of clusters, k .
- Apply the HINoV algorithm, using the k clusters.
- Select the more meaningful variables based on the TOPRI plots.

After doing this for all nine years, Table 5.2 was created. The table represents the frequency with which each variable appeared in the top variables. In Appendix C, one may find the TOPRI plots for all years.

Variable	Frequency
NACE3D	9
ISCO4D	9
HATLEVEL	9
EARNINGDEC	8
TELEARB	6
DEGURBA	6
SIZEFIRM	4
EARNSATISF	1
SUPERVISOR	1
WISHMORE	1
SEX	0
AGE	0
NUMJOB	0
FIPT	0
LOOKOJ	0
YSTARTWK	0

Table 5.2: Number of years each variable appeared among the most relevant variables using the HINoV algorithm

By analysing the results of Table 5.2, the choice was to proceed with the variables that appear at the top in six or more years. It is quite surprising that DEGURBA was one of the selected variables since it is not directly related to the labour market. However, it is a social-economic variable which can have some kind of relation with a person's position in the labour market. For example, a person who earns more money might be more willing to live in a more expensive area. After deciding which variables were to be used for the analysis, new symbolic data tables were created. These can be found in Appendix C, Figure C.10 to Figure C.18.

To sum up, to get the results presented in Chapter 6, the Hierarchical algorithm with Complete linkage using the χ^2 divergence distance was used, and the model's input variables were DEGURBA, TELEARB, NACE3D, ISCO4D, HATLEVEL, and EARNINGDEC.

Chapter 6

Analysis of Results

This chapter begins with the presentation of the clustering results along with a description of the clusters. Then, considering the cluster results and the distribution of the variables, some general conclusions that portray the immigrant groups' situation in the labour market over the years are drawn. In the end, the evolution of the clusters is monitored.

It is important to mention that the clustering algorithm selected in the previous chapter was applied every year from 2014 to 2022 to a set of 21 symbolic objects that were described by the distributions of six categorical modal variables. These 21 objects correspond to groups in the population and not to individuals.

Furthermore, it is important to emphasise that, as discussed before in Subsubsection 2.3.1, in 2014 the methodology adopted by Statec was a little bit different when compared with the methodology that has been used since 2015. This may have some influence on the results differences observed between 2014 and 2015.

6.1 Cluster Description

After doing all the pre-processing steps described in Section 4.2, aggregating the microdata and selecting the algorithm and the relevant variables, the dendrograms and the validity indices plots were computed for each of the nine years.

Since this analysis is done for all years, to simplify the interpretation and the reading, the description of the clusters for each year is summarised in tables. First, to create these tables, some other tables that are in Appendix - Table D.1 to Table D.9 were created. The tables in the Appendix were built from the cluster descriptions obtained in R. More specifically, each cell represents the average frequencies, that is, the average weight of each category of the variable calculated for objects belonging to the cluster. Then, the frequencies that stood out were highlighted, indicating that there is some relation between a specific cluster and a particular category of a categorical modal variable.

Starting with 2014, the validity indices plots presented in Figure D.1 together with the analysis of the dendrogram in Figure 6.1 suggest a partition into four clusters (Silhouette = 0.45; $G_2 = 0.80$; $G_3 = 0.03$). For the year 2014, the four clusters obtained may be described as presented in Table D.10.

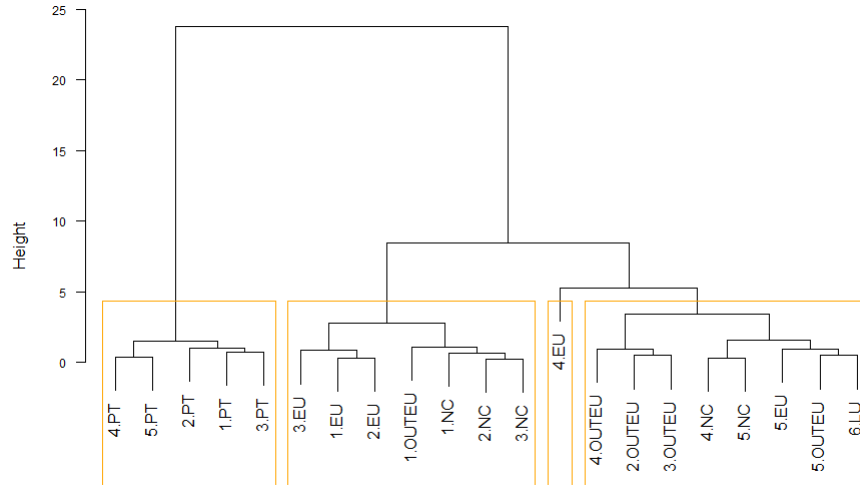


Figure 6.1: Dendrogram 2014

Continuing with 2015, by analysing the plots in Figure D.2, the choice was to select a separation into three clusters since the Index G2 and G3 have already high/low values and the Silhouette value is the second largest (Silhouette = 0.52; G2 = 0.77; G3 = 0.03). Furthermore, the dendrogram in Figure 6.2 seems to also suggest a division into three clusters. Additionally, to make a description of each one of the clusters, Table D.11 was created.

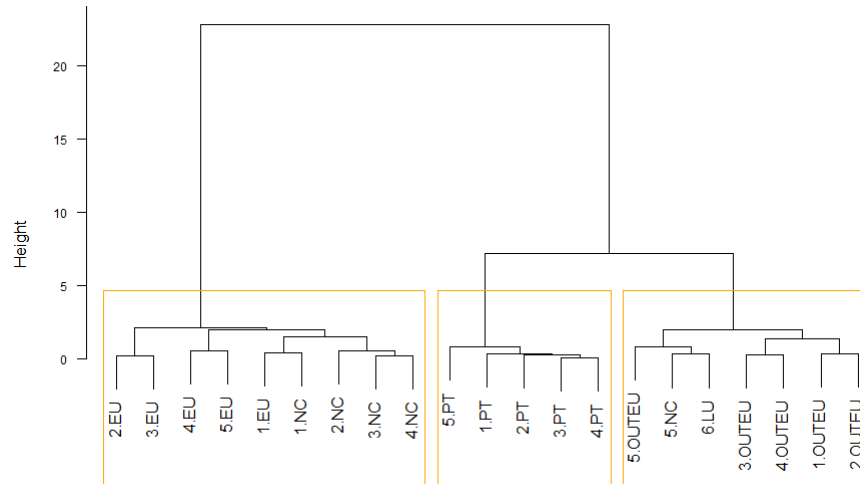


Figure 6.2: Dendrogram 2015

Unlike the previous two years and based on Figure D.3, for 2016, it was quite easy to choose three as the optimal number of clusters (Silhouette = 0.57; G2 = 0.83; G3 = 0.02). The cluster structure for 2016 can be seen in Figure 6.3. A description of the three clusters is made in Table D.12.

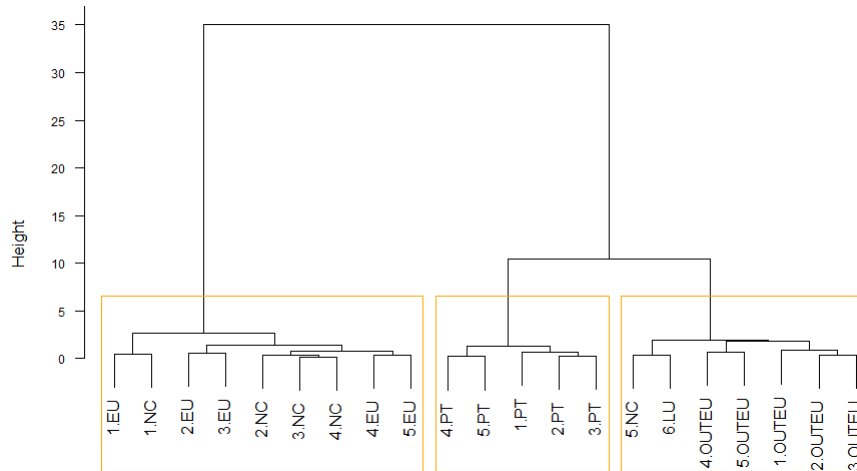


Figure 6.3: Dendrogram 2016

Regarding 2017, by looking at the three indices (see Figure D.4) it is a little bit challenging to select an adequate number of clusters. While a division into four clusters seems reasonable when looking for the indices G2 and G3, the same does not happen when looking for the Silhouette plot. If one only considers the indices and their relevance according to the literature presented in Subsection 3.1.3, perhaps the choice would be a division into four clusters. However, the dendrogram plot in Figure 6.4 suggests a division into three clusters. Comparing the indices values for the division into three clusters (Silhouette = 0.61; G2 = 0.84; G3 = 0.02) with the values for the other years (and not their variation in 2017) together with the dendrogram information, the choice was to divide the objects into three groups. To describe them, Table D.13 was created.

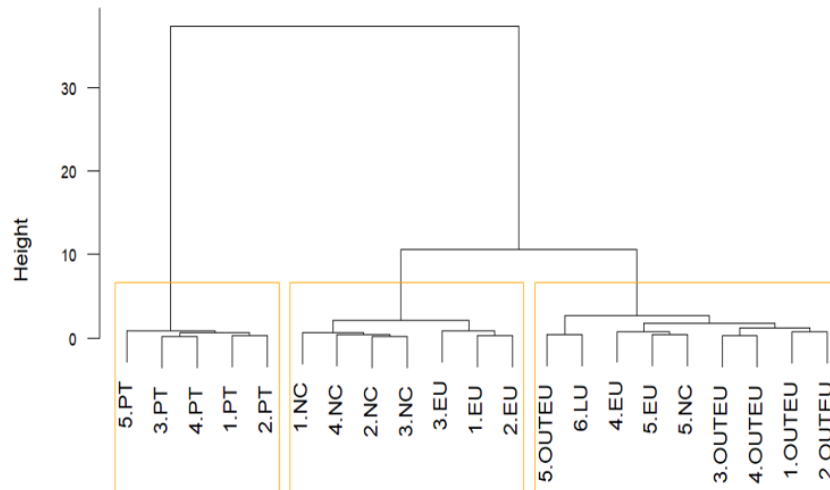


Figure 6.4: Dendrogram 2017

For 2018, the decision was to divide the objects into three clusters (Silhouette = 0.58; G2 =

0.82; $G3 = 0.04$). This choice was not straightforward since the index $G2$ is the second lowest and the index $G3$ is the second highest. However, a big improvement in the values of these two indices from the division into two clusters for the division into three clusters is visible, while there is not a huge improvement if the division is in more clusters. Furthermore, the Silhouette value reaches its peak in the division into three clusters (see Figure D.5). Moreover, the dendrogram in Figure 6.5 also suggests a division into three clusters. Finally, one may see in the dendrogram that the Portuguese immigrant groups are no longer in a cluster alone as they have been until now. A description of the four groups is presented in Table D.14.

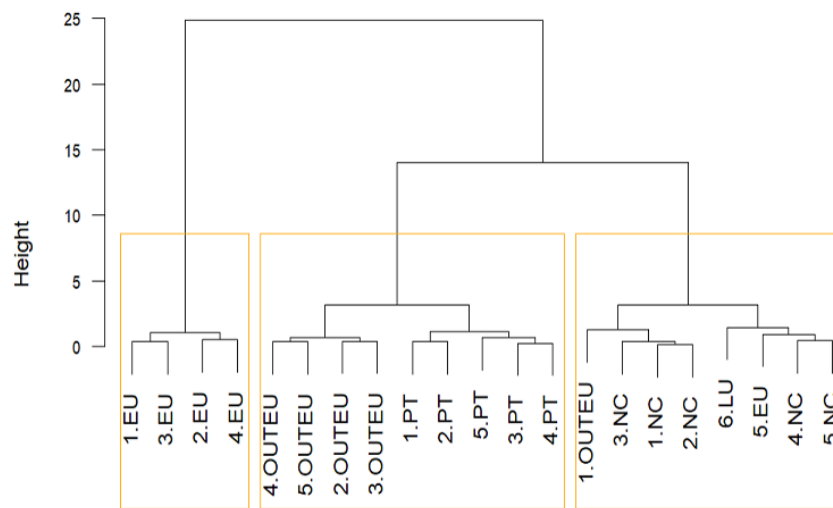


Figure 6.5: Dendrogram 2018

Based on the validity indices for 2019 (see Figure D.6), a split into four clusters was selected (Silhouette = 0.56; $G2 = 0.82$; $G3 = 0.03$). The cluster structure, which also appears to indicate a separation into four clusters, is given by the dendrogram in Figure 6.6. Additionally, the interpretation of each one of the clusters is provided by Table D.15.

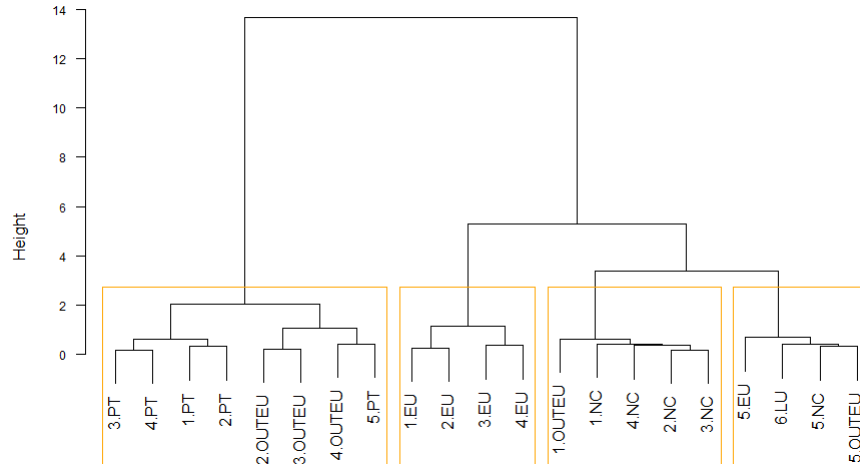


Figure 6.6: Dendrogram 2019

For the year 2020, based on the plots of the indices (Figure D.7) it was a bit difficult to make a decision since when the Index G2 and G3 are at their best, the Silhouette has its worst value. However, the choice was to proceed with four clusters since the indices values as a whole for four clusters are similar to the values of previous years (Silhouette = 0.47; G2 = 0.83; G3 = 0.03). Furthermore, looking at the dendrogram in Figure 6.7, it seems reasonable to do a division into four clusters. Lastly, in Table D.16 a characterization of the clusters is made.

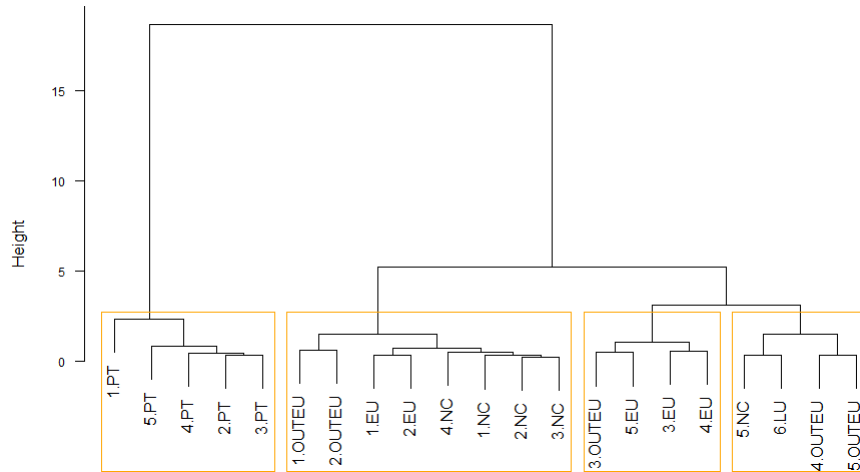


Figure 6.7: Dendrogram 2020

According to Figure D.8, for 2021 the choice was to split the symbolic objects into six clusters (Silhouette = 0.45; G2 = 0.86; G3 = 0.01). The objects were grouped as shown in the dendrogram in Figure 6.8 and a characterization of each one of the six groups can be found in Table D.17.

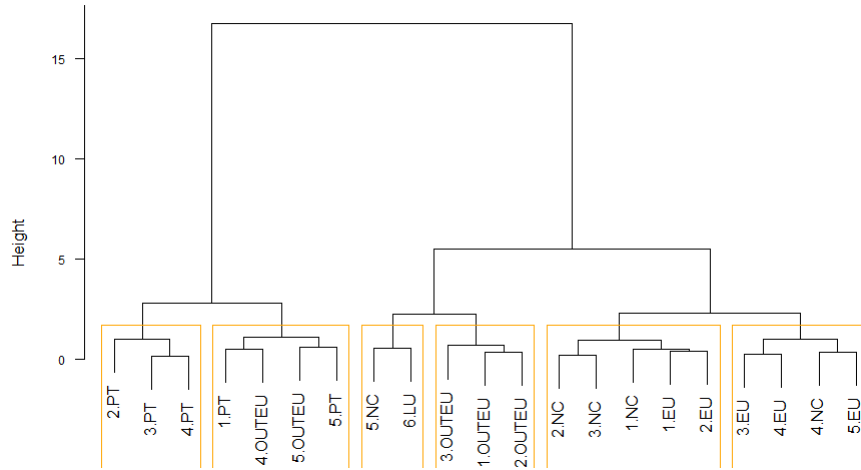


Figure 6.8: Dendrogram 2021

Lastly, concerning 2022, the decision was to select five clusters (Silhouette = 0.48; $G2 = 0.91$; $G3 = 0.01$). Even though the Silhouette value is one of the lowest for the division into five clusters, it is similar to the value in some of the other years. Moreover, by selecting five clusters the other two indices show very good results. Furthermore, by analysing the dendrogram in Figure 6.9 and the values for the indices, one may see that a division into four clusters would not be a bad choice as well. However, the choice was to proceed with five clusters due to the very good results in the indices $G2$ and $G3$. The description of the cluster can be seen in Table D.18.

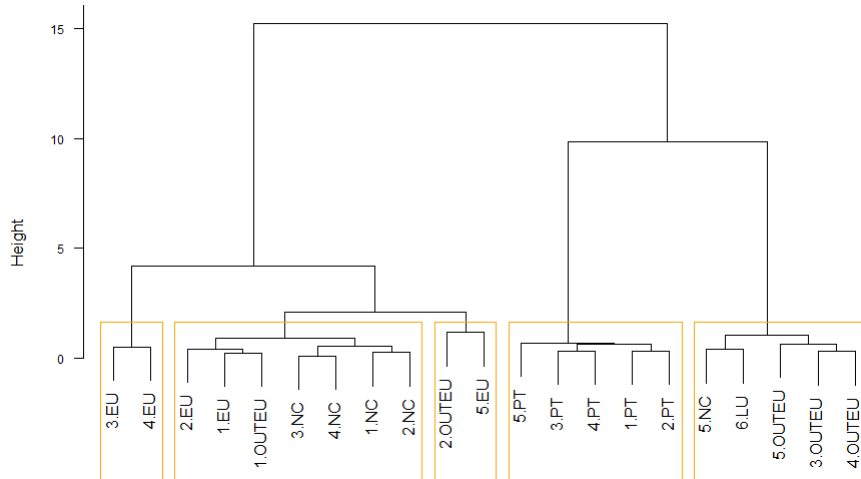


Figure 6.9: Dendrogram 2022

By analysing the clustering results, the clusters' descriptions and the distribution of the symbolic variables for each object many conclusions may be drawn.

Firstly, by analyzing the maximum heights of each of the dendrograms (23.78, 22.84, 35.07, 37.35, 24.92, 13.62, 18.67, 16.77, 15.23, from 2014 to 2022 respectively) it is possible to say

that, in general, the objects have been getting closer. That is, in recent years (2019-2022) the maximum heights are lower than the maximum heights from 2014 to 2018. This could mean that the disparities between groups are decreasing.

One interesting thing is that for six out of the nine years, all the Portuguese groups together form a cluster, a phenomenon that does not happen with the other places of birth. Consequently, from the clusters' descriptions, it is possible to do almost a straightforward characterization of the Portuguese people's profile in the labour market in Luxembourg.

Starting with the sector of activity. The EU symbolic objects are almost always in clusters where the average percentage of people working in Arts, entertainment and recreation and other service activities is significant, suggesting that it is very common for people who were born in an EU country and moved to Luxembourg to work in this sector. Furthermore, if a closer look is taken at the distribution of the variable NACE3D for each EU group (see Figure C.10 to Figure C.18), it is possible to conclude that working in Financial and insurance services is also very common for the EU groups. Regarding the Portuguese community, the results suggest that the Construction and Trade, transport and hospitality sectors represent a significant percentage of the activities of Portuguese workers. Furthermore, the distribution of the variable NACE3D and the cluster results show that many people from non-EU countries also work in Trade, transport and hospitality. People from the NC are always in clusters where Financial and insurance activities play a major role. Concerning the natives, the outcomes indicate that they mainly work in Public administration, defence, education, health and social work which is in line with what was mentioned in Subsection 2.2.2, but this time using the LFS data. Lastly, by analysing only the NACE3D variable distribution per object (see Figure C.10 to Figure C.18), one may conclude that the Business service sector is usually widespread among recently arrived people (YEARESID=1), regardless of their place of birth. By doing a similar analysis, it was possible to figure out that Public administration, defence, education, health and social work activities are not only common among the natives but also common among the longer established people (YEARESID=5), regardless of the place of birth. The fact that only people who have lived in the country the longest may work in Public administration, defence, education, health and social work could be due to the fact that to work in these activities there are language requirements that only with time people will have.

Regarding income, Portuguese objects always belong to clusters characterized by low income. On the other hand, the clusters in which EU and NC groups belong are usually characterized by upper-income earners. Sometimes for the recent groups, it may be a little bit less, perhaps because people may be younger and with less work experience. As for the natives and the groups of people from countries outside the EU it is a little bit more challenging to understand. This may be explained by the fact that the LU group includes a mix of people with different years of experience in the labour market and by the fact that the OUTEU groups include people from many different countries with different development states.

With regard to the education level, it is possible to conclude that people from the NC and EU are highly educated, belonging to clusters where the average percentage of people having a bachelor's or a master's/PhD predominates. On the other hand, Portuguese objects always belong to clusters where the highest level of education that predominates is high school. By analyzing the variable HATLEVEL per object it is also possible to make the following conclusions:

- Portuguese have lower levels of education; nevertheless, the profile of the immigrants who come to Luxembourg is changing a bit. Even though the percentage of people who have a low education level is still high, it is possible to see an increase in the percentage of recently arrived immigrants (1.PT) who have a Master's or a PhD, see Figure 6.10.

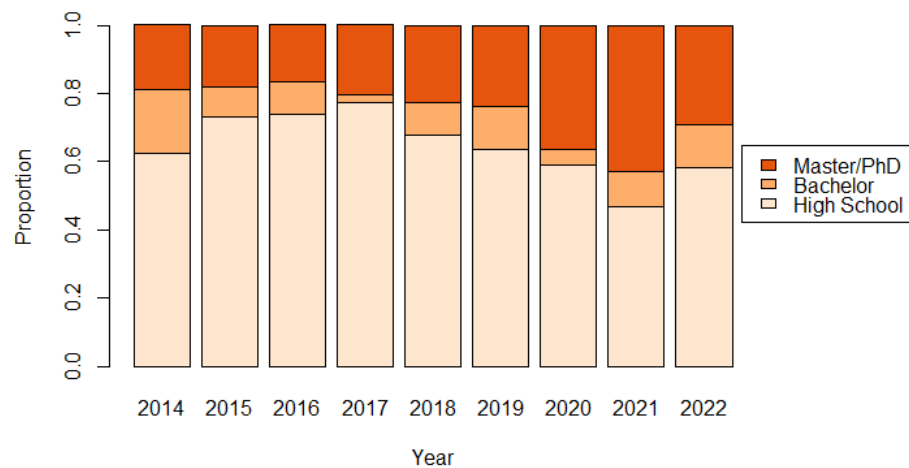


Figure 6.10: Education attainment of recently arrived Portuguese immigrants (1.PT) 2014-2022

- The majority of groups of people from the EU and the NC have always stood out for their percentage of people holding a Master's or PhD. However, this percentage is increasingly higher. Through the description of the clusters of each year, it is possible to observe that initially the clusters where these objects were inserted were characterized only by higher education (bachelor and over) and, recently, there are already some clusters that are characterized by the achievement of a Master's or PhD.
- Among the natives, the percentage of people with education up to secondary school has been declining in recent years. As it is possible to see in Figure 6.11, it has varied between about 55% and a little bit above 65% over the last 9 years. The natives are not as educated as the NC and EU people, but more educated than the Portuguese overall.

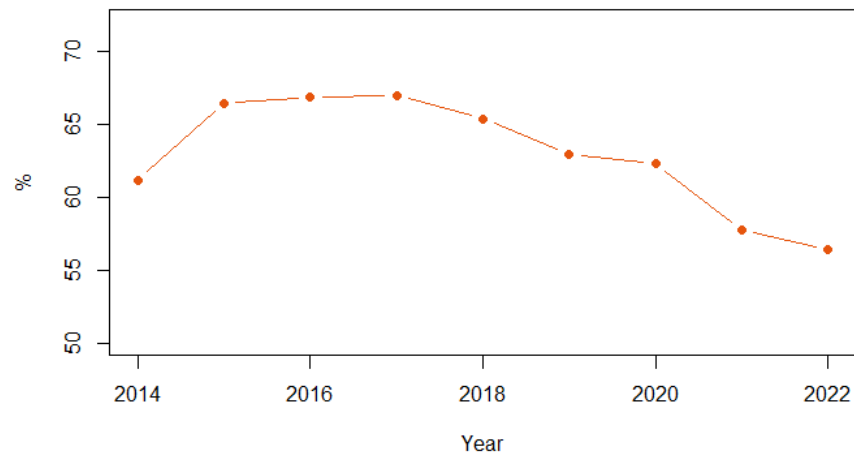


Figure 6.11: Percentage of Luxembourgers whose highest level of education attained is secondary education 2014-2022

- Regarding the OUTEU groups, what happens more than one year and in more than one object is that the two extremes have similar proportions, that is similar percentages of people having only high school to the percentages of people having a Master/PhD, as an example see Figure 6.12.

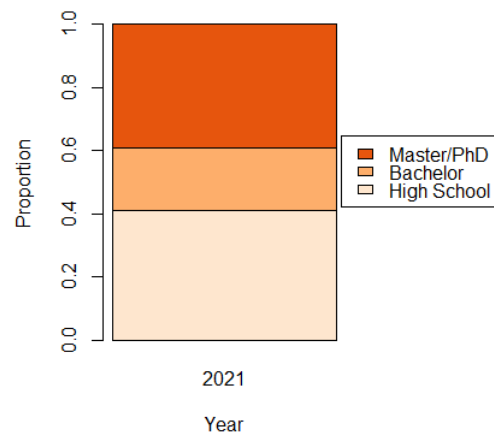


Figure 6.12: Education attainment level of people born in a non-EU country and living in Luxembourg for between 9 and 15 years (3.OUTEU) - 2021.

Some of these findings regarding the HATLEVEL, for the period 2014-2022, are in line with

what was explored in Ametepe et al. (2022) (this article was discussed in Subsection 2.2.2) for the year 2021.

By looking at the clusters through time, one can observe that the Portuguese groups are always in clusters where there is a large percentage of people working in blue-collar professions, whereas the NC and EU objects belong to clusters where white-collar high-skilled professions are predominant. Object 6.LU almost always belong to clusters characterized by white-collar professions (high and low-skilled). Concerning the people from OUTEU, once again it is a little bit difficult to understand what characterizes these groups. Additionally, by looking at ISCO4D variable distribution through time it is possible to conclude that after a strong decline from 2014 to 2016, things changed a bit for the recently arrived Portuguese group (1.PT), meaning that the percentage of people working in white-collar high-skilled professions increased, reaching its peak in 2021, see Figure 6.13.

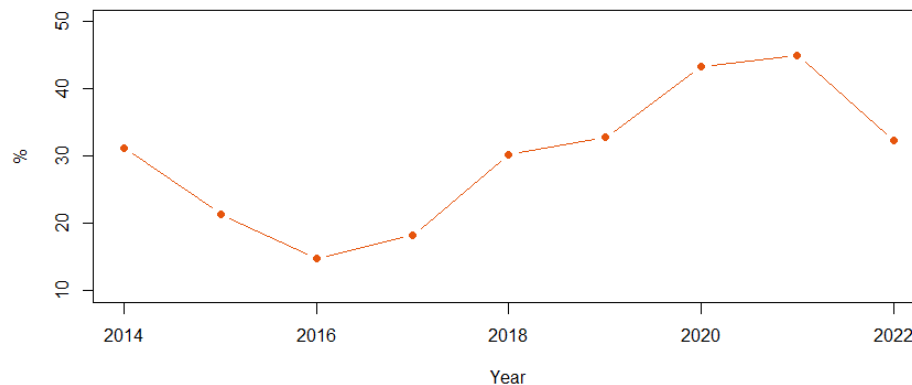


Figure 6.13: Percentage of recently arrived Portuguese immigrants (1.PT) working in white-collar high-skilled professions 2014-2022

By analysing Figure C.10 to Figure C.18 it is possible to conclude that it is very common for the natives to live in Rural areas. When looking at the clusters, one may see that it is very common for the NC and EU objects to fall in clusters where living in Cities is very common and that the Portuguese groups are more associated with living in Towns and Suburbs. Furthermore, the people from EU and from the NC start by living in Cities and as time passes they start moving to areas with other degrees of urbanization (people from the NC countries do it sooner). For the year 2018, Figure 6.14 illustrates well that fact.

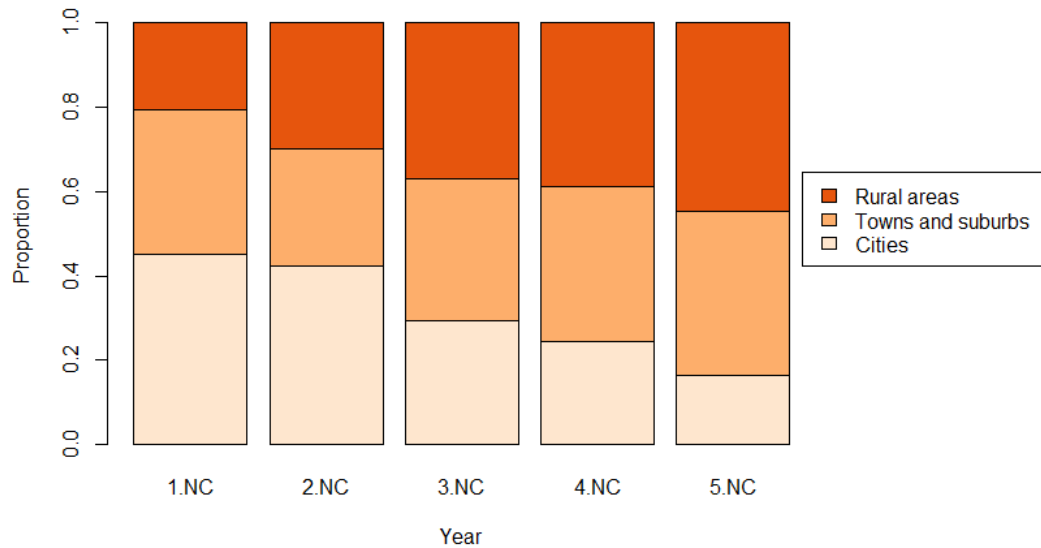


Figure 6.14: Percentage of people living in each of the different types of urbanisation for the different NC groups - 2018

For the people born in a non-EU country, once again it is difficult to find a pattern, however, since 2017 it is possible to verify that the recently arrived (1.0UTEU) mainly live in cities, however as time passes, they start moving to Towns and Suburbs. In fact, regardless of the place of birth, the percentage of individuals living in cities declines as YEARESID rises. This might be explained by the fact that as time passes, families may grow which leads to an increase in expenses and people begin to relocate to less expensive areas. To demonstrate this idea the plot in Figure 6.15 was made. Looking at the cities, the darker dots are above the respective lighter dots, showing that the percentage of people in each group living in cities is higher for those who have recently arrived to Luxembourg compared to those who are more established. Conversely, looking at the rural areas, one can see the exact opposite. The percentage of people of each group living in rural areas is higher in the more established groups in comparison to the recently arrived ones.

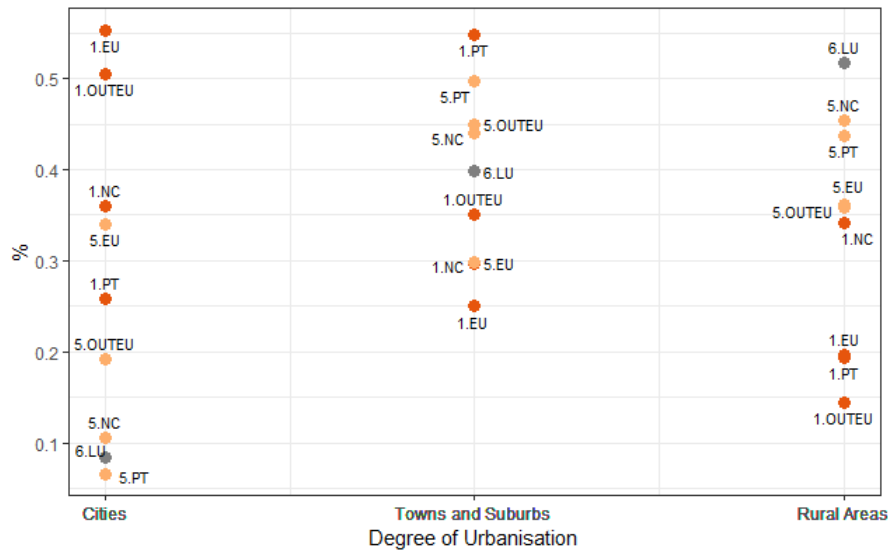


Figure 6.15: Degree of Urbanization across the among the recently arrived and longer established immigrants - 2022

Lastly, as it is known many things have changed due to the pandemic and being able to work remotely is one of those things. Observing the tables that describe the clusters for each year it is possible to verify that since 2020 the ability to do telework is an important characteristic to distinguish the groups. Furthermore, since 2020, the TELEARB variable has always been in the top 3 of the most important variables to form the clusters (see Figures Figure D.10 to Figure D.12), while in the past it did not have such big influence, being in the majority of the previous years the least important variable. Considering this, a question arises: ‘Is telework here to stay, or was it a temporary trend and things are already getting back to what was common in 2019?’. To further explore this topic and to understand why in some groups there was a huge increase and in others not, statistical tests were performed and plots were computed.

Bar plots for the variable TELEARB from 2019 to 2022 were computed (Figures 6.16 to 6.20). By observing them, it is possible to confirm that some groups, such as all EU and NC groups, experienced large increases in the proportion of workers who started doing telework in 2020 and that this percentage remained high until 2022. On the other hand, there are other groups where the increase does not seem very meaningful, for example among the Portuguese groups. In the years after 2020, in some groups, the percentages had little increases/decreases, but in others, there were big decreases, and in 2022, the percentages reached levels similar to 2019.

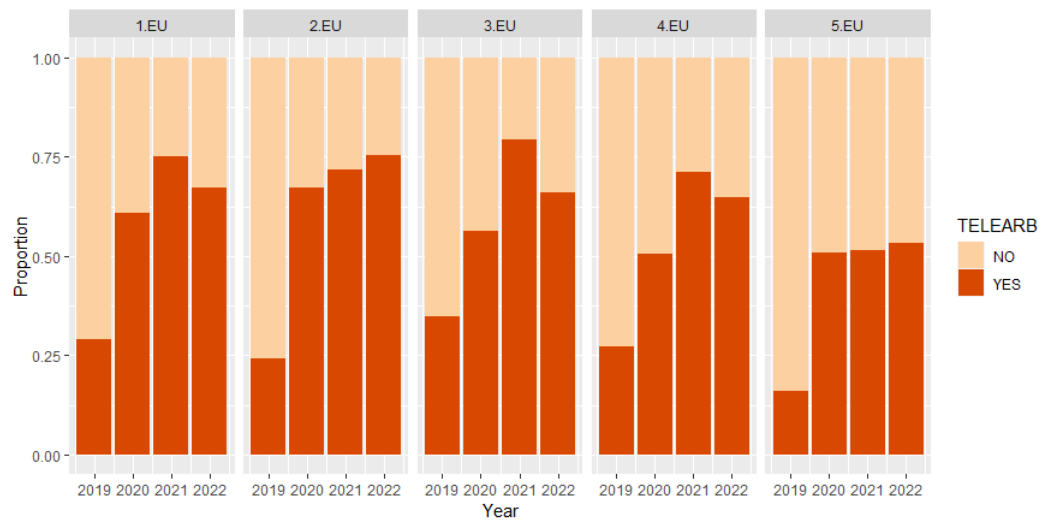


Figure 6.16: Telework variation European Union community 2019 - 2022

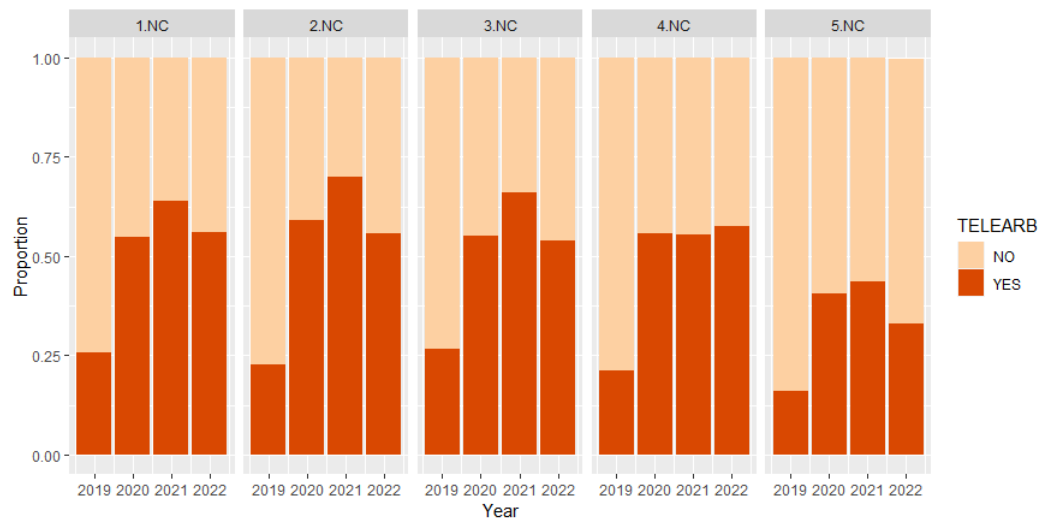


Figure 6.17: Telework variation Neighbouring countries community 2019 - 2022

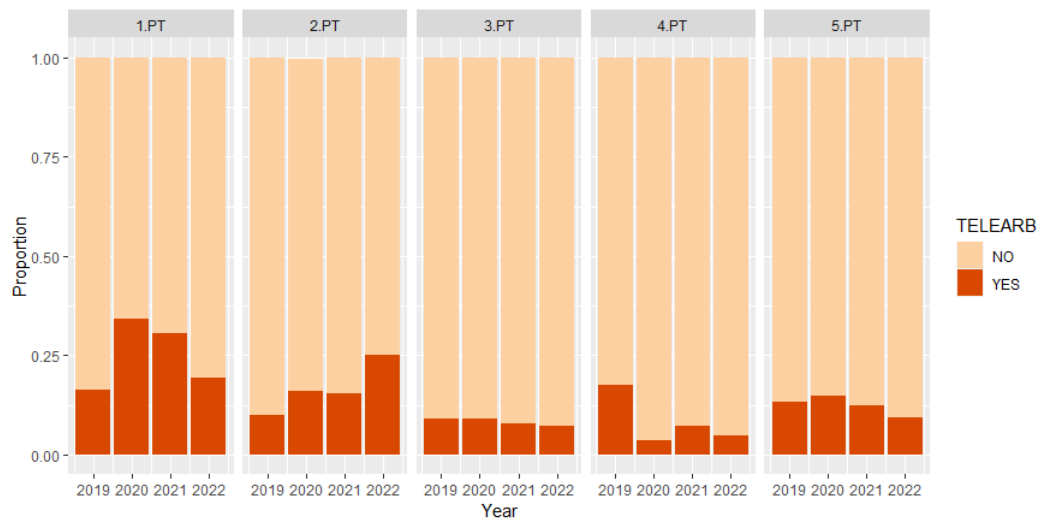


Figure 6.18: Telework variation Portuguese community 2019 - 2022

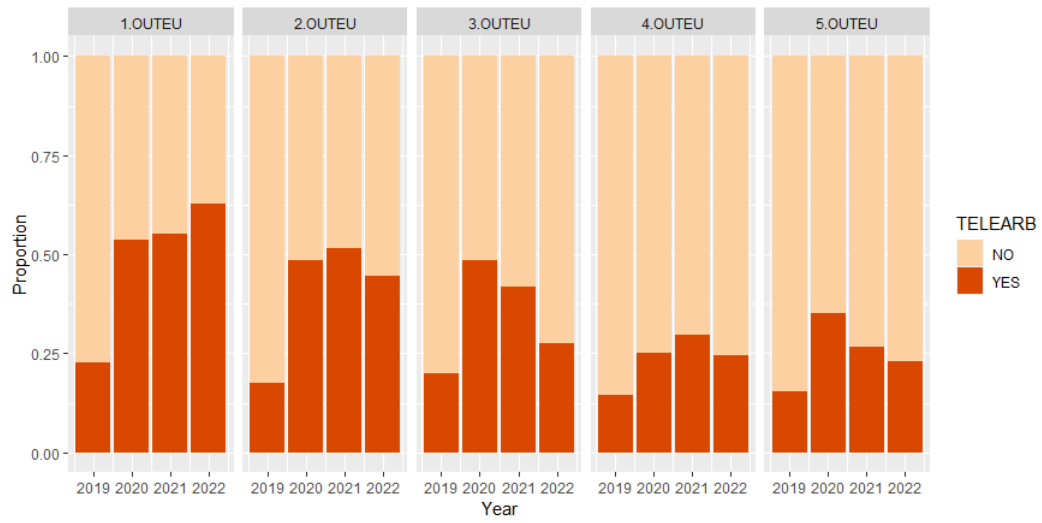


Figure 6.19: Telework variation Outside the European Union community 2019 - 2022

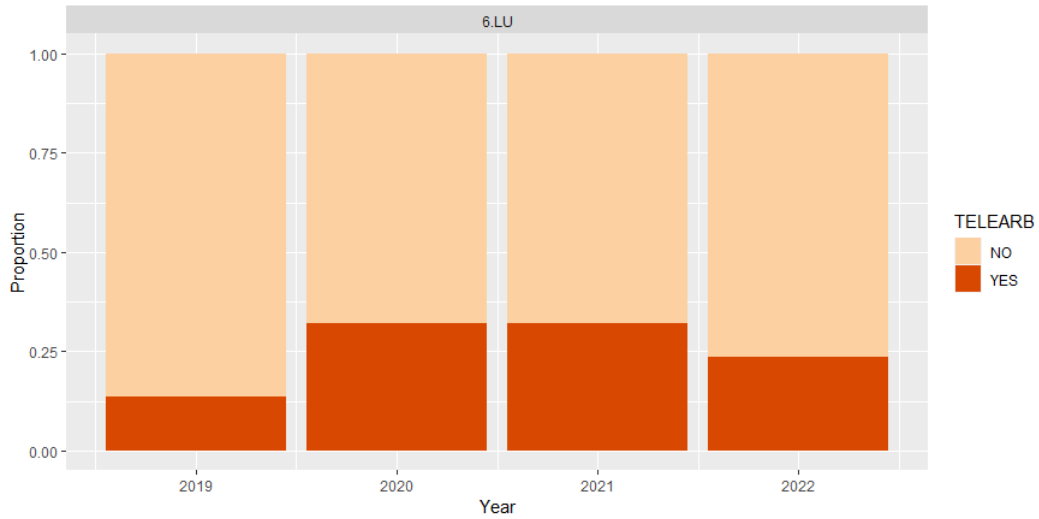


Figure 6.20: Telework variation Luxembourgish natives 2019 - 2022

To see if the proportions are significantly different from 2019 to 2020, two-tailed proportion tests (5% significance level) were applied with the following hypothesis:

- H_0 : The proportions in 2019 and 2020 are equal
- H_1 : The proportions in 2019 and 2020 are different

To perform the tests there was a need to recover microdata information. More concretely, the number of people doing telework in each group was recovered by multiplying the proportion of people doing telework in each symbolic object by the total number of classical observations in each symbolic object.

The tests results (see Appendix D, Table D.19), allow to conclude that telework has significantly increased for all the EU, LU, NC and OUTEU (except 4.OUTEU) groups. However, that is not true for the Portuguese groups. This may be explained by the fact that, as previously mentioned, blue-collar professions, which generally refer to jobs that require physical labour, are the most prevalent among the Portuguese, making the implementation of telework challenging.

A similar procedure was also applied to see if the decreases from 2020 to 2022 were significant which can help understanding if telework came to stay or if it is decreasing as life gets back to 'normality'. In this case, the results (see Appendix D, Table D.20) suggest that for all EU and NC groups telework is something to stay (it was not possible to conclude that the proportions are significantly different) and for the natives it was just a temporary situation (the proportions were significantly different). Regarding the OUTEU groups, the results differ across groups.

6.2 Cluster Monitoring

The graph in Figure 6.21 was obtained by applying the MEC framework to monitor the cluster transitions. To make it easier to follow the next analysis, Table D.21 should be used so that it

is possible to see which objects are in each cluster each year.

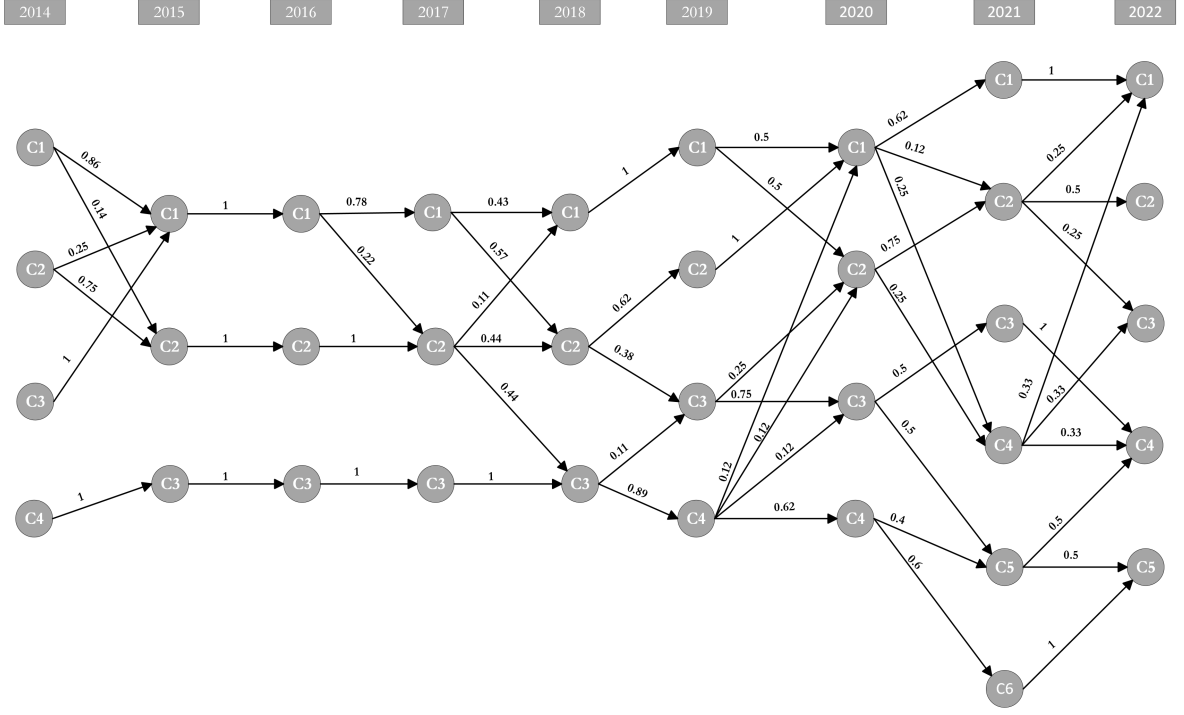


Figure 6.21: Cluster Monitoring Graph 2014-2022

Applying the MEC nomenclature, the transitions that follow have been identified. As previously mentioned in Subsection 3.2, to analyse the compactness transitions, the heights of the different classes in the dendrogram were carefully studied using some R functions. Furthermore, one should remember the notations presented in Appendix A.

2014-2015

- Two survivals: $C_2(2014) \rightarrow C_2(2015)$. $C_4(2014) \rightarrow C_3(2015)$.
Within the survivals, it is possible to identify a shrinkage and two compressions: $C_2(2014) \searrow C_2(2015)$, $C_2(2014) \xrightarrow{\bullet} C_2(2015)$, $C_4(2014) \xrightarrow{\bullet} C_3(2015)$.
- One merge: $\{C_1(2014), C_3(2014)\} \xrightarrow{\subset} C_1(2015)$.

2015-2016:

- All clusters survived: $C_1(2015) \rightarrow C_1(2016)$, $C_2(2015) \rightarrow C_2(2016)$, $C_3(2015) \rightarrow C_3(2016)$.
Within the survivals, one may distinguish three endogenous compactness transitions, more concretely, two dispersals and one compression: $C_1(2015) \xrightarrow{*} C_1(2016)$, $C_3(2015) \xrightarrow{*} C_3(2016)$, $C_2(2015) \xrightarrow{\bullet} C_2(2016)$, respectively.

2016-2017

- All clusters survived: $C_1(2016) \rightarrow C_1(2017)$, $C_2(2016) \rightarrow C_2(2017)$, $C_3(2016) \rightarrow C_3(2017)$.
Two endogenous size transitions may be identified: $C_1(2016) \searrow C_1(2017)$, $C_2(2016) \nearrow C_2(2017)$.
Moreover, three endogenous compactness transitions may also be identified: $C_1(2016) \xrightarrow{\bullet} C_1(2017)$, $C_2(2016) \xrightarrow{*} C_2(2017)$, $C_3(2016) \xrightarrow{\bullet} C_3(2017)$.

2017-2018

- Two survivals: $C_1(2017) \rightarrow C_2(2018)$, $C_3(2017) \rightarrow C_3(2018)$.
Within the survivals, the following endogenous transitions can be identified: $C_1(2017) \nearrow C_2(2018)$, $C_3(2017) \nearrow C_3(2018)$, $C_1(2017) \xrightarrow{*} C_2(2018)$, $C_3(2017) \xrightarrow{*} C_3(2018)$.
- One birth: $\emptyset \rightarrow C_1(2018)$.
- One split: $C_2(2017) \xrightarrow{\subseteq} \{C_2(2018), C_3(2018)\}$.

2018-2019

- Three survivals: $C_1(2018) \rightarrow C_1(2019)$, $C_2(2018) \rightarrow C_2(2019)$, $C_3(2018) \rightarrow C_4(2019)$.
Two endogenous size transitions may be identified: $C_2(2018) \searrow C_2(2019)$, $C_3(2018) \searrow C_4(2019)$.
Moreover, three endogenous compactness transitions may also be identified: $C_1(2018) \xrightarrow{*} C_1(2019)$, $C_2(2018) \xrightarrow{\bullet} C_2(2019)$, $C_3(2018) \xrightarrow{\bullet} C_4(2019)$.
- One birth: $\emptyset \rightarrow C_3(2019)$

2019-2020

- Two survivals: $C_3(2019) \rightarrow C_3(2020)$, $C_4(2019) \rightarrow C_4(2020)$.
The following endogenous size transition may be identified: $C_4(2019) \searrow C_4(2020)$.
Furthermore, two endogenous compactness transitions can be identified: $C_3(2019) \xrightarrow{*} C_3(2020)$, $C_4(2019) \xrightarrow{*} C_4(2020)$.
- One merge: $\{C_1(2019), C_2(2019)\} \xrightarrow{\subseteq} C_1(2020)$.

2020-2021

- Three survivals: $C_1(2020) \rightarrow C_1(2021)$, $C_2(2020) \rightarrow C_2(2021)$, $C_4(2020) \rightarrow C_6(2021)$.
It is possible to identify two size transitions: $C_1(2020) \searrow C_1(2021)$ and $C_4(2020) \searrow C_6(2021)$.
Additionally, three compactness transitions may also be identified, more concretely, three compressions: $C_1(2020) \xrightarrow{\bullet} C_1(2021)$, $C_2(2020) \xrightarrow{\bullet} C_2(2021)$, $C_4(2020) \xrightarrow{\bullet} C_6(2021)$.

2021-2022

- Two survivals: $C_1(2021) \rightarrow C_1(2022)$, $C_2(2021) \rightarrow C_2(2022)$.
Within these survivals, endogenous size and compactness transitions can be identified:
 $C_1(2021) \nearrow C_1(2022)$, $C_2(2021) \searrow C_2(2022)$, $C_1(2021) \xrightarrow{\bullet} C_1(2022)$, $C_2(2021) \xrightarrow{\bullet} C_2(2022)$.
- One birth: $\emptyset \rightarrow C_3(2022)$.
- One split: $C_4(2021) \xrightarrow{\subseteq} \{C_1(2022), C_3(2022), C_4(2022)\}$.
- Two merges: $\{C_3(2021), C_5(2021)\} \xrightarrow{\subseteq} C_4(2022)$, $\{C_5(2021), C_6(2021)\} \xrightarrow{\subseteq} C_5(2022)$.

Starting by analysing some of the transitions identified above, one may see that from 2015 to 2016 all the clusters survived, there were no size transitions and the composition of the clusters remained exactly the same (see Table D.21). Nevertheless, some of the groups became more/less dispersed. The height of the Clusters C_1 and C_2 increased (from 2.12 to 2.62 and from 0.82 to 1.28, respectively). This means that the objects are more dispersed in 2016. On the other hand, the height of Cluster C_2 decreased very slightly (1.97 to 1.94), meaning that the objects are a bit less dispersed. A similar thing happened with Cluster $C_1(2018)$ and $C_1(2019)$. The objects in the cluster remained exactly the same from one year to the other but the objects are slightly more dispersed in 2019.

Furthermore, by looking at the clusters' height it is possible to see that, even though the Portuguese objects are almost always together in a cluster (except for 2021), the height of the class to which they belong to varies over time. This means that there are years in which the profile of the different groups of Portuguese people is more similar than in others. Part of this may be seen through the endogenous compactness transitions. More concretely, for each year, the height of the surviving clusters to which the Portuguese objects belong was as follows: 1.52, 0.82, 1.28, 0.81, 3.19, 2.04, 2.36, 1.00. The previous values are from 2014 to 2021 respectively, years between which there were clusters of Portuguese objects that survived. Furthermore, the height of the Portuguese cluster in 2022 was 0.70. This last cluster was the result of a merge and not a survival. By analysing the heights, one may see that from 2017 to 2018 there was a big increase, from 0.81 to 3.19. This is due to the fact that in 2018 the cluster was composed not only of Portuguese objects, as it happened in the previous years, but also of objects referring to people from non-Eu countries. A similar thing happened in 2019, which may explain why the height is still high compared to other years. Nevertheless, in 2020, the height of the Portuguese cluster (composed only of the five Portuguese objects) is still above two and it has even increased compared to the previous year (2019). This is probably explained by the pandemic adaptations in the labour market, for example, the introduction of telework. This might have led the object 1.PT to become more distant from the other objects. As was already seen in Figure 6.13, 2020 was the year where there was the second highest percentage of 1.PT working in white-collar high-skilled professions, jobs where telework might be easier.

Observing Figure 6.21, one may verify that in some years there are many movements from one cluster to another, especially from 2019 to 2022. The high number of movements can be explained by events that occurred in the world or changes in methodology.

From 2019 to 2020, lots of movements occurred, which may be explained by the beginning of the Pandemic which led to labour market adaptations to this new situation as well as by the Brexit. In 2020 the most essential variables for cluster formation were HATLEVEL, TELEARB and ISCO4D (Figure D.10). Overall, from 2019 to 2020 there were major changes in EU and OUTEU groups due to Brexit. The United Kingdom people, who are usually highly educated and work mainly in white-collar high-skilled professions moved from the EU groups to the OUTEU ones, representing the largest OUTEU community. This contributed to an increase in ISCO4D=1 (white-collar high-skilled professions) and HATLEVEL=3 (master/Phd) and 2 (bachelor) in the OUTEU objects, causing them to separate from the Portuguese groups. Furthermore, there was a merge of two clusters that, as previously mentioned, did not differ much, and the TELEARB variable came to accentuate the equalities between these objects. Nevertheless 3.EU and 4.EU were not present in the merge. These objects saw decreases in the percentages of highly educated people and in the percentage of people working in white-collar high-skilled professions, which might be explained by the fact that people from the UK are no longer included in these objects.

From 2020 to 2021 there were some methodological changes which may have contributed to some of the movements. Moreover, even though the transitions were all survivals, not all the objects remained in the same cluster. What may have given rise to all the movements from one year to the other was the change in the importance of the variables for the formation of the clusters, more concretely, the fact that ISCO4D is the most important variable alone (see Figure D.11) and not together with HATLEVEL and TELEARB as in 2020 (Figure D.10). Focusing on the clusters that survived, it was possible to conclude that for all the elements that did not move to the corresponding cluster, the distribution of the ISCO4D variable was already a little different from the distribution observed in the other elements of the same cluster; only this time the ISCO4D variable is the most important for the formation of clusters. Even though one cannot say that some external transition occurred in cluster $C_3(2020)$, the truth is that the objects moved in pairs. Already in 2020, within cluster C_3 , it was possible to see that 4.OUT and 5.OUT were more similar to each other just as 5.NC and 6.LU in terms of the distributions of the variable ISCO4D, see Figure 6.22; however, other variables played also an important role making them belong to the same cluster. Nevertheless, in 2021, ISCO4D was the most important variable among all and may explain the movements.

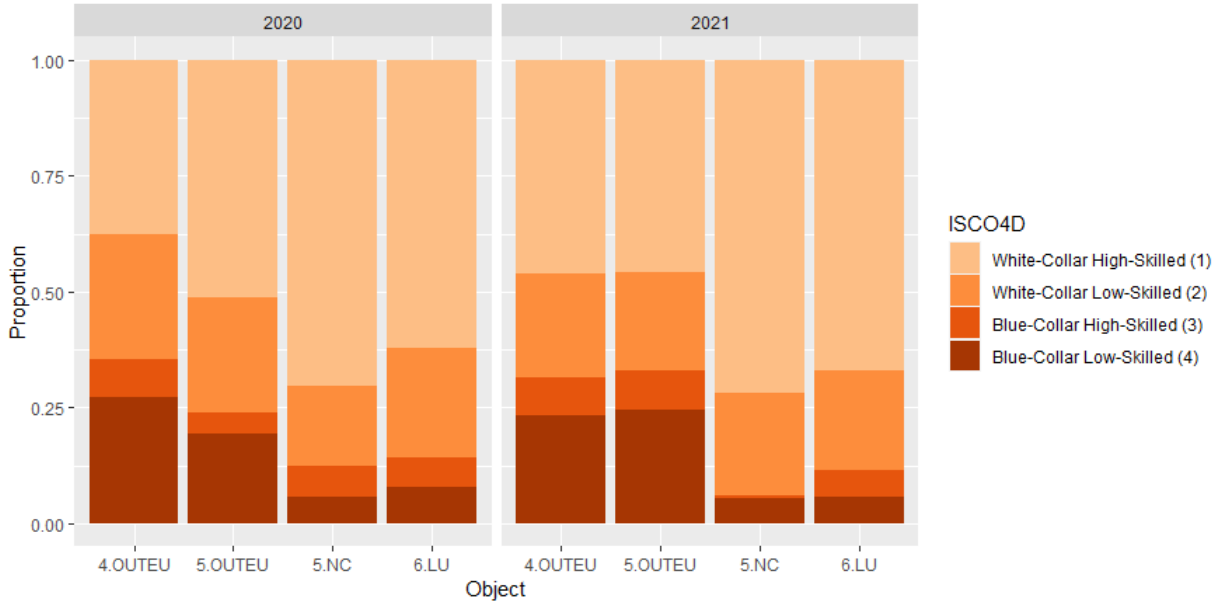


Figure 6.22: ISCO4D across $C_3(2020)$ objects 2020-2021

Similar to what happened from 2020 to 2021, the variable's importance for cluster formation may explain some of the movements from 2021 to 2022. In 2022 there are three variables that are more or less equally important for the cluster structure, TELEARB, ISCO4D and DEGURBA (see Figure D.12). It is likely that teleworking has a high importance in the formation of clusters because, as already seen, in 2022, with the pandemic already stabilized, teleworking became something more optional. For that reason, in the activities/professions where it is not so convenient, it gradually began to return to normality. Consequently, in some objects, a decrease in the percentage of people teleworking is visible (see Figures 6.16 to 6.20). This accentuated differences in the distribution of the TELEARB variable between some objects in the same cluster from one year to the other. In fact, the two cluster division of the dendrogram in Figure 6.9, roughly separates the objects where telework remained a common practice of those where it did not. There are two reasons that may have caused most of the movements from 2021 to 2022. First, in 2021 there were already differences between objects in the same cluster in the variables TELEARB, ISCO4D and DEGURBA. In 2022, these variables are together the most important ones making the previously mentioned differences the cause of the movements of the objects to clusters where the set of distributions of the three variables was closer. Secondly, some differences that already existed have been further accentuated, such as in the cluster $C_4(2021)$ the differences in TELEARB were already visible, nevertheless, there was an increase in TELEARB=YES in 1.OUT and a decrease in the other two objects (2.OUT and 3.OUT), see Figure 6.23.

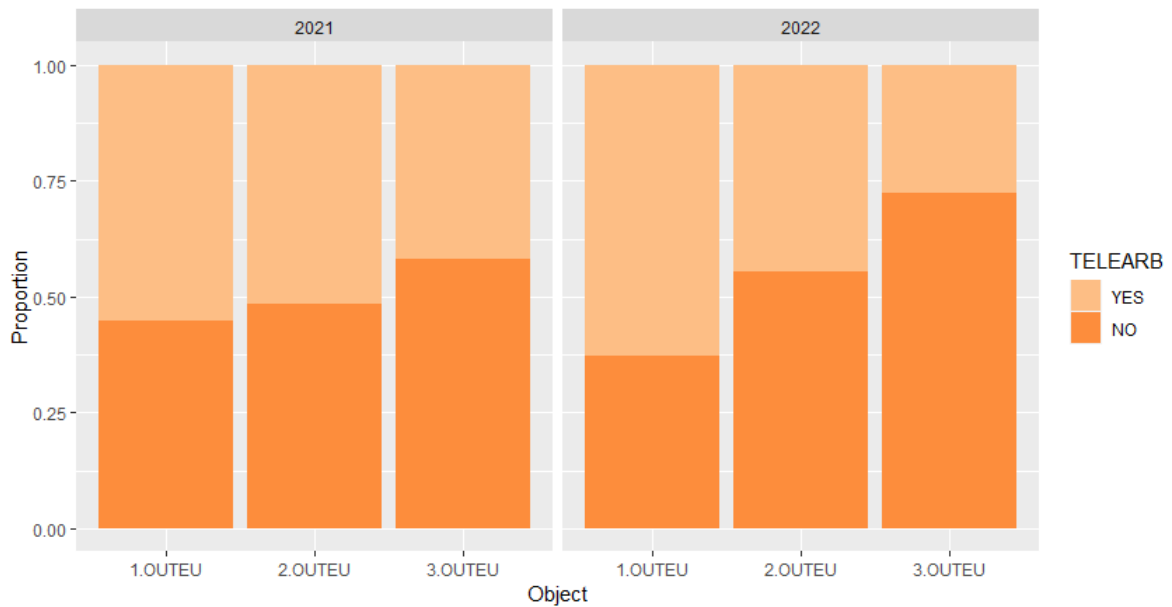


Figure 6.23: TELEARB across $C_4(2021)$ objects 2021-2022

Through the analysis of the clusters as well as of the movements of the objects from one year to the other it is possible to conclude that the people born in the neighbouring countries, especially those in groups 1.NC to 3.NC, have similar profiles and these profiles are similar to the profiles of people born in another EU country who have been in Luxembourg for 8 years or less (1.EU and 2.EU), they are in the same cluster in 7 out of 9 years. As for those who have been in Luxembourg for a long time (5.NC), conditions are generally similar to those of Luxembourg natives (they are in the same cluster for all years). Moreover, regardless of how long people have been living in Luxembourg, all Portuguese have similar profiles in the labour market, with only a slight divergence in 2021. In addition, they form a group on their own almost every year. The groups of people from countries outside the EU are the ones that move around more and are more spread out among the clusters. In general, those who have been living in Luxembourg the longest (5.OUT) have profiles that are closer not only to those of the 5.NC group but also to the natives. It is important to note that what is visible now in the recently arrived groups, may be seen, in the future, in the groups of longer established people as with time, people move from symbolic object to symbolic object.

Chapter 7

Conclusions

7.1 Main conclusions

Luxembourg has a long history of immigration. Throughout this work, the objective was to understand the different pathways in the labour market for the different immigrant groups using Symbolic Data Analysis. Six variables were chosen as the most important ones to describe the different symbolic objects. After all the analysis, it is possible to do the labour market profile of some of the groups included in the study.

First, all the EU and NC groups have similar profiles. People who were born in an EU country or in the Neighbouring countries can be characterized by their high level of education (which has been getting higher and higher), their work in white-collar professions, especially high-skilled, and their ability to do telework. When considering the amount of time that people from these groups have been living in Luxembourg, it is possible to verify that, generally, people start living in Cities but over time start moving to places with a different degree of urbanisation. People from the NC start doing it earlier. In addition, it was possible to conclude that those who have been in Luxembourg for less time, have a medium income in comparison to the rest of the society, but as time goes by they start to earn more and more and a large portion ends up being in the top 30% earners of the society. Lastly, something that distinguishes workers born in countries of the EU from workers born in the NC is the sector of activity that predominates among them. For the EU people who live in Luxembourg, it is very common to work in Arts, entertainment recreation and other service activities and in Financial and insurance activities. For the people born in a NC, Financial and insurance activities are the most common activities.

Completely in opposition to the groups just mentioned, there is the Portuguese community in Luxembourg which is mostly characterised by low levels of education and working in blue-collar professions. The most predominant activity sectors among them are Construction and Trade, transport and hospitality and due to their type of work they are not able to do telework. The Portuguese mostly settle in Towns and Suburbs. Furthermore, throughout the years the Portuguese are characterized by their low income, mainly earning amounts below the 3rd income decile. Some of this characterization of the Portuguese community in Luxembourg is in line with what was already mentioned in the article of Hildebrand et al. (2017), presented in Section 2.2.1, but this time using data from the LFS. Nevertheless, from the data, it was possible to verify that

things may be slowly changing. Nowadays, even though immigration is still mainly characterized by lower-educated people, it is possible to verify that highly educated people are also starting to immigrate to Luxembourg and work in white-collar professions. Lastly, Portuguese immigration to Luxembourg is strongly associated with work.

One can say that the profile of Luxembourgers is somewhere in between that of the Portuguese and the EU and NC people as they have medium education levels. This group is mainly characterized by working in white-collar professions and more concretely in Public administration, defence, education, health and social work activities. Nonetheless, teleworking is not predominant among them. In terms of where natives prefer to live, Rural areas stand out.

Regarding the OUTEU groups, it is a little bit difficult to characterize them as they include people born in countries with different states of development. However, Trade, transport and hospitality are very common among them and this community stands out for its high levels of unemployment compared to the other groups. Furthermore, throughout this work, it was possible to conclude that there is a strong association between people born in a non-EU country and immigrating for asylum reasons.

Considering all the 21 symbolic objects one may draw the conclusion that, in general, regardless of the country of birth, working in Business services is common among recently arrived people and for long-established people (more than 24 years living in Luxembourg) it is common to work in Public administration, defence, education, health and social work. In addition, as time of residence passes, it is possible to see a decrease in the percentage of people living in cities compared to when they first arrived in the country.

Lastly, by using the MEC framework it was possible to see that between the years 2019 to 2022 there were many movements of the objects and the majority of them can be explained by events that occurred globally. For example, from 2019 to 2020, lots of movements occurred and they may be explained by the beginning of the Pandemic which led to labour market adaptations as well as by the Brexit. Moreover, from 2021 to 2022, it is likely that the return to ‘normality’ after two years of the pandemic has had an impact on the formation of the groups. This means that it is likely that teleworking has had a high importance in the formation of clusters in 2022 because, as seen throughout the work, with the pandemic already stabilized, teleworking became something more optional. Additionally, by looking at the clusters’ height it is possible to see that, even though the Portuguese objects are almost always together in a cluster (except for 2021), the height of the class to which they belong varies over time. This means that there are years in which the profile of the different Portuguese objects is more similar than in others. Finally, in recent years (2019-2022) the maximum heights of each of the dendrograms are lower than the maximum heights of each of the dendrograms from 2014 to 2018. This could mean that the disparities between groups are decreasing.

7.2 Final Considerations and Future Work

First of all, it is important to note that some of the changes that are observed among the newcomers groups, for example, the evolution of the education attainment level, will eventually be noticed in the long term in the other groups if the trends continue.

The limitations of this work lie in the inherent constraints of the data and techniques used. These limitations are discussed throughout the work. One of the major constraints lies in the tools that are currently available to aggregate microdata and analyse symbolic data. The present work is limited by the type of symbolic variables available in the R packages as well as the clustering algorithms for symbolic data that are developed in R. Thus, for future work, it is suggested the exploration and development of other clustering algorithms in R that have already been described in the literature but are not yet implemented in any package.

Regardless of the limitations, it is important to note that the greatest contribution of this work lies in the use of a methodology that is not yet widely used in Official Statistics. Although in Portugal there are already some works that combine SDA with Official Statistical data, namely Campos (2007); Grilo (2012), there is no work that combines data from the Luxembourgish LFS with SDA. In fact, this methodology was a novelty in the Statec unit where this work was developed. However, it has an extremely important application in the field of Official Statistics since, by law and for confidentiality reasons, the National Statistics Institutes are forbidden to share individual information, thus analyzing aggregated data allows for freely sharing the results. Additionally, the aggregation of the data also allows different independent surveys made on the same population to be combined, if the same variables are used to form groups. This means that it is possible to combine the LFS with another survey that may be relevant to extend the analysis done in this work.

Additionally, it could be relevant to do a similar analysis but only consider people born in a non-EU country since not many conclusions about this group could be drawn in this work due to the fact that it includes people from countries with extremely different characteristics.

Finally, it would be interesting to continue with this analysis in the coming years to monitor the changes that occur. With this, it would be possible to understand, for example, if in fact there is a gradual change in the profile of Portuguese immigrants or not. Moreover, given the replicability of this analysis in countries where the EU-LFS is performed, it would be relevant to do something similar in countries where immigration is also a large phenomenon, such as France and Switzerland, and see if the patterns are similar to those of Luxembourg. Nevertheless, some of the variables used in the present work are from the Luxembourgish Labour Force Survey, because it was the country we were aiming to study, and are not harmonised at EU level. Consequently, some changes must be made in order to make a fair comparison between countries, e.g., choosing only variables from the EU-LFS.

Appendix

A

Transitions Taxonomy	Notation	Formal definition
Birth	$\emptyset \rightarrow C_u(t_j)$	$0 < weight(C_m(t_i), C_u(t_j)) < \tau \forall_m$
Death	$C_m(t_i) \rightarrow \emptyset$	$weight(C_m(t_i), C_u(t_j)) < \rho \forall_u$
Split	$C_m(t_i) \xrightarrow{\subseteq} \{C_1(t_j), \dots, C_r(t_j)\}$	$(\exists_u \exists_v : weight(C_m(t_i), C_u(t_j)) \geq \rho \wedge$ $(weight(C_m(t_i), C_v(t_j)) \geq \rho) \wedge$ $\sum_{u=1}^r weight(C_m(t_i), C_u(t_j)) \geq \tau$
Merge	$\{C_1(t_i), \dots, C_p(t_i)\} \xrightarrow{\subseteq} C_u(t_j)$	$(weight(C_m(t_i), C_u(t_j)) \geq \tau) \wedge$ $\exists C_p \in P_i \setminus \{C_m\} : weight(C_p(t_i), C_u(t_j)) \geq \tau$
Survival	$C_m(t_i) \rightarrow C_u(t_j)$	$(weight(C_m(t_i), C_u(t_j)) \geq \tau) \wedge$ $\nexists C_p \in P_i \setminus \{C_m\} : weight(C_p(t_i), C_u(t_j)) \geq \tau$

Table A.1: Exogenous transitions formal definitions
Adapted from Oliveira (2010)

Transition Taxonomy	Notation
Expansion	$C_m(t_i) \nearrow C_u(t_j)$
Shrinkage	$C_m(t_i) \searrow C_u(t_j)$
Compression	$C_m(t_i) \xrightarrow{\bullet} C_u(t_j)$
Dispersal	$C_m(t_i) \xrightarrow{*} C_u(t_j)$

Table A.2: Endogenous transitions notation
Adapted from Oliveira (2010)

B

```
> tes<-chisq.test(temp,simulate.p.value = TRUE)
> tes
```

Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data: temp
X-squared = 442.71, df = NA, p-value = 0.0004998

```
> resid(tes)
```

	EU	NC	OUTEU	PT
1	2.09820106	-1.05749218	-6.19918191	5.98284556
3	-0.64604156	1.75936063	3.84087980	-5.78512448
4	0.76954780	-1.71435524	1.16694671	0.11738470
5	0.02590639	2.71514655	-1.44855007	-1.90076235
6	-3.55863003	-5.96999341	14.36611200	-4.47083066
7	-3.19330341	2.78988338	0.36155564	-0.67468165

Figure B.1: Association between migration reasons and place of birth - χ^2 test results

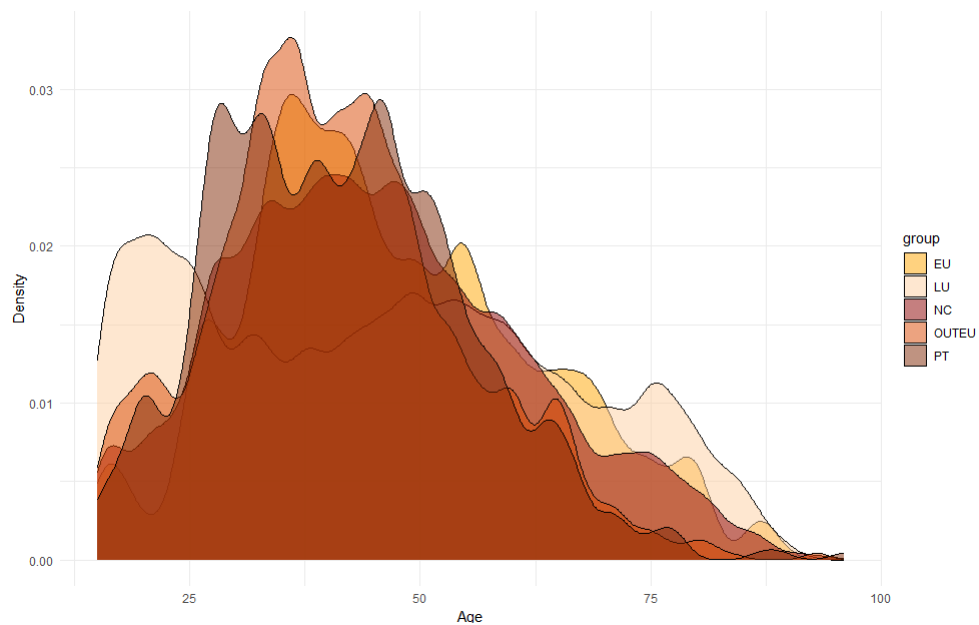


Figure B.2: Age Density Plot per Group 2014

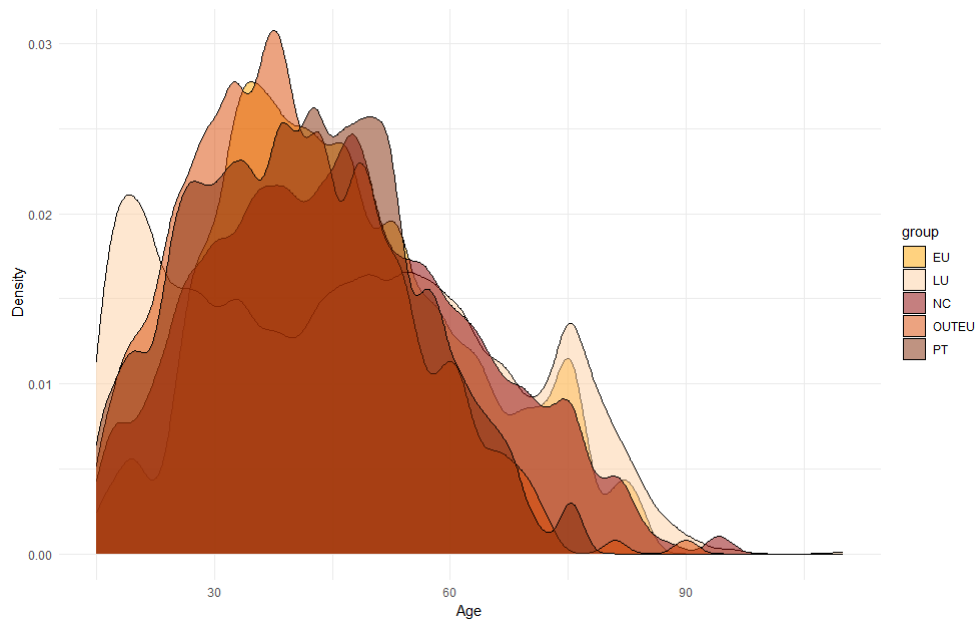


Figure B.3: Age Density Plot per Group 2015

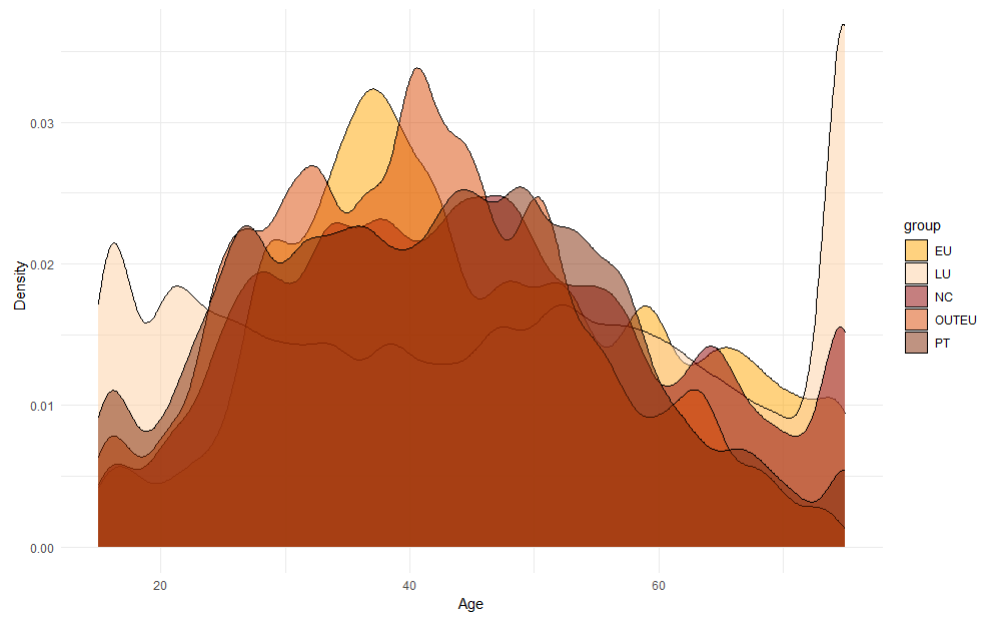


Figure B.4: Age Density Plot per Group 2016

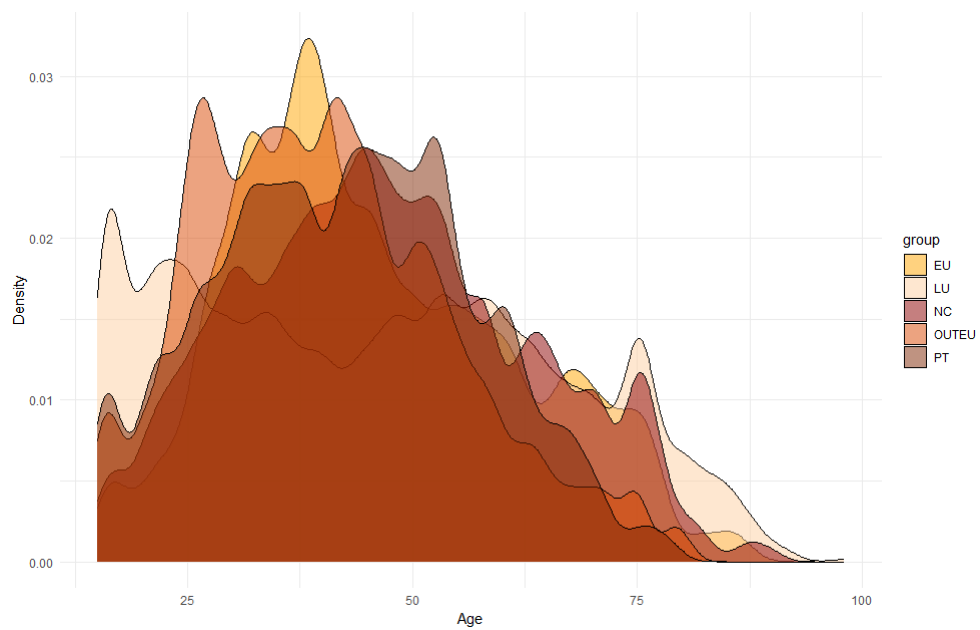


Figure B.5: Age Density Plot per Group 2017

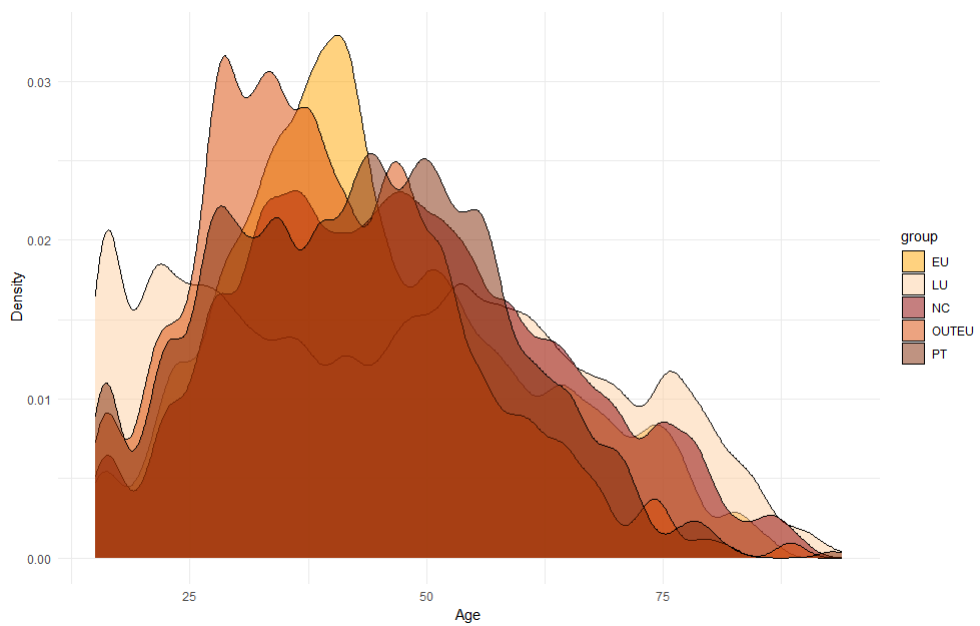


Figure B.6: Age Density Plot per Group 2018

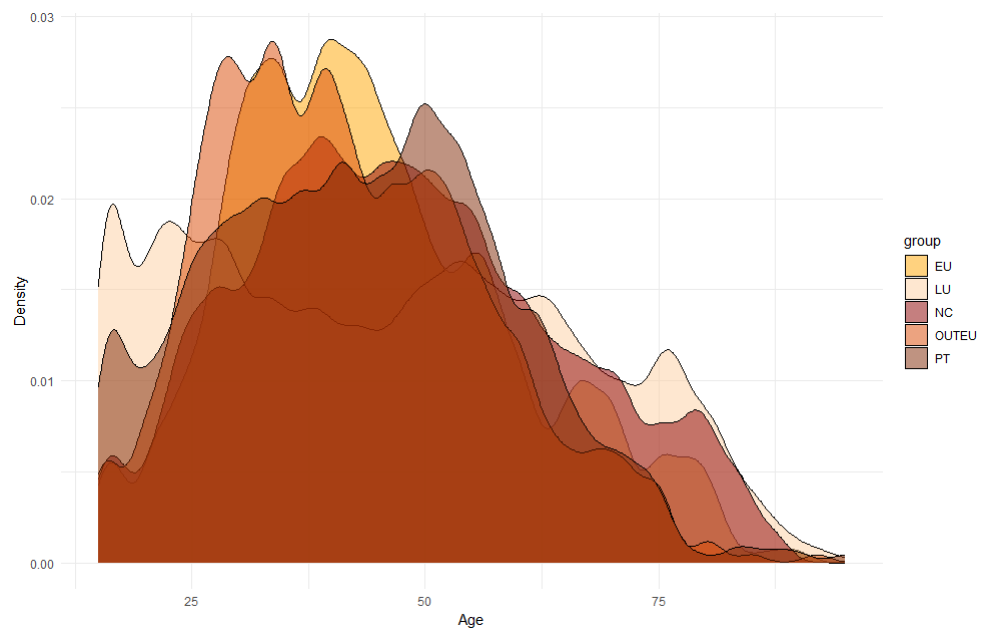


Figure B.7: Age Density Plot per Group 2019

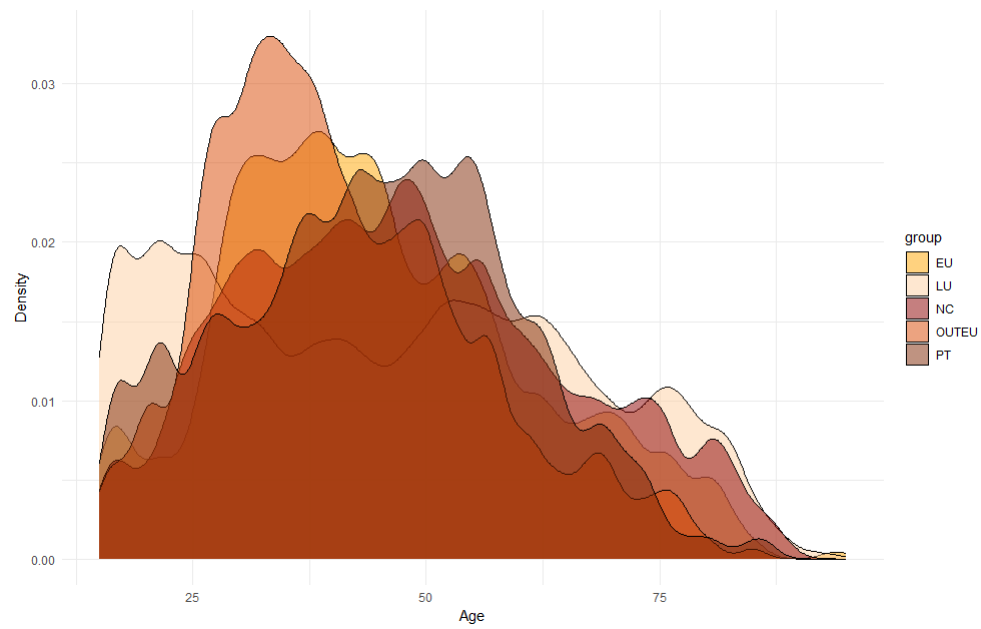


Figure B.8: Age Density Plot per Group 2020

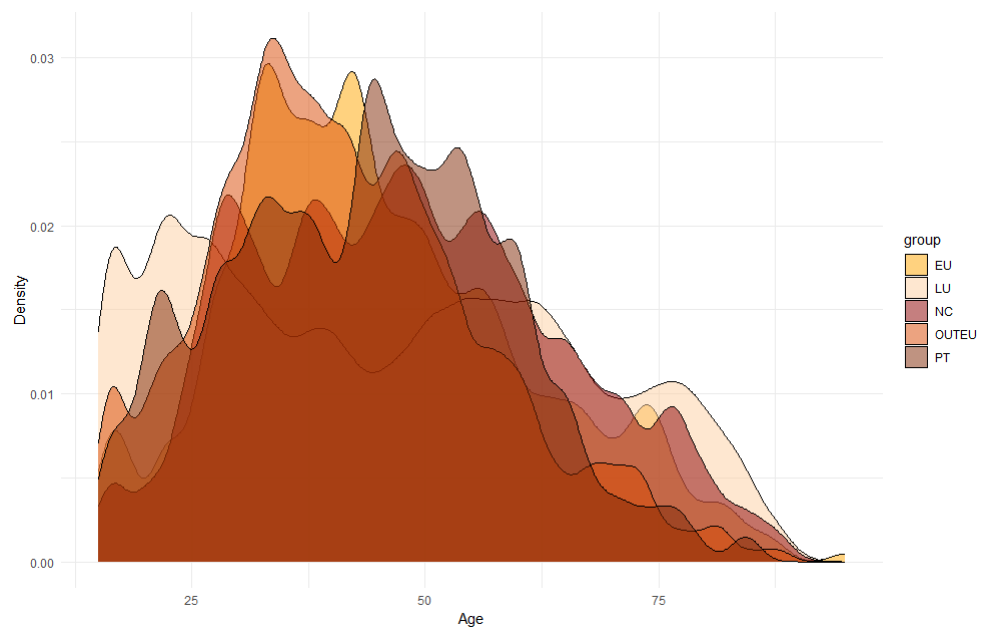


Figure B.9: Age Density Plot per Group 2021

C

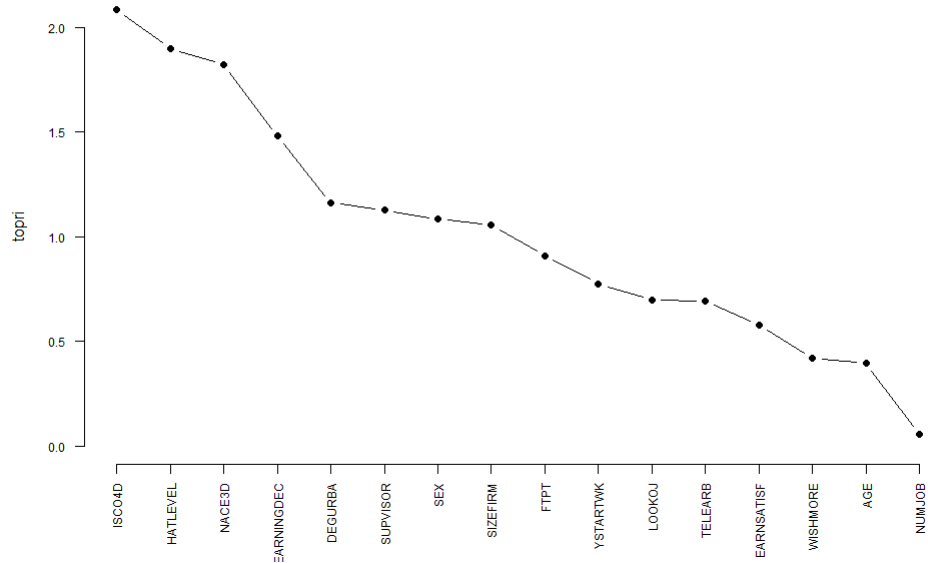


Figure C.1: TOPRI plot 2014 (all variables)

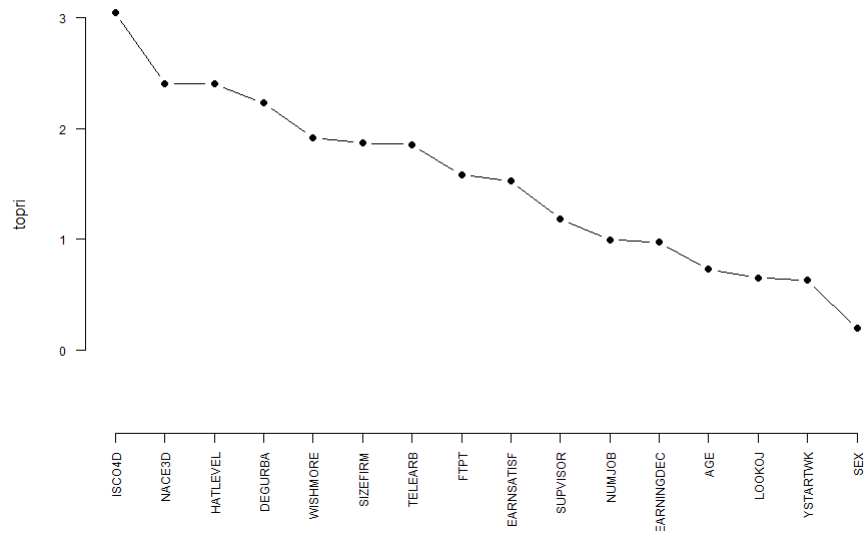


Figure C.2: TOPRI plot 2015 (all variables)

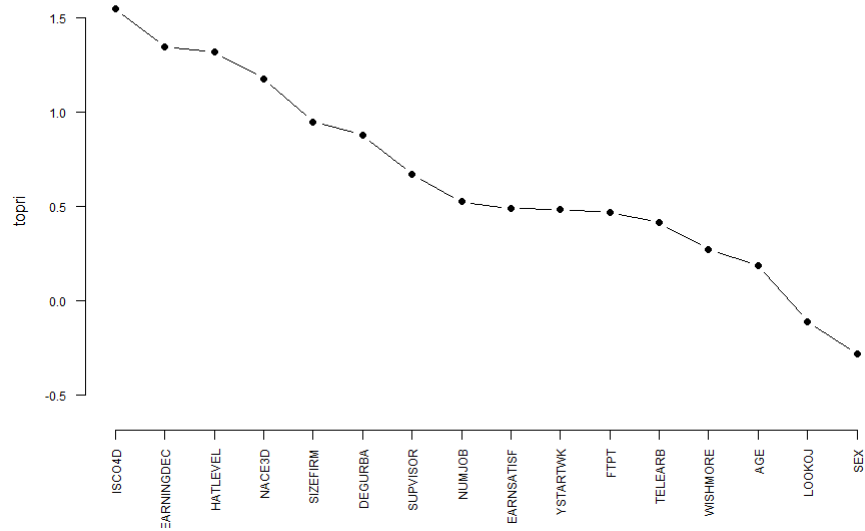


Figure C.3: TOPRI plot 2016 (all variables)

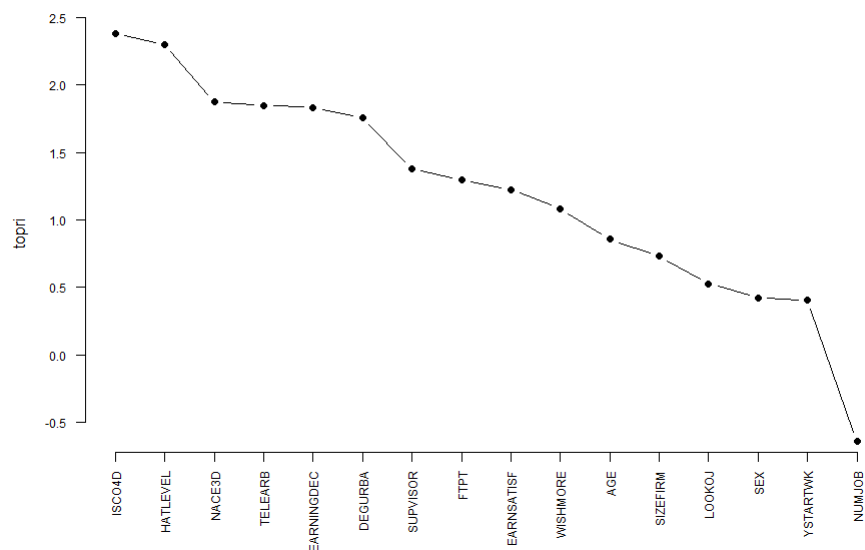


Figure C.4: TOPRI plot 2017 (all variables)

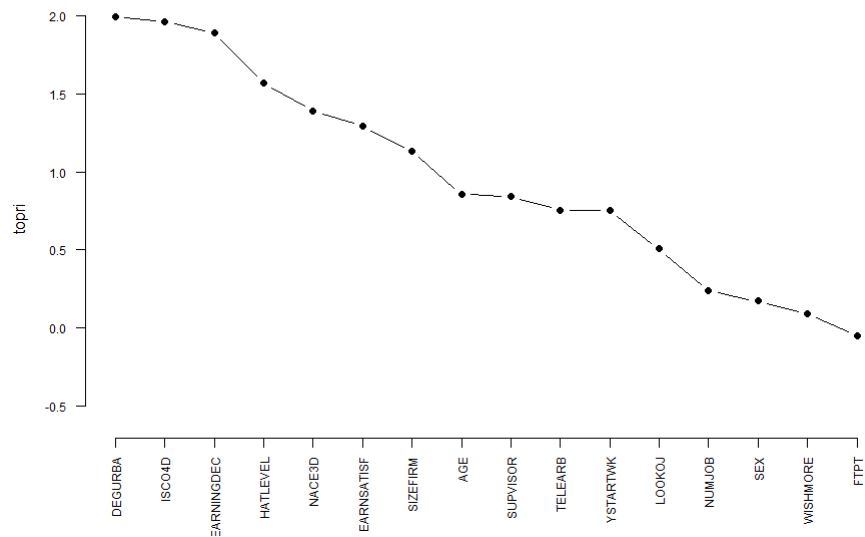


Figure C.5: TOPRI plot 2018 (all variables)

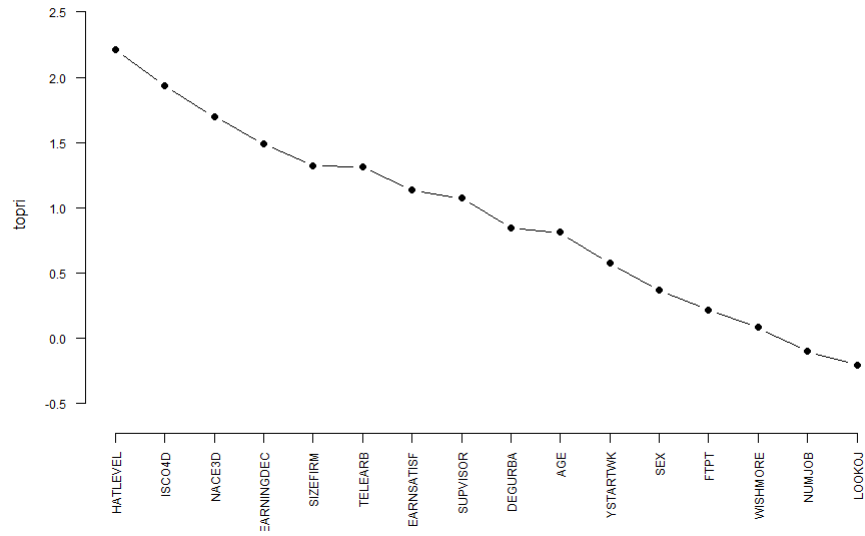


Figure C.6: TOPRI plot 2019 (all variables)

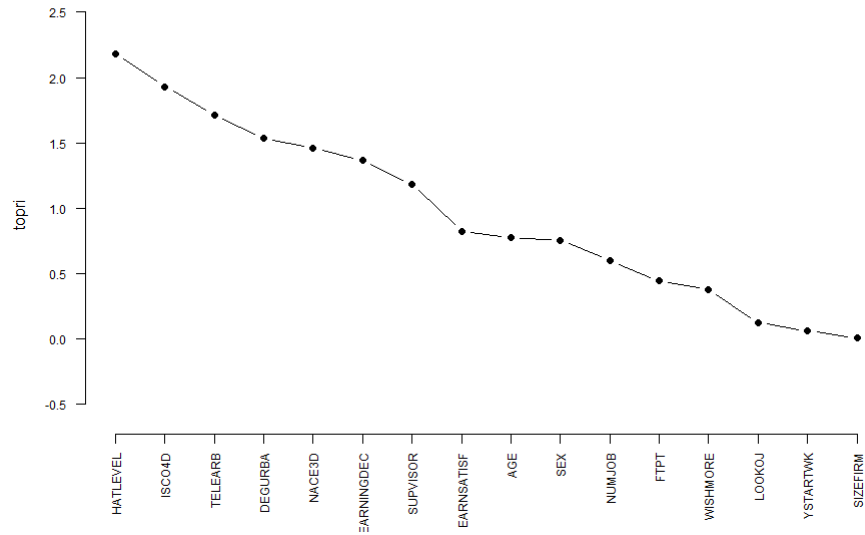


Figure C.7: TOPRI plot 2020 (all variables)

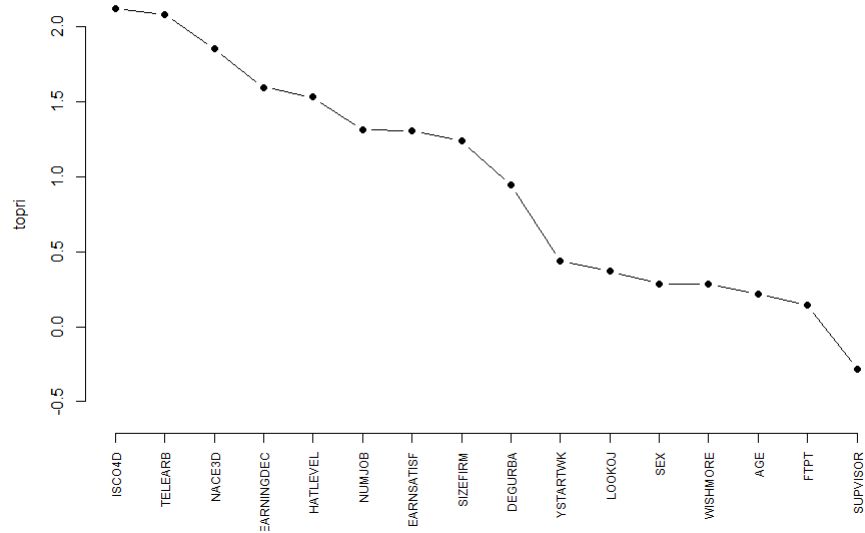


Figure C.8: TOPRI plot 2021 (all variables)

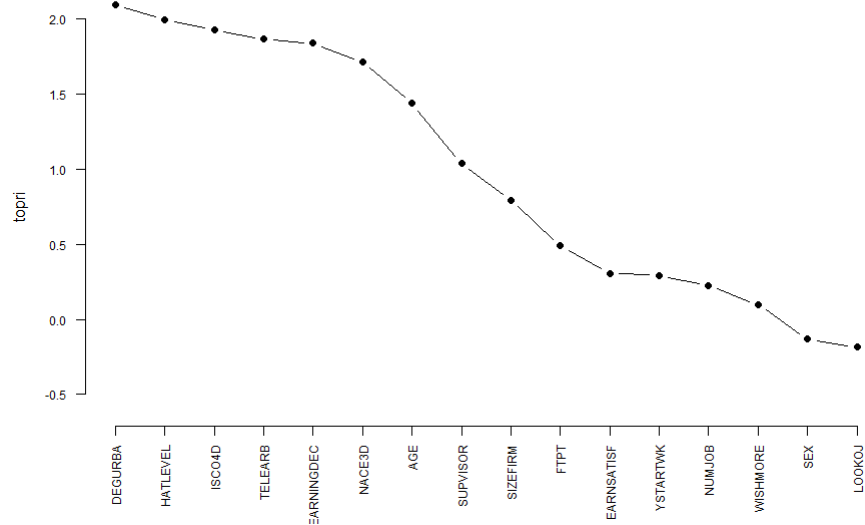


Figure C.9: TOPRI plot 2022 (all variables)

Based on C.1, the variables selected in 2014 are: ISCO4D, HATLEVEL, NACE3D and EARNINGDEC.

Based on C.2, the variables selected in 2015 are: ISCO4D, NACE3D, HATLEVEL, DEGURBA, WISHMORE, SIZEFIRM and TELEARB.

Based on C.3, the variables selected in 2016 are: ISCO4D, EARNINGDEC, HATLEVEL, NACE3D, SIZEFIRM and DEGURBA.

Based on C.4, the variables selected in 2017 are: ISCO4D, HATLEVEL, NACE3D, TELEARB, EARNINGDEC and DEGURBA.

Based on C.5, the variables selected in 2018 are: DEGURBA, ISCO4D, EARNINGDEC, HATLEVEL, NACE3D, EARNSATIF and SIZEFIRM.

Based on C.6, the variables selected in 2019 are: HATLEVEL, ISCO4D, NACE3D, EARNINGDEC, SIZEFIRM and TELEARB.

Based on C.7, the variables selected in 2020 are: HATLEVEL, ISCO4D, TELEARB, DEGURBA, NACE3D, EARNINGDEC and SUPERVISOR.

Based on C.8, the variables selected in 2021 are: ISCO4D, TELEARB, NACE3D, EARNINGDEC and HATLEVEL.

Based on C.9, the variables selected in 2022 are: DEGURBA, HATLEVEL, ISCO4D, TELEARB, EARNINGDEC and NACE3D.

	SH DEGURBA			3 SH TELEARB			2 SH NACE3D			1 10 2 3 4 5 6 7 8 9 SH TSC04D														
	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9									
1.EU	SH	3	0.683	0.133	0.183	SH	2	0.217	0.783	SH	10	0.000	0.400	0.050	0.000	0.083	0.050	0.150	0.000	0.117	0.150	SH	4	0.917
1.NC	SH	3	0.500	0.208	0.292	SH	2	0.117	0.883	SH	10	0.000	0.100	0.058	0.017	0.158	0.075	0.267	0.000	0.150	0.175	SH	4	0.842
1.OUTEU	SH	3	0.367	0.233	0.400	SH	2	0.033	0.967	SH	10	0.000	0.133	0.100	0.000	0.167	0.033	0.133	0.000	0.233	0.200	SH	4	0.767
1.PT	SH	3	0.062	0.312	0.625	SH	2	0.062	0.938	SH	10	0.000	0.188	0.125	0.062	0.250	0.066	0.000	0.000	0.250	0.062	SH	4	0.312
2.EU	SH	3	0.541	0.180	0.279	SH	2	0.295	0.705	SH	10	0.000	0.492	0.000	0.016	0.082	0.066	0.246	0.000	0.033	0.066	SH	4	0.934
2.NC	SH	3	0.423	0.221	0.356	SH	2	0.183	0.817	SH	10	0.000	0.106	0.038	0.019	0.144	0.029	0.288	0.010	0.212	0.154	SH	4	0.865
2.OUTEU	SH	3	0.475	0.250	0.275	SH	2	0.225	0.775	SH	10	0.000	0.150	0.050	0.000	0.200	0.025	0.200	0.000	0.175	0.200	SH	4	0.725
2.PT	SH	3	0.190	0.476	0.333	SH	2	0.095	0.905	SH	10	0.000	0.190	0.048	0.143	0.238	0.048	0.000	0.000	0.238	0.095	SH	4	0.048
3.EU	SH	3	0.400	0.231	0.369	SH	2	0.308	0.692	SH	10	0.000	0.400	0.031	0.000	0.092	0.046	0.262	0.000	0.046	0.123	SH	4	0.846
3.NC	SH	3	0.375	0.229	0.396	SH	2	0.167	0.833	SH	10	0.000	0.132	0.021	0.028	0.146	0.083	0.243	0.014	0.160	0.174	SH	4	0.833
3.OUTEU	SH	3	0.289	0.289	0.422	SH	2	0.244	0.756	SH	10	0.000	0.156	0.022	0.067	0.089	0.044	0.289	0.000	0.156	0.178	SH	4	0.578
3.PT	SH	3	0.244	0.311	0.444	SH	2	0.022	0.978	SH	10	0.000	0.267	0.067	0.067	0.356	0.044	0.022	0.000	0.133	0.044	SH	4	0.111
4.EU	SH	3	0.347	0.163	0.490	SH	2	0.163	0.837	SH	10	0.000	0.469	0.000	0.020	0.041	0.000	0.286	0.000	0.020	0.163	SH	4	0.898
4.NC	SH	3	0.255	0.314	0.431	SH	2	0.176	0.824	SH	10	0.000	0.167	0.039	0.029	0.176	0.049	0.196	0.000	0.098	0.245	SH	4	0.784
4.OUTEU	SH	3	0.298	0.447	0.255	SH	2	0.085	0.915	SH	10	0.000	0.149	0.021	0.043	0.234	0.064	0.085	0.000	0.149	0.255	SH	4	0.426
4.PT	SH	3	0.114	0.414	0.471	SH	2	0.014	0.986	SH	10	0.000	0.257	0.100	0.029	0.157	0.043	0.043	0.000	0.086	0.286	SH	4	0.171
5.EU	SH	3	0.302	0.302	0.396	SH	2	0.245	0.755	SH	10	0.000	0.245	0.132	0.000	0.151	0.019	0.132	0.019	0.038	0.264	SH	4	0.660
5.NC	SH	3	0.136	0.200	0.664	SH	2	0.118	0.882	SH	10	0.000	0.145	0.036	0.018	0.200	0.018	0.227	0.000	0.073	0.282	SH	4	0.718
5.OUTEU	SH	3	0.152	0.545	0.303	SH	2	0.091	0.909	SH	10	0.000	0.061	0.061	0.030	0.152	0.030	0.121	0.000	0.030	0.515	SH	4	0.485
5.PT	SH	3	0.143	0.324	0.533	SH	2	0.067	0.933	SH	10	0.000	0.181	0.143	0.076	0.267	0.029	0.048	0.000	0.048	0.210	SH	4	0.200
6.LU	SH	3	0.098	0.335	0.567	SH	2	0.092	0.908	SH	10	0.001	0.043	0.059	0.016	0.112	0.040	0.101	0.003	0.057	0.568	SH	4	0.645
	2 3 4 SH HATLEVEL			1 2 3 SH EARNINGDEC			1 10 2 3 4 5 6 7 8 9																	
1.EU	SH	0.050	0.000	0.033	SH	3	0.083	0.317	0.600	SH	10	0.100	0.050	0.050	0.033	0.217	0.017	0.133	0.100	0.200	0.100			
1.NC	SH	0.117	0.008	0.033	SH	3	0.192	0.350	0.458	SH	10	0.042	0.092	0.125	0.150	0.208	0.075	0.083	0.125	0.083	0.017			
1.OUTEU	SH	0.033	0.033	0.167	SH	3	0.233	0.233	0.533	SH	10	0.167	0.033	0.067	0.133	0.067	0.067	0.167	0.167	0.067	0.067			
1.PT	SH	0.188	0.000	0.500	SH	3	0.625	0.188	0.188	SH	10	0.312	0.000	0.375	0.125	0.062	0.000	0.062	0.000	0.062	0.000			
2.EU	SH	0.066	0.000	0.000	SH	3	0.049	0.328	0.623	SH	10	0.049	0.049	0.066	0.049	0.180	0.016	0.131	0.098	0.230	0.131			
2.NC	SH	0.077	0.019	0.038	SH	3	0.163	0.394	0.442	SH	10	0.048	0.067	0.096	0.144	0.135	0.048	0.125	0.058	0.173	0.106			
2.OUTEU	SH	0.075	0.050	0.150	SH	3	0.300	0.400	0.300	SH	10	0.150	0.175	0.175	0.075	0.050	0.025	0.125	0.100	0.075	0.050			
2.PT	SH	0.238	0.190	0.524	SH	3	0.952	0.048	0.000	SH	10	0.524	0.000	0.095	0.190	0.143	0.048	0.000	0.000	0.000	0.000			
3.EU	SH	0.108	0.015	0.031	SH	3	0.231	0.231	0.538	SH	10	0.031	0.123	0.077	0.062	0.108	0.046	0.062	0.185	0.262	0.046			
3.NC	SH	0.111	0.042	0.014	SH	3	0.264	0.417	0.319	SH	10	0.083	0.097	0.049	0.076	0.188	0.076	0.146	0.062	0.146	0.076			
3.OUTEU	SH	0.133	0.111	0.178	SH	3	0.489	0.222	0.289	SH	10	0.200	0.067	0.200	0.089	0.089	0.089	0.111	0.111	0.022	0.022			
3.PT	SH	0.222	0.089	0.578	SH	3	0.889	0.044	0.067	SH	10	0.244	0.000	0.356	0.178	0.089	0.000	0.044	0.022	0.067	0.000			
4.EU	SH	0.041	0.020	0.041	SH	3	0.204	0.367	0.429	SH	10	0.020	0.429	0.000	0.082	0.122	0.000	0.082	0.041	0.184	0.041			
4.NC	SH	0.157	0.010	0.049	SH	3	0.392	0.314	0.294	SH	10	0.088	0.098	0.078	0.098	0.137	0.059	0.069	0.098	0.216	0.059			
4.OUTEU	SH	0.277	0.043	0.255	SH	3	0.574	0.319	0.106	SH	10	0.277	0.064	0.191	0.043	0.106	0.021	0.085	0.085	0.064	0.064			
4.PT	SH	0.271	0.043	0.514	SH	3	0.886	0.071	0.043	SH	10	0.314	0.014	0.257	0.143	0.157	0.014	0.043	0.000	0.043	0.014			
5.EU	SH	0.208	0.019	0.113	SH	3	0.585	0.208	0.208	SH	10	0.132	0.208	0.094	0.075	0.057	0.019	0.189	0.094	0.075	0.057			
5.NC	SH	0.255	0.000	0.027	SH	3	0.436	0.345	0.218	SH	10	0.118	0.100	0.082	0.073	0.064	0.045	0.018	0.164	0.282	0.055			
5.OUTEU	SH	0.303	0.030	0.182	SH	3	0.636	0.242	0.121	SH	10	0.152	0.061	0.152	0.152	0.212	0.030	0.061	0.061	0.061	0.061			
5.PT	SH	0.286	0.076	0.438	SH	3	0.895	0.067	0.038	SH	10	0.133	0.019	0.295	0.152	0.210	0.038	0.076	0.019	0.038	0.019			
6.LU	SH	0.234	0.051	0.071	SH	3	0.612	0.239	0.148	SH	10	0.084	0.071	0.091	0.110	0.145	0.033	0.120	0.108	0.132	0.104			

Figure C.10: Final symbolic data table - 2014

Figure C.11: Final symbolic data table - 2015

	SH DEURBA			1	2	3	SH TELEARB	1	2	3	SH NACE3D	1	10	2	3	4	5	6	7	8	9	SH ISCO4D								
1.EU	SH	3	0.537	0.212	0.250	SH	2	0.238	0.762	SH	10	0.013	0.250	0.025	0.038	0.138	0.100	0.175	0.013	0.200	0.050	SH	4	0.825						
1.NC	SH	3	0.409	0.213	0.378	SH	2	0.291	0.709	SH	10	0.000	0.118	0.031	0.000	0.118	0.039	0.323	0.024	0.244	0.102	SH	4	0.835						
1.OUTEU	SH	3	0.308	0.354	0.338	SH	2	0.138	0.862	SH	10	0.015	0.123	0.077	0.031	0.292	0.046	0.138	0.000	0.154	0.123	SH	4	0.569						
1.PT	SH	3	0.093	0.500	0.407	SH	2	0.167	0.833	SH	10	0.000	0.056	0.074	0.167	0.352	0.000	0.000	0.296	0.056	SH	4	0.148							
2.EU	SH	3	0.364	0.258	0.379	SH	2	0.258	0.742	SH	10	0.000	0.318	0.030	0.045	0.061	0.000	0.273	0.000	0.212	0.061	SH	4	0.803						
2.NC	SH	3	0.267	0.279	0.453	SH	2	0.326	0.674	SH	10	0.000	0.128	0.023	0.000	0.209	0.058	0.163	0.000	0.233	0.186	SH	4	0.849						
2.OUTEU	SH	3	0.218	0.364	0.418	SH	2	0.236	0.764	SH	10	0.000	0.000	0.055	0.055	0.236	0.073	0.200	0.000	0.218	0.164	SH	4	0.527						
2.PT	SH	3	0.083	0.483	0.433	SH	2	0.117	0.883	SH	10	0.000	0.133	0.100	0.317	0.200	0.017	0.017	0.000	0.167	0.050	SH	4	0.133						
3.EU	SH	3	0.256	0.300	0.444	SH	2	0.278	0.722	SH	10	0.000	0.444	0.022	0.022	0.156	0.033	0.189	0.011	0.056	0.067	SH	4	0.867						
3.NC	SH	3	0.208	0.342	0.450	SH	2	0.175	0.825	SH	10	0.000	0.100	0.050	0.050	0.200	0.050	0.275	0.008	0.150	0.117	SH	4	0.700						
3.OUTEU	SH	3	0.295	0.246	0.459	SH	2	0.246	0.754	SH	10	0.000	0.115	0.049	0.049	0.230	0.033	0.148	0.000	0.180	0.197	SH	4	0.459						
3.PT	SH	3	0.071	0.464	0.464	SH	2	0.107	0.893	SH	10	0.000	0.107	0.060	0.310	0.274	0.012	0.000	0.000	0.179	0.060	SH	4	0.083						
4.EU	SH	3	0.125	0.375	0.500	SH	2	0.232	0.768	SH	10	0.000	0.232	0.018	0.071	0.125	0.018	0.286	0.000	0.089	0.161	SH	4	0.750						
4.NC	SH	3	0.165	0.243	0.591	SH	2	0.209	0.791	SH	10	0.000	0.104	0.017	0.035	0.235	0.078	0.235	0.000	0.183	0.113	SH	4	0.704						
4.OUTEU	SH	3	0.068	0.493	0.438	SH	2	0.151	0.849	SH	10	0.000	0.110	0.014	0.068	0.288	0.027	0.096	0.000	0.137	0.260	SH	4	0.301						
4.PT	SH	3	0.100	0.356	0.544	SH	2	0.089	0.911	SH	10	0.011	0.167	0.056	0.256	0.267	0.033	0.033	0.000	0.044	0.133	SH	4	0.144						
5.EU	SH	3	0.086	0.271	0.643	SH	2	0.229	0.771	SH	10	0.014	0.214	0.057	0.114	0.071	0.014	0.171	0.000	0.114	0.229	SH	4	0.671						
5.NC	SH	3	0.071	0.333	0.596	SH	2	0.121	0.879	SH	10	0.010	0.111	0.030	0.051	0.152	0.040	0.182	0.000	0.081	0.343	SH	4	0.687						
5.OUTEU	SH	3	0.082	0.449	0.469	SH	2	0.143	0.857	SH	10	0.020	0.041	0.082	0.082	0.245	0.000	0.122	0.041	0.102	0.265	SH	4	0.388						
5.PT	SH	3	0.079	0.447	0.474	SH	2	0.092	0.908	SH	10	0.007	0.112	0.099	0.257	0.197	0.007	0.039	0.007	0.099	0.178	SH	4	0.164						
6.LU	SH	3	0.066	0.316	0.618	SH	2	0.171	0.829	SH	10	0.004	0.039	0.049	0.029	0.161	0.030	0.084	0.006	0.056	0.543	SH	4	0.566						
																							2	3	4	5	6	7	8	9
1.EU	SH	3	0.075	0.038	0.062	SH	3	0.112	0.275	0.613	SH	10	0.050	0.125	0.112	0.150	0.062	0.088	0.112	0.100	0.138	0.062								
1.NC	SH	3	0.142	0.016	0.008	SH	3	0.228	0.244	0.528	SH	10	0.039	0.142	0.071	0.134	0.150	0.118	0.087	0.094	0.102	0.063								
1.OUTEU	SH	3	0.231	0.015	0.185	SH	3	0.400	0.200	0.400	SH	10	0.215	0.062	0.154	0.123	0.169	0.015	0.077	0.077	0.046	0.062								
1.PT	SH	3	0.222	0.130	0.500	SH	3	0.741	0.093	0.167	SH	10	0.222	0.000	0.333	0.185	0.130	0.037	0.019	0.019	0.037	0.019								
2.EU	SH	3	0.106	0.030	0.061	SH	3	0.242	0.288	0.470	SH	10	0.061	0.258	0.015	0.061	0.076	0.045	0.091	0.076	0.227	0.091								
2.NC	SH	3	0.105	0.000	0.047	SH	3	0.233	0.326	0.442	SH	10	0.058	0.105	0.081	0.105	0.151	0.035	0.070	0.151	0.151	0.093								
2.OUTEU	SH	3	0.145	0.073	0.255	SH	3	0.400	0.182	0.418	SH	10	0.164	0.091	0.182	0.127	0.091	0.109	0.073	0.018	0.036	0.109								
2.PT	SH	3	0.100	0.317	0.450	SH	3	0.900	0.050	0.050	SH	10	0.217	0.000	0.200	0.233	0.100	0.117	0.067	0.017	0.017	0.033								
3.EU	SH	3	0.067	0.000	0.067	SH	3	0.178	0.300	0.522	SH	10	0.078	0.278	0.022	0.022	0.033	0.022	0.022	0.178	0.133	0.211								
3.NC	SH	3	0.208	0.050	0.042	SH	3	0.267	0.383	0.350	SH	10	0.033	0.183	0.083	0.075	0.100	0.108	0.083	0.117	0.142	0.075								
3.OUTEU	SH	3	0.197	0.049	0.295	SH	3	0.541	0.262	0.197	SH	10	0.197	0.180	0.213	0.115	0.115	0.016	0.000	0.033	0.066	0.066								
3.PT	SH	3	0.190	0.214	0.512	SH	3	0.964	0.012	0.024	SH	10	0.190	0.012	0.250	0.202	0.202	0.107	0.012	0.024	0.000	0.000								
4.EU	SH	3	0.161	0.036	0.054	SH	3	0.339	0.304	0.357	SH	10	0.054	0.304	0.143	0.036	0.071	0.054	0.054	0.089	0.107	0.089								
4.NC	SH	3	0.209	0.035	0.052	SH	3	0.313	0.391	0.296	SH	10	0.061	0.174	0.104	0.061	0.113	0.035	0.070	0.148	0.104	0.130								
4.OUTEU	SH	3	0.260	0.082	0.356	SH	3	0.740	0.192	0.068	SH	10	0.205	0.027	0.164	0.178	0.151	0.068	0.055	0.096	0.000	0.055								
4.PT	SH	3	0.189	0.211	0.456	SH	3	0.911	0.067	0.022	SH	10	0.189	0.011	0.111	0.156	0.300	0.111	0.078	0.044	0.000	0.000								
5.EU	SH	3	0.129	0.057	0.143	SH	3	0.486	0.371	0.143	SH	10	0.086	0.257	0.071	0.057	0.086	0.043	0.057	0.071	0.100	0.171								
5.NC	SH	3	0.222	0.020	0.071	SH	3	0.475	0.242	0.283	SH	10	0.051	0.192	0.071	0.071	0.071	0.040	0.162	0.091	0.131	0.121								
5.OUTEU	SH	3	0.286	0.102	0.224	SH	3	0.633	0.163	0.204	SH	10	0.061	0.041	0.102	0.184	0.082	0.122	0.143	0.184	0.020	0.061								
5.PT	SH	3	0.164	0.151	0.520	SH	3	0.921	0.059	0.020	SH	10	0.164	0.007	0.151	0.112	0.178	0.178	0.105	0.046	0.033	0.026								
6.LU	SH	3	0.282	0.058	0.094	SH	3	0.669	0.181	0.150	SH	10	0.097	0.101	0.057	0.077	0.094	0.088	0.116	0.125	0.118	0.127								

Figure C.12: Final symbolic data table - 2016

	SH DEURBA																					1	2	3	SH TELEARB	1	2	3	SH NACE3D	1	10	2	3	4	5	6	7	8	9	SH ISCO4D	1
1.1.EU	SH	3	0.667	0.198	0.135	SH	2	0.294	0.706	SH	10	0.000	0.238	0.016	0.008	0.159	0.056	0.190	0.008	0.278	0.048	SH	4	0.849																	
1.1.NC	SH	3	0.450	0.298	0.252	SH	2	0.351	0.649	SH	10	0.000	0.107	0.046	0.031	0.115	0.069	0.214	0.015	0.267	0.137	SH	4	0.824																	
1.1.UTEU	SH	3	0.476	0.317	0.207	SH	2	0.159	0.841	SH	10	0.000	0.049	0.037	0.024	0.146	0.146	0.159	0.012	0.317	0.110	SH	4	0.671																	
1.1.PT	SH	3	0.182	0.500	0.318	SH	2	0.205	0.795	SH	10	0.000	0.068	0.091	0.273	0.182	0.000	0.045	0.000	0.273	0.068	SH	4	0.182																	
2.2.EU	SH	3	0.511	0.250	0.239	SH	2	0.284	0.716	SH	10	0.000	0.341	0.000	0.023	0.125	0.045	0.205	0.000	0.182	0.080	SH	4	0.852																	
2.2.NC	SH	3	0.422	0.284	0.294	SH	2	0.333	0.667	SH	10	0.000	0.098	0.039	0.020	0.186	0.078	0.225	0.010	0.196	0.147	SH	4	0.814																	
2.2.UTEU	SH	3	0.304	0.429	0.268	SH	2	0.304	0.696	SH	10	0.000	0.071	0.054	0.054	0.250	0.107	0.161	0.018	0.143	0.143	SH	4	0.571																	
2.2.PT	SH	3	0.051	0.576	0.373	SH	2	0.068	0.932	SH	10	0.000	0.102	0.102	0.322	0.237	0.000	0.034	0.000	0.102	0.102	SH	4	0.051																	
3.3.EU	SH	3	0.436	0.351	0.213	SH	2	0.383	0.617	SH	10	0.000	0.521	0.032	0.011	0.085	0.032	0.149	0.011	0.096	0.064	SH	4	0.862																	
3.3.NC	SH	3	0.299	0.340	0.361	SH	2	0.229	0.771	SH	10	0.000	0.139	0.056	0.021	0.139	0.035	0.264	0.000	0.174	0.174	SH	4	0.799																	
3.3.UTEU	SH	3	0.180	0.492	0.328	SH	2	0.148	0.852	SH	10	0.000	0.115	0.066	0.049	0.311	0.000	0.131	0.000	0.164	0.164	SH	4	0.475																	
3.3.PT	SH	3	0.098	0.625	0.277	SH	2	0.116	0.884	SH	10	0.000	0.152	0.089	0.259	0.179	0.000	0.000	0.000	0.196	0.125	SH	4	0.107																	
4.4.EU	SH	3	0.290	0.339	0.371	SH	2	0.161	0.839	SH	10	0.000	0.258	0.000	0.016	0.097	0.048	0.210	0.000	0.129	0.242	SH	4	0.758																	
4.4.NC	SH	3	0.181	0.315	0.503	SH	2	0.208	0.792	SH	10	0.000	0.114	0.020	0.013	0.195	0.081	0.289	0.027	0.087	0.174	SH	4	0.799																	
4.4.UTEU	SH	3	0.188	0.435	0.377	SH	2	0.174	0.826	SH	10	0.000	0.087	0.014	0.029	0.304	0.043	0.145	0.029	0.087	0.261	SH	4	0.522																	
4.4.PT	SH	3	0.165	0.474	0.361	SH	2	0.082	0.918	SH	10	0.000	0.186	0.062	0.268	0.237	0.010	0.010	0.000	0.093	0.134	SH	4	0.103																	
5.5.EU	SH	3	0.190	0.476	0.333	SH	2	0.190	0.810	SH	10	0.000	0.206	0.032	0.032	0.190	0.016	0.222	0.000	0.095	0.206	SH	4	0.603																	
5.5.NC	SH	3	0.150	0.327	0.523	SH	2	0.206	0.794	SH	10	0.009	0.112	0.075	0.028	0.168	0.056	0.140	0.009	0.084	0.318	SH	4	0.673																	
5.5.UTEU	SH	3	0.130	0.449	0.420	SH	2	0.174	0.826	SH	10	0.000	0.029	0.058	0.058	0.188	0.000	0.116	0.000	0.145	0.406	SH	4	0.391																	
5.5.PT	SH	3	0.096	0.515	0.389	SH	2	0.054	0.946	SH	10	0.000	0.126	0.060	0.234	0.222	0.018	0.060	0.024	0.048	0.210	SH	4	0.126																	
6.6.LU	SH	3	0.082	0.400	0.518	SH	2	0.150	0.850	SH	10	0.005	0.031	0.041	0.027	0.142	0.037	0.095	0.004	0.063	0.555	SH	4	0.566																	

	2	3	4	SH HATLEVEL	1	2	3	SH EARNINGDEC	1	10	2	3	4	5	6	7	8	9		
1.1.EU	0.063	0.000	0.087	SH	3	0.167	0.159	0.675	SH	10	0.056	0.103	0.103	0.056	0.103	0.119	0.079	0.135	0.167	0.079
1.1.NC	0.145	0.023	0.008	SH	3	0.198	0.168	0.634	SH	10	0.046	0.137	0.069	0.122	0.137	0.145	0.107	0.115	0.061	0.061
1.1.UTEU	0.061	0.037	0.232	SH	3	0.293	0.159	0.549	SH	10	0.207	0.085	0.110	0.122	0.220	0.073	0.037	0.049	0.049	0.049
1.1.PT	0.159	0.182	0.477	SH	3	0.773	0.023	0.205	SH	10	0.250	0.023	0.364	0.114	0.114	0.045	0.023	0.023	0.023	0.023
2.2.EU	0.080	0.034	0.034	SH	3	0.182	0.205	0.614	SH	10	0.045	0.205	0.057	0.023	0.091	0.102	0.114	0.091	0.193	0.080
2.2.NC	0.108	0.020	0.059	SH	3	0.216	0.167	0.618	SH	10	0.029	0.137	0.118	0.049	0.098	0.108	0.157	0.137	0.098	0.069
2.2.UTEU	0.125	0.036	0.268	SH	3	0.446	0.161	0.393	SH	10	0.161	0.161	0.268	0.018	0.071	0.036	0.036	0.107	0.054	0.089
2.2.PT	0.153	0.288	0.508	SH	3	0.898	0.017	0.085	SH	10	0.237	0.017	0.390	0.119	0.102	0.102	0.017	0.017	0.000	0.000
3.3.EU	0.085	0.021	0.032	SH	3	0.149	0.319	0.532	SH	10	0.096	0.298	0.053	0.011	0.032	0.064	0.096	0.117	0.138	0.096
3.3.NC	0.132	0.035	0.035	SH	3	0.181	0.250	0.569	SH	10	0.049	0.174	0.069	0.049	0.083	0.083	0.090	0.139	0.174	0.090
3.3.UTEU	0.164	0.098	0.262	SH	3	0.574	0.148	0.279	SH	10	0.148	0.082	0.213	0.066	0.082	0.033	0.098	0.148	0.082	0.049
3.3.PT	0.161	0.232	0.500	SH	3	0.920	0.027	0.054	SH	10	0.241	0.009	0.259	0.196	0.143	0.089	0.018	0.027	0.009	0.009
4.4.EU	0.113	0.016	0.113	SH	3	0.274	0.242	0.484	SH	10	0.081	0.323	0.097	0.016	0.113	0.081	0.113	0.032	0.048	0.097
4.4.NC	0.128	0.013	0.060	SH	3	0.262	0.242	0.497	SH	10	0.054	0.134	0.087	0.101	0.074	0.087	0.094	0.107	0.148	0.114
4.4.UTEU	0.246	0.043	0.188	SH	3	0.565	0.174	0.261	SH	10	0.116	0.087	0.130	0.116	0.159	0.029	0.130	0.130	0.029	0.072
4.4.PT	0.144	0.216	0.536	SH	3	0.938	0.021	0.041	SH	10	0.165	0.000	0.258	0.175	0.186	0.134	0.041	0.010	0.010	0.021
5.5.EU	0.190	0.048	0.159	SH	3	0.571	0.159	0.270	SH	10	0.063	0.206	0.079	0.063	0.048	0.143	0.063	0.111	0.095	0.127
5.5.NC	0.252	0.009	0.065	SH	3	0.514	0.196	0.290	SH	10	0.103	0.206	0.093	0.093	0.084	0.047	0.093	0.093	0.112	0.075
5.5.UTEU	0.232	0.087	0.290	SH	3	0.696	0.116	0.188	SH	10	0.145	0.043	0.116	0.000	0.145	0.159	0.188	0.058	0.101	0.043
5.5.PT	0.240	0.210	0.425	SH	3	0.940	0.018	0.042	SH	10	0.126	0.012	0.293	0.108	0.144	0.126	0.090	0.072	0.006	0.024
6.6.LU	0.276	0.075	0.083	SH	3	0.670	0.176	0.155	SH	10	0.094	0.090	0.095	0.072	0.087	0.106	0.115	0.118	0.108	0.117

Figure C.13: Final symbolic data table - 2017

Figure C.14: Final symbolic data table - 2018

	SH	DEGURBA	1	2	3	SH	TELEARB	1	2	SH	NACE3D	1	10	2	3	4	5	6	7	8	9	SH	ISCO4D	1
1.EU	\$H	3	0.620	0.223	0.157	\$H	2	0.289	0.711	\$H	10	0.006	0.265	0.006	0.018	0.163	0.084	0.169	0.012	0.193	0.084	\$H	4	0.880
1.NC	\$H	3	0.510	0.282	0.209	\$H	2	0.257	0.743	\$H	10	0.000	0.121	0.029	0.010	0.199	0.078	0.214	0.010	0.243	0.097	\$H	4	0.835
1.OUTEU	\$H	3	0.469	0.352	0.179	\$H	2	0.228	0.772	\$H	10	0.000	0.069	0.041	0.034	0.262	0.103	0.179	0.000	0.228	0.083	\$H	4	0.697
1.PT	\$H	3	0.200	0.455	0.345	\$H	2	0.164	0.836	\$H	10	0.000	0.127	0.073	0.218	0.291	0.000	0.109	0.000	0.109	0.073	\$H	4	0.327
2.EU	\$H	3	0.452	0.266	0.282	\$H	2	0.242	0.758	\$H	10	0.016	0.282	0.000	0.024	0.129	0.081	0.194	0.000	0.169	0.105	\$H	4	0.774
2.NC	\$H	3	0.311	0.353	0.336	\$H	2	0.227	0.773	\$H	10	0.000	0.118	0.034	0.050	0.126	0.017	0.277	0.008	0.176	0.193	\$H	4	0.824
2.OUTEU	\$H	3	0.230	0.459	0.311	\$H	2	0.176	0.824	\$H	10	0.000	0.068	0.081	0.068	0.243	0.068	0.122	0.000	0.230	0.122	\$H	4	0.514
2.PT	\$H	3	0.136	0.568	0.296	\$H	2	0.099	0.901	\$H	10	0.000	0.136	0.012	0.259	0.259	0.025	0.049	0.000	0.210	0.049	\$H	4	0.222
3.EU	\$H	3	0.422	0.293	0.286	\$H	2	0.347	0.653	\$H	10	0.007	0.524	0.014	0.014	0.088	0.020	0.129	0.007	0.088	0.109	\$H	4	0.857
3.NC	\$H	3	0.277	0.370	0.353	\$H	2	0.266	0.734	\$H	10	0.000	0.098	0.065	0.038	0.120	0.038	0.277	0.005	0.223	0.136	\$H	4	0.745
3.OUTEU	\$H	3	0.288	0.450	0.262	\$H	2	0.200	0.800	\$H	10	0.000	0.125	0.075	0.075	0.238	0.038	0.125	0.013	0.138	0.175	\$H	4	0.450
3.PT	\$H	3	0.099	0.559	0.342	\$H	2	0.090	0.910	\$H	10	0.000	0.117	0.054	0.216	0.207	0.018	0.036	0.000	0.198	0.153	\$H	4	0.135
4.EU	\$H	3	0.400	0.300	0.300	\$H	2	0.271	0.729	\$H	10	0.000	0.457	0.000	0.000	0.114	0.071	0.171	0.000	0.129	0.057	\$H	4	0.857
4.NC	\$H	3	0.258	0.391	0.352	\$H	2	0.211	0.789	\$H	10	0.000	0.164	0.016	0.016	0.172	0.055	0.211	0.031	0.133	0.203	\$H	4	0.734
4.OUTEU	\$H	3	0.145	0.609	0.246	\$H	2	0.145	0.855	\$H	10	0.000	0.043	0.058	0.087	0.275	0.043	0.087	0.000	0.087	0.319	\$H	4	0.319
4.PT	\$H	3	0.072	0.546	0.381	\$H	2	0.175	0.825	\$H	10	0.010	0.144	0.062	0.186	0.206	0.021	0.031	0.000	0.165	0.175	\$H	4	0.155
5.EU	\$H	3	0.175	0.397	0.429	\$H	2	0.159	0.841	\$H	10	0.000	0.333	0.032	0.048	0.127	0.016	0.095	0.000	0.048	0.302	\$H	4	0.667
5.NC	\$H	3	0.127	0.405	0.468	\$H	2	0.159	0.841	\$H	10	0.000	0.151	0.079	0.040	0.151	0.024	0.111	0.000	0.119	0.325	\$H	4	0.677
5.OUTEU	\$H	3	0.200	0.400	0.400	\$H	2	0.154	0.846	\$H	10	0.000	0.092	0.092	0.077	0.154	0.031	0.138	0.000	0.108	0.308	\$H	4	0.477
5.PT	\$H	3	0.144	0.521	0.335	\$H	2	0.132	0.868	\$H	10	0.006	0.120	0.048	0.198	0.228	0.012	0.030	0.012	0.084	0.263	\$H	4	0.222
6.LU	\$H	3	0.083	0.438	0.479	\$H	2	0.135	0.865	\$H	10	0.002	0.044	0.050	0.030	0.136	0.036	0.075	0.005	0.056	0.566	\$H	4	0.596
1.EU	\$H	3	0.078	0.018	0.024	\$H	3	0.175	0.301	0.524	\$H	10	0.084	0.108	0.084	0.108	0.090	0.078	0.133	0.120	0.108	0.084	\$H	9
1.NC	\$H	3	0.097	0.019	0.049	\$H	3	0.214	0.243	0.544	\$H	10	0.092	0.146	0.097	0.083	0.087	0.160	0.087	0.117	0.068	0.063	\$H	9
1.OUTEU	\$H	3	0.083	0.014	0.207	\$H	3	0.241	0.331	0.428	\$H	10	0.152	0.062	0.145	0.110	0.097	0.110	0.097	0.055	0.076	0.097	\$H	9
1.PT	\$H	3	0.236	0.182	0.255	\$H	3	0.636	0.127	0.236	\$H	10	0.182	0.000	0.327	0.182	0.091	0.055	0.036	0.055	0.036	0.036	\$H	9
2.EU	\$H	3	0.113	0.065	0.048	\$H	3	0.234	0.266	0.500	\$H	10	0.089	0.073	0.089	0.089	0.081	0.073	0.185	0.065	0.161	0.097	\$H	9
2.NC	\$H	3	0.109	0.025	0.042	\$H	3	0.227	0.252	0.521	\$H	10	0.076	0.101	0.059	0.076	0.076	0.126	0.118	0.134	0.151	0.084	\$H	9
2.OUTEU	\$H	3	0.081	0.081	0.324	\$H	3	0.486	0.230	0.284	\$H	10	0.149	0.081	0.230	0.149	0.081	0.068	0.068	0.054	0.068	0.054	\$H	9
2.PT	\$H	3	0.136	0.210	0.432	\$H	3	0.753	0.062	0.185	\$H	10	0.210	0.025	0.284	0.185	0.160	0.062	0.037	0.012	0.012	0.012	\$H	9
3.EU	\$H	3	0.088	0.034	0.020	\$H	3	0.163	0.204	0.633	\$H	10	0.041	0.333	0.061	0.034	0.048	0.068	0.082	0.082	0.156	0.109	\$H	9
3.NC	\$H	3	0.136	0.054	0.065	\$H	3	0.277	0.217	0.505	\$H	10	0.092	0.174	0.103	0.054	0.071	0.098	0.076	0.109	0.103	0.120	\$H	9
3.OUTEU	\$H	3	0.188	0.050	0.312	\$H	3	0.500	0.188	0.312	\$H	10	0.138	0.112	0.300	0.100	0.062	0.088	0.038	0.050	0.062	0.050	\$H	9
3.PT	\$H	3	0.261	0.189	0.414	\$H	3	0.901	0.045	0.054	\$H	10	0.189	0.009	0.306	0.216	0.126	0.072	0.027	0.009	0.027	0.018	\$H	9
4.EU	\$H	3	0.100	0.000	0.043	\$H	3	0.186	0.329	0.486	\$H	10	0.029	0.300	0.086	0.071	0.114	0.014	0.086	0.037	0.100	0.143	\$H	9
4.NC	\$H	3	0.172	0.023	0.070	\$H	3	0.305	0.328	0.367	\$H	10	0.078	0.172	0.086	0.070	0.125	0.102	0.102	0.094	0.078	0.094	\$H	9
4.OUTEU	\$H	3	0.232	0.159	0.290	\$H	3	0.710	0.203	0.087	\$H	10	0.130	0.014	0.217	0.203	0.159	0.087	0.043	0.058	0.043	0.043	\$H	9
4.PT	\$H	3	0.227	0.196	0.423	\$H	3	0.907	0.052	0.041	\$H	10	0.175	0.021	0.247	0.196	0.175	0.072	0.031	0.041	0.021	0.021	\$H	9
5.EU	\$H	3	0.190	0.016	0.127	\$H	3	0.429	0.333	0.238	\$H	10	0.079	0.286	0.063	0.095	0.063	0.032	0.079	0.095	0.095	0.111	\$H	9
5.NC	\$H	3	0.190	0.032	0.111	\$H	3	0.484	0.222	0.294	\$H	10	0.048	0.159	0.087	0.056	0.095	0.079	0.095	0.111	0.127	0.143	\$H	9
5.OUTEU	\$H	3	0.246	0.062	0.215	\$H	3	0.554	0.200	0.246	\$H	10	0.108	0.077	0.108	0.092	0.123	0.077	0.123	0.154	0.077	0.062	\$H	9
5.PT	\$H	3	0.234	0.150	0.395	\$H	3	0.862	0.072	0.066	\$H	10	0.168	0.024	0.168	0.150	0.162	0.144	0.102	0.042	0.024	0.018	\$H	9
6.LU	\$H	3	0.257	0.067	0.079	\$H	3	0.630	0.204	0.166	\$H	10	0.087	0.094	0.078	0.070	0.091	0.105	0.119	0.111	0.118	0.128	\$H	9

Figure C.15: Final symbolic data table - 2019

	SH DEGURBA			3 SH TELEARB			2 SH NACE3D			1 10 2 3 4 5 6 7 8 9			SH ISCO4D						
	1	2	3	1	2	3	1	2	3	4	5	6	7	8	9				
1.EU \$H	3 0.662	0.162	0.176	\$H	2 0.610	0.390	\$H	10 0.000	0.228	0.037	0.022	0.147	0.051	0.199	0.000	0.250	0.066	\$H	4 0.860
1.NC \$H	3 0.452	0.280	0.268	\$H	2 0.548	0.452	\$H	10 0.006	0.089	0.071	0.042	0.131	0.077	0.208	0.000	0.226	0.149	\$H	4 0.875
1.0UTEU \$H	3 0.500	0.321	0.179	\$H	2 0.537	0.463	\$H	10 0.006	0.056	0.031	0.062	0.216	0.093	0.160	0.000	0.247	0.130	\$H	4 0.772
1.PT \$H	3 0.273	0.523	0.205	\$H	2 0.341	0.659	\$H	10 0.000	0.182	0.045	0.159	0.364	0.023	0.136	0.023	0.045	0.023	\$H	4 0.432
2.EU \$H	3 0.477	0.297	0.227	\$H	2 0.672	0.328	\$H	10 0.000	0.234	0.016	0.008	0.156	0.070	0.258	0.008	0.133	0.117	\$H	4 0.852
2.NC \$H	3 0.359	0.385	0.256	\$H	2 0.590	0.410	\$H	10 0.009	0.103	0.051	0.009	0.137	0.026	0.239	0.009	0.197	0.222	\$H	4 0.855
2.0UTEU \$H	3 0.389	0.375	0.236	\$H	2 0.486	0.514	\$H	10 0.000	0.056	0.083	0.042	0.375	0.083	0.139	0.014	0.139	0.069	\$H	4 0.625
2.PT \$H	3 0.081	0.532	0.387	\$H	2 0.161	0.839	\$H	10 0.000	0.081	0.048	0.210	0.210	0.000	0.048	0.000	0.339	0.065	\$H	4 0.145
3.EU \$H	3 0.461	0.273	0.266	\$H	2 0.562	0.438	\$H	10 0.008	0.453	0.031	0.023	0.094	0.031	0.141	0.023	0.094	0.102	\$H	4 0.805
3.NC \$H	3 0.325	0.298	0.377	\$H	2 0.550	0.450	\$H	10 0.005	0.120	0.058	0.037	0.141	0.021	0.241	0.005	0.215	0.157	\$H	4 0.812
3.0UTEU \$H	3 0.267	0.406	0.327	\$H	2 0.485	0.515	\$H	10 0.000	0.139	0.040	0.099	0.168	0.030	0.208	0.000	0.178	0.139	\$H	4 0.584
3.PT \$H	3 0.067	0.611	0.322	\$H	2 0.089	0.911	\$H	10 0.000	0.133	0.022	0.333	0.211	0.022	0.033	0.000	0.122	0.122	\$H	4 0.189
4.EU \$H	3 0.403	0.313	0.284	\$H	2 0.507	0.493	\$H	10 0.000	0.358	0.060	0.015	0.134	0.030	0.224	0.000	0.030	0.149	\$H	4 0.701
4.NC \$H	3 0.316	0.271	0.414	\$H	2 0.556	0.444	\$H	10 0.000	0.165	0.008	0.045	0.158	0.068	0.286	0.000	0.113	0.158	\$H	4 0.805
4.0UTEU \$H	3 0.146	0.500	0.354	\$H	2 0.250	0.750	\$H	10 0.010	0.104	0.042	0.031	0.250	0.042	0.104	0.000	0.198	0.219	\$H	4 0.375
4.PT \$H	3 0.080	0.598	0.321	\$H	2 0.036	0.964	\$H	10 0.000	0.107	0.054	0.161	0.321	0.036	0.045	0.009	0.098	0.170	\$H	4 0.116
5.EU \$H	3 0.246	0.351	0.404	\$H	2 0.509	0.491	\$H	10 0.018	0.228	0.018	0.053	0.175	0.000	0.140	0.000	0.070	0.298	\$H	4 0.649
5.NC \$H	3 0.188	0.348	0.464	\$H	2 0.406	0.594	\$H	10 0.007	0.130	0.101	0.029	0.138	0.014	0.116	0.007	0.080	0.377	\$H	4 0.703
5.0UTEU \$H	3 0.205	0.398	0.398	\$H	2 0.352	0.648	\$H	10 0.000	0.102	0.034	0.045	0.261	0.011	0.057	0.023	0.170	0.295	\$H	4 0.511
5.PT \$H	3 0.167	0.522	0.311	\$H	2 0.148	0.852	\$H	10 0.000	0.129	0.057	0.206	0.239	0.019	0.053	0.000	0.100	0.196	\$H	4 0.234
6.LU \$H	3 0.077	0.410	0.513	\$H	2 0.320	0.680	\$H	10 0.003	0.034	0.038	0.029	0.145	0.031	0.067	0.008	0.058	0.587	\$H	4 0.622
	2	3	4	SH HATLEVEL	1	2	3	SH EARNINGDEC	1	10	2	3	4	5	6	7	8	9	
1.EU	0.074	0.022	0.044	\$H	3 0.140	0.206	0.654	\$H	10 0.059	0.103	0.088	0.088	0.176	0.088	0.110	0.110	0.074	0.103	
1.NC	0.089	0.006	0.030	\$H	3 0.173	0.274	0.554	\$H	10 0.042	0.143	0.077	0.137	0.143	0.131	0.101	0.083	0.083	0.060	
1.0UTEU	0.056	0.056	0.117	\$H	3 0.216	0.352	0.432	\$H	10 0.117	0.062	0.086	0.136	0.142	0.142	0.093	0.031	0.068	0.123	
1.PT	0.159	0.159	0.250	\$H	3 0.591	0.045	0.364	\$H	10 0.182	0.068	0.318	0.091	0.091	0.068	0.091	0.045	0.045	0.000	
2.EU	0.078	0.023	0.047	\$H	3 0.125	0.297	0.578	\$H	10 0.031	0.172	0.094	0.055	0.062	0.109	0.102	0.070	0.148	0.156	
2.NC	0.103	0.009	0.034	\$H	3 0.222	0.179	0.598	\$H	10 0.034	0.111	0.060	0.111	0.085	0.051	0.137	0.179	0.111	0.120	
2.0UTEU	0.222	0.056	0.097	\$H	3 0.319	0.250	0.431	\$H	10 0.194	0.111	0.097	0.069	0.139	0.111	0.069	0.069	0.069	0.069	
2.PT	0.113	0.258	0.484	\$H	3 0.887	0.016	0.097	\$H	10 0.323	0.000	0.274	0.145	0.113	0.081	0.048	0.000	0.000	0.016	
3.EU	0.070	0.031	0.094	\$H	3 0.234	0.195	0.570	\$H	10 0.062	0.250	0.109	0.055	0.047	0.039	0.047	0.039	0.164	0.188	
3.NC	0.089	0.016	0.084	\$H	3 0.267	0.220	0.513	\$H	10 0.089	0.126	0.047	0.094	0.079	0.084	0.089	0.099	0.168	0.126	
3.0UTEU	0.129	0.089	0.198	\$H	3 0.416	0.238	0.347	\$H	10 0.149	0.099	0.188	0.099	0.059	0.040	0.099	0.069	0.109	0.089	
3.PT	0.144	0.267	0.400	\$H	3 0.889	0.033	0.078	\$H	10 0.278	0.011	0.233	0.244	0.100	0.033	0.022	0.011	0.033	0.033	
4.EU	0.209	0.030	0.060	\$H	3 0.388	0.209	0.403	\$H	10 0.075	0.343	0.149	0.164	0.030	0.030	0.030	0.030	0.060	0.090	
4.NC	0.165	0.008	0.023	\$H	3 0.248	0.301	0.451	\$H	10 0.098	0.158	0.030	0.030	0.098	0.120	0.083	0.068	0.083	0.098	
4.0UTEU	0.271	0.083	0.271	\$H	3 0.667	0.177	0.156	\$H	10 0.229	0.052	0.208	0.198	0.052	0.115	0.031	0.052	0.021	0.042	
4.PT	0.152	0.152	0.580	\$H	3 0.946	0.018	0.036	\$H	10 0.214	0.009	0.214	0.205	0.214	0.089	0.027	0.018	0.009	0.000	
5.EU	0.193	0.018	0.140	\$H	3 0.561	0.175	0.263	\$H	10 0.105	0.140	0.053	0.123	0.070	0.105	0.088	0.053	0.158	0.105	
5.NC	0.174	0.065	0.058	\$H	3 0.428	0.261	0.312	\$H	10 0.043	0.123	0.072	0.101	0.123	0.123	0.101	0.087	0.101	0.123	
5.0UTEU	0.250	0.045	0.193	\$H	3 0.636	0.182	0.182	\$H	10 0.136	0.091	0.148	0.102	0.091	0.148	0.068	0.091	0.057	0.068	
5.PT	0.201	0.172	0.392	\$H	3 0.871	0.081	0.048	\$H	10 0.163	0.024	0.177	0.153	0.167	0.115	0.086	0.067	0.029	0.019	
6.LU	0.235	0.065	0.078	\$H	3 0.623	0.200	0.177	\$H	10 0.076	0.069	0.079	0.091	0.088	0.112	0.108	0.123	0.114	0.140	

Figure C.16: Final symbolic data table - 2020

SH DEURBA		1	2	3	SH TELEARB	1	2	SH NACE3D	1	10	2	3	4	5	6	7	8	9	SH ISCO4D					
1.EU	SH	3	0.560	0.284	0.155	SH	2	0.750	0.250	SH	10	0.009	0.181	0.017	0.034	0.181	0.052	0.216	0.000	0.250	0.060	SH	4	0.871
1.NC	SH	3	0.362	0.376	0.262	SH	2	0.638	0.362	SH	10	0.000	0.094	0.020	0.027	0.154	0.060	0.221	0.007	0.215	0.201	SH	4	0.893
1.UTEU	SH	3	0.412	0.419	0.169	SH	2	0.551	0.449	SH	10	0.000	0.081	0.022	0.059	0.213	0.074	0.162	0.000	0.243	0.147	SH	4	0.787
1.PT	SH	3	0.245	0.367	0.388	SH	2	0.306	0.694	SH	10	0.000	0.061	0.061	0.143	0.286	0.061	0.082	0.000	0.184	0.122	SH	4	0.449
2.EU	SH	3	0.521	0.253	0.226	SH	2	0.719	0.281	SH	10	0.000	0.212	0.055	0.007	0.137	0.089	0.274	0.007	0.130	0.089	SH	4	0.897
2.NC	SH	3	0.293	0.374	0.333	SH	2	0.699	0.301	SH	10	0.000	0.122	0.041	0.033	0.089	0.098	0.260	0.016	0.203	0.138	SH	4	0.894
2.UTEU	SH	3	0.340	0.320	0.340	SH	2	0.515	0.485	SH	10	0.000	0.146	0.029	0.039	0.214	0.097	0.155	0.010	0.136	0.175	SH	4	0.718
2.PT	SH	3	0.127	0.563	0.310	SH	2	0.155	0.845	SH	10	0.000	0.028	0.070	0.239	0.239	0.014	0.085	0.000	0.183	0.141	SH	4	0.239
3.EU	SH	3	0.385	0.326	0.289	SH	2	0.793	0.207	SH	10	0.007	0.481	0.059	0.022	0.081	0.044	0.163	0.000	0.074	0.067	SH	4	0.889
3.NC	SH	3	0.365	0.328	0.307	SH	2	0.661	0.339	SH	10	0.005	0.116	0.042	0.021	0.090	0.058	0.238	0.021	0.265	0.143	SH	4	0.852
3.UTEU	SH	3	0.227	0.464	0.309	SH	2	0.418	0.582	SH	10	0.000	0.136	0.045	0.036	0.218	0.064	0.145	0.009	0.200	0.145	SH	4	0.564
3.PT	SH	3	0.086	0.617	0.297	SH	2	0.078	0.922	SH	10	0.008	0.094	0.047	0.250	0.219	0.031	0.016	0.008	0.148	0.180	SH	4	0.156
4.EU	SH	3	0.340	0.340	0.319	SH	2	0.713	0.287	SH	10	0.000	0.511	0.011	0.000	0.064	0.053	0.128	0.011	0.085	0.138	SH	4	0.851
4.NC	SH	3	0.284	0.324	0.392	SH	2	0.554	0.446	SH	10	0.000	0.203	0.027	0.041	0.122	0.034	0.223	0.014	0.142	0.196	SH	4	0.818
4.UTEU	SH	3	0.207	0.459	0.333	SH	2	0.297	0.703	SH	10	0.000	0.081	0.072	0.036	0.342	0.009	0.135	0.009	0.126	0.189	SH	4	0.459
4.PT	SH	3	0.103	0.619	0.278	SH	2	0.072	0.928	SH	10	0.010	0.155	0.041	0.289	0.227	0.021	0.021	0.000	0.134	0.103	SH	4	0.134
5.EU	SH	3	0.243	0.471	0.286	SH	2	0.514	0.486	SH	10	0.000	0.329	0.057	0.057	0.129	0.029	0.143	0.000	0.057	0.200	SH	4	0.729
5.NC	SH	3	0.174	0.349	0.477	SH	2	0.436	0.564	SH	10	0.007	0.094	0.034	0.034	0.161	0.020	0.208	0.027	0.081	0.336	SH	4	0.718
5.UTEU	SH	3	0.170	0.511	0.319	SH	2	0.266	0.734	SH	10	0.000	0.138	0.021	0.085	0.181	0.043	0.096	0.000	0.106	0.330	SH	4	0.457
5.PT	SH	3	0.080	0.500	0.420	SH	2	0.125	0.875	SH	10	0.000	0.062	0.068	0.188	0.222	0.023	0.045	0.000	0.085	0.307	SH	4	0.278
6.LU	SH	3	0.083	0.396	0.521	SH	2	0.322	0.678	SH	10	0.004	0.047	0.045	0.024	0.134	0.033	0.075	0.004	0.064	0.571	SH	4	0.669
SH HATLEVEL		1	2	3	SH EARNINGDEC	1	10	2	3	4	5	6	7	8	9									
1.EU	SH	3	0.052	0.043	0.034	SH	3	0.121	0.198	0.681	SH	10	0.034	0.069	0.086	0.103	0.095	0.164	0.172	0.147	0.069	0.060		
1.NC	SH	3	0.067	0.007	0.034	SH	3	0.154	0.215	0.631	SH	10	0.060	0.107	0.054	0.121	0.121	0.154	0.174	0.087	0.087	0.034		
1.UTEU	SH	3	0.059	0.044	0.110	SH	3	0.169	0.279	0.551	SH	10	0.088	0.110	0.147	0.074	0.103	0.096	0.154	0.110	0.059	0.059		
1.PT	SH	3	0.143	0.102	0.306	SH	3	0.469	0.102	0.429	SH	10	0.143	0.020	0.245	0.204	0.122	0.061	0.102	0.061	0.000	0.041		
2.EU	SH	3	0.068	0.000	0.034	SH	3	0.151	0.171	0.678	SH	10	0.034	0.082	0.055	0.041	0.130	0.130	0.123	0.123	0.178	0.103		
2.NC	SH	3	0.073	0.016	0.016	SH	3	0.179	0.203	0.618	SH	10	0.041	0.073	0.073	0.073	0.073	0.089	0.130	0.122	0.138	0.187		
2.UTEU	SH	3	0.107	0.039	0.136	SH	3	0.282	0.165	0.553	SH	10	0.155	0.146	0.126	0.068	0.068	0.058	0.136	0.058	0.078	0.107		
2.PT	SH	3	0.155	0.225	0.380	SH	3	0.718	0.169	0.113	SH	10	0.141	0.028	0.366	0.282	0.113	0.014	0.028	0.014	0.000	0.014		
3.EU	SH	3	0.081	0.000	0.030	SH	3	0.141	0.185	0.674	SH	10	0.089	0.200	0.037	0.052	0.052	0.089	0.104	0.126	0.119	0.133		
3.NC	SH	3	0.079	0.037	0.032	SH	3	0.190	0.212	0.598	SH	10	0.069	0.132	0.032	0.095	0.079	0.058	0.058	0.122	0.175	0.180		
3.UTEU	SH	3	0.155	0.055	0.227	SH	3	0.409	0.200	0.391	SH	10	0.091	0.109	0.255	0.136	0.118	0.073	0.064	0.009	0.064	0.082		
3.PT	SH	3	0.180	0.203	0.461	SH	3	0.906	0.047	0.047	SH	10	0.164	0.008	0.344	0.195	0.141	0.062	0.016	0.016	0.023	0.031		
4.EU	SH	3	0.096	0.021	0.032	SH	3	0.213	0.245	0.543	SH	10	0.032	0.277	0.064	0.064	0.064	0.053	0.106	0.085	0.096	0.160		
4.NC	SH	3	0.101	0.047	0.034	SH	3	0.277	0.324	0.399	SH	10	0.088	0.155	0.074	0.108	0.068	0.081	0.034	0.101	0.142	0.149		
4.UTEU	SH	3	0.225	0.081	0.234	SH	3	0.586	0.207	0.207	SH	10	0.171	0.090	0.198	0.090	0.126	0.072	0.072	0.081	0.045	0.054		
4.PT	SH	3	0.186	0.216	0.464	SH	3	0.938	0.031	0.031	SH	10	0.206	0.021	0.289	0.165	0.165	0.103	0.021	0.010	0.010	0.010		
5.EU	SH	3	0.171	0.029	0.071	SH	3	0.414	0.200	0.386	SH	10	0.086	0.200	0.114	0.043	0.057	0.057	0.057	0.157	0.071	0.157		
5.NC	SH	3	0.221	0.007	0.054	SH	3	0.423	0.282	0.295	SH	10	0.107	0.101	0.027	0.087	0.101	0.154	0.114	0.114	0.094	0.101		
5.UTEU	SH	3	0.213	0.085	0.245	SH	3	0.617	0.223	0.160	SH	10	0.096	0.106	0.149	0.149	0.160	0.117	0.032	0.064	0.085	0.043		
5.PT	SH	3	0.199	0.199	0.324	SH	3	0.881	0.085	0.034	SH	10	0.108	0.017	0.205	0.142	0.170	0.188	0.080	0.051	0.028	0.011		
6.LU	SH	3	0.217	0.058	0.056	SH	3	0.578	0.219	0.203	SH	10	0.105	0.098	0.073	0.087	0.091	0.106	0.106	0.110	0.113	0.111		

Figure C.17: Final symbolic data table - 2021

	SH DEGURBA				3 SH TELEARB				2 SH NACE3D				3				5				7				9 SH ISCO4D					
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4		
1.EU \$H	0.553	0.250	0.197	\$H	2	0.671	0.329	\$H	10	0.000	0.224	0.026	0.000	0.224	0.053	0.211	0.013	0.184	0.066	\$H	4	0.855								
1.NC \$H	0.360	0.297	0.342	\$H	2	0.559	0.441	\$H	10	0.000	0.054	0.018	0.018	0.162	0.072	0.279	0.000	0.216	0.180	\$H	4	0.865								
1.UTEU \$H	0.505	0.351	0.144	\$H	2	0.629	0.371	\$H	10	0.000	0.093	0.021	0.010	0.237	0.113	0.227	0.000	0.247	0.052	\$H	4	0.814								
1.PT \$H	0.258	0.548	0.194	\$H	2	0.194	0.806	\$H	10	0.000	0.065	0.032	0.129	0.258	0.065	0.000	0.032	0.194	0.226	\$H	4	0.323								
2.EU \$H	0.582	0.227	0.191	\$H	2	0.755	0.245	\$H	10	0.000	0.227	0.027	0.000	0.200	0.109	0.245	0.000	0.145	0.045	\$H	4	0.882								
2.NC \$H	0.403	0.371	0.226	\$H	2	0.556	0.444	\$H	10	0.000	0.137	0.040	0.024	0.129	0.048	0.250	0.024	0.161	0.185	\$H	4	0.919								
2.UTEU \$H	0.325	0.423	0.252	\$H	2	0.447	0.553	\$H	10	0.000	0.098	0.033	0.016	0.236	0.114	0.163	0.000	0.220	0.122	\$H	4	0.683								
2.PT \$H	0.188	0.562	0.250	\$H	2	0.250	0.750	\$H	10	0.000	0.062	0.031	0.219	0.156	0.062	0.062	0.000	0.281	0.125	\$H	4	0.375								
3.EU \$H	0.430	0.289	0.281	\$H	2	0.659	0.341	\$H	10	0.000	0.496	0.037	0.037	0.104	0.074	0.156	0.000	0.074	0.022	\$H	4	0.881								
3.NC \$H	0.297	0.338	0.366	\$H	2	0.538	0.462	\$H	10	0.000	0.159	0.048	0.021	0.124	0.076	0.269	0.014	0.117	0.172	\$H	4	0.848								
3.UTEU \$H	0.194	0.510	0.296	\$H	2	0.276	0.724	\$H	10	0.000	0.173	0.020	0.082	0.184	0.051	0.143	0.010	0.122	0.214	\$H	4	0.531								
3.PT \$H	0.028	0.697	0.275	\$H	2	0.073	0.927	\$H	10	0.000	0.101	0.083	0.248	0.193	0.028	0.018	0.000	0.156	0.174	\$H	4	0.193								
4.EU \$H	0.382	0.324	0.294	\$H	2	0.647	0.353	\$H	10	0.000	0.515	0.000	0.000	0.103	0.029	0.176	0.000	0.088	0.088	\$H	4	0.779								
4.NC \$H	0.333	0.300	0.367	\$H	2	0.575	0.425	\$H	10	0.000	0.192	0.033	0.025	0.133	0.075	0.242	0.000	0.117	0.183	\$H	4	0.825								
4.UTEU \$H	0.143	0.520	0.337	\$H	2	0.245	0.755	\$H	10	0.010	0.122	0.031	0.102	0.133	0.020	0.143	0.000	0.153	0.286	\$H	4	0.500								
4.PT \$H	0.047	0.581	0.372	\$H	2	0.047	0.953	\$H	10	0.000	0.081	0.035	0.233	0.314	0.000	0.023	0.000	0.186	0.128	\$H	4	0.186								
5.EU \$H	0.340	0.298	0.362	\$H	2	0.532	0.468	\$H	10	0.000	0.340	0.043	0.000	0.149	0.106	0.106	0.000	0.043	0.213	\$H	4	0.787								
5.NC \$H	0.106	0.440	0.454	\$H	2	0.333	0.667	\$H	10	0.000	0.113	0.071	0.035	0.156	0.035	0.170	0.021	0.078	0.319	\$H	4	0.716								
5.UTEU \$H	0.192	0.449	0.359	\$H	2	0.231	0.769	\$H	10	0.000	0.090	0.090	0.051	0.269	0.051	0.077	0.000	0.103	0.269	\$H	4	0.462								
5.PT \$H	0.066	0.497	0.437	\$H	2	0.093	0.907	\$H	10	0.013	0.113	0.053	0.192	0.219	0.040	0.033	0.007	0.079	0.252	\$H	4	0.258								
6.LU \$H	0.084	0.398	0.518	\$H	2	0.237	0.763	\$H	10	0.006	0.040	0.053	0.025	0.128	0.042	0.066	0.005	0.052	0.583	\$H	4	0.656								
1 2 3 4 5 6 7 8 9																														
1.EU	0.066	0.026	0.053	\$H	3	0.118	0.211	0.671	\$H	10	0.066	0.066	0.079	0.079	0.079	0.158	0.132	0.145	0.118	0.079										
1.NC	0.108	0.027	0.000	\$H	3	0.171	0.198	0.631	\$H	10	0.045	0.072	0.063	0.090	0.171	0.126	0.090	0.117	0.144	0.081										
1.UTEU	0.103	0.000	0.082	\$H	3	0.165	0.237	0.598	\$H	10	0.072	0.072	0.113	0.072	0.082	0.082	0.186	0.144	0.113	0.062										
1.PT	0.161	0.194	0.323	\$H	3	0.581	0.129	0.290	\$H	10	0.161	0.065	0.258	0.355	0.097	0.032	0.000	0.032	0.000	0.000										
2.EU	0.091	0.000	0.027	\$H	3	0.173	0.227	0.600	\$H	10	0.036	0.073	0.036	0.036	0.055	0.100	0.127	0.209	0.182	0.145										
2.NC	0.048	0.008	0.024	\$H	3	0.129	0.202	0.669	\$H	10	0.032	0.097	0.089	0.048	0.089	0.097	0.121	0.137	0.145	0.145										
2.UTEU	0.146	0.033	0.138	\$H	3	0.228	0.236	0.537	\$H	10	0.163	0.073	0.122	0.008	0.081	0.106	0.114	0.122	0.106	0.106										
2.PT	0.094	0.094	0.438	\$H	3	0.750	0.062	0.188	\$H	10	0.125	0.031	0.250	0.281	0.094	0.094	0.000	0.031	0.031	0.062										
3.EU	0.059	0.007	0.052	\$H	3	0.081	0.215	0.704	\$H	10	0.059	0.104	0.037	0.089	0.067	0.052	0.111	0.089	0.215	0.178										
3.NC	0.083	0.034	0.034	\$H	3	0.172	0.200	0.628	\$H	10	0.062	0.103	0.041	0.097	0.062	0.062	0.076	0.110	0.159	0.228										
3.UTEU	0.163	0.061	0.245	\$H	3	0.520	0.122	0.357	\$H	10	0.224	0.092	0.143	0.122	0.082	0.041	0.082	0.071	0.071	0.071										
3.PT	0.147	0.183	0.477	\$H	3	0.844	0.092	0.064	\$H	10	0.257	0.009	0.294	0.156	0.183	0.055	0.009	0.018	0.009	0.009										
4.EU	0.132	0.029	0.059	\$H	3	0.221	0.235	0.544	\$H	10	0.074	0.221	0.074	0.074	0.074	0.074	0.059	0.074	0.206	0.074										
4.NC	0.092	0.017	0.067	\$H	3	0.200	0.292	0.508	\$H	10	0.050	0.158	0.058	0.092	0.083	0.050	0.083	0.108	0.125	0.192										
4.UTEU	0.102	0.122	0.276	\$H	3	0.480	0.276	0.245	\$H	10	0.214	0.082	0.204	0.082	0.112	0.092	0.051	0.061	0.041	0.061										
4.PT	0.256	0.151	0.407	\$H	3	0.872	0.116	0.012	\$H	10	0.198	0.000	0.233	0.302	0.140	0.081	0.023	0.000	0.012	0.012										
5.EU	0.106	0.021	0.085	\$H	3	0.511	0.213	0.277	\$H	10	0.149	0.128	0.043	0.064	0.043	0.064	0.085	0.106	0.128	0.191										
5.NC	0.170	0.035	0.078	\$H	3	0.475	0.191	0.333	\$H	10	0.085	0.050	0.050	0.099	0.085	0.085	0.128	0.142	0.071	0.170										
5.UTEU	0.244	0.103	0.192	\$H	3	0.667	0.179	0.154	\$H	10	0.090	0.128	0.141	0.115	0.154	0.141	0.103	0.026	0.064	0.038										
5.PT	0.185	0.185	0.371	\$H	3	0.881	0.079	0.040	\$H	10	0.199	0.026	0.152	0.199	0.159	0.099	0.060	0.040	0.040	0.026										
6.LU	0.211	0.068	0.066	\$H	3	0.564	0.235	0.201	\$H	10	0.094	0.089	0.076	0.093	0.096	0.115	0.126	0.093	0.114	0.104										

Figure C.18: Final symbolic data table - 2022

D

		Cluster			
		1	2	3	4
TELEARB	Yes	0.189	0.160	0.163	0.052
	No	0.811	0.841	0.837	0.948
ISCO4D	White-collar high-skilled professions	0.858	0.628	0.898	0.168
	White-collar low-skilled professions	0.080	0.205	0.041	0.241
	Blue-collar high-skilled professions	0.017	0.039	0.020	0.080
	Blue-collar low-skilled professions	0.045	0.128	0.041	0.511
HATLEVEL	High School	0.174	0.503	0.204	0.849
	Bachelor	0.324	0.286	0.367	0.084
	Master/PhD	0.502	0.210	0.429	0.067
NACE3D	Agriculture	0.000	0.000	0.000	0.000
	Industry	0.043	0.052	0.000	0.097
	Construction	0.011	0.025	0.020	0.075
	Trade, transport and hospitality	0.125	0.164	0.041	0.254
	Information and communication	0.055	0.036	0.000	0.045
	Financial and insurance activities	0.227	0.169	0.286	0.023
	Real state activities	0.003	0.003	0.000	0.000
	Business services	0.136	0.097	0.020	0.151
	Public administration,... ¹	0.149	0.313	0.163	0.139
	Arts, entertainment and recreation,... ²	0.252	0.139	0.469	0.217
DEGURBA	Cities	0.470	0.251	0.347	0.151
	Towns and Suburbs	0.205	0.335	0.163	0.367
	Rural areas	0.325	0.414	0.490	0.481
EARNINGDEC	1	0.074	0.150	0.020	0.305
	2	0.076	0.133	0.000	0.276
	3	0.092	0.089	0.082	0.158
	4	0.158	0.108	0.122	0.132
	5	0.049	0.040	0.000	0.020
	6	0.121	0.097	0.082	0.045
	7	0.114	0.103	0.041	0.008
	8	0.166	0.116	0.184	0.042
	9	0.078	0.059	0.041	0.007
	10	0.073	0.106	0.429	0.007

Table D.1: Average weight of each category of a variable for the objects belonging to a cluster - 2014

¹Public administration, defence, education, health and social work

²Arts, entertainment and recreation and other service activities

		Cluster		
		1	2	3
TELEARB	Yes	0.216	0.172	0.109
	No	0.784	0.828	0.891
ISCO4D	White-collar high-skilled professions	0.813	0.571	0.142
	White-collar low-skilled professions	0.099	0.201	0.182
	Blue-collar high-skilled professions	0.028	0.079	0.213
	Blue-collar low-skilled professions	0.059	0.204	0.463
HATLEVEL	High School	0.245	0.588	0.890
	Bachelor	0.338	0.211	0.050
	Master/PhD	0.418	0.200	0.060
NACE3D	Agriculture	0.030	0.006	0.001
	Industry	0.045	0.052	0.068
	Construction	0.023	0.081	0.241
	Trade, transport and hospitality	0.139	0.221	0.240
	Information and communication	0.063	0.034	0.021
	Financial and insurance activities	0.207	0.120	0.031
	Real state activities	0.004	0.002	0.002
	Business services	0.138	0.124	0.148
	Public administration,...	0.129	0.265	0.131
	Arts, entertainment and recreation,...	0.248	0.094	0.117
DEGURBA	Cities	0.432	0.259	0.158
	Towns and Suburbs	0.249	0.340	0.472
	Rural Areas	0.319	0.400	0.370
EARNINGDEC	1	0.044	0.126	0.211
	2	0.065	0.137	0.221
	3	0.070	0.117	0.196
	4	0.087	0.116	0.171
	5	0.123	0.122	0.083
	6	0.067	0.054	0.040
	7	0.124	0.087	0.037
	8	0.083	0.062	0.010
	9	0.129	0.071	0.015
	10	0.206	0.109	0.016

Table D.2: Average weight of each category of a variable for the objects belonging to a cluster - 2015

		Cluster		
		1	2	3
TELEARB	Yes	0.248	0.172	0.114
	No	0.752	0.828	0.886
ISCO4D	White-collar high-skilled professions	0.778	0.500	0.134
	White-collar low-skilled professions	0.134	0.232	0.173
	Blue-collar high-skilled professions	0.029	0.057	0.205
	Blue-collar low-skilled professions	0.060	0.211	0.488
HATLEVEL	High School	0.266	0.551	0.887
	Bachelor	0.320	0.203	0.056
	Master/PhD	0.413	0.246	0.057
NACE3D	Agriculture	0.003	0.007	0.004
	Industry	0.030	0.051	0.078
	Construction	0.042	0.052	0.261
	Trade, transport and hospitality	0.146	0.229	0.258
	Information and communication	0.043	0.036	0.014
	Financial and insurance activities	0.232	0.139	0.018
	Real state activities	0.006	0.007	0.001
	Business services	0.165	0.133	0.157
	Public administration,...	0.121	0.271	0.095
	Arts, entertainment and recreation,...	0.212	0.077	0.115
DEGURBA	Cities	0.269	0.158	0.085
	Towns and Suburbs	0.277	0.365	0.450
	Rural Areas	0.454	0.477	0.464
EARNINGDEC	1	0.058	0.141	0.196
	2	0.078	0.135	0.209
	3	0.078	0.125	0.178
	4	0.094	0.110	0.182
	5	0.061	0.065	0.110
	6	0.072	0.089	0.056
	7	0.114	0.089	0.030
	8	0.134	0.060	0.017
	9	0.109	0.086	0.016
	10	0.203	0.099	0.006

Table D.3: Average weight of each category of a variable for the objects belonging to a cluster - 2016

		Cluster		
		1	2	3
TELEARB	Yes	0.297	0.185	0.105
	No	0.703	0.815	0.895
ISCO4D	White-collar high-skilled professions	0.828	0.581	0.114
	White-collar low-skilled professions	0.106	0.184	0.171
	Blue-collar high-skilled professions	0.021	0.050	0.226
	Blue-collar low-skilled professions	0.045	0.184	0.489
HATLEVEL	High School	0.194	0.511	0.894
	Bachelor	0.216	0.170	0.021
	Master/PhD	0.591	0.319	0.085
NACE3D	Agriculture	0.000	0.002	0.000
	Industry	0.030	0.042	0.081
	Construction	0.018	0.035	0.271
	Trade, transport and hospitality	0.143	0.200	0.211
	Information and communication	0.057	0.050	0.006
	Financial and insurance activities	0.219	0.153	0.030
	Real state activities	0.010	0.008	0.005
	Business services	0.183	0.136	0.142
	Public administration,...	0.118	0.267	0.128
	Arts, entertainment and recreation,...	0.223	0.106	0.127
DEGURBA	Cities	0.424	0.221	0.118
	Towns and Suburbs	0.291	0.407	0.538
	Rural Areas	0.285	0.372	0.344
EARNINGDEC	1	0.054	0.124	0.204
	2	0.079	0.133	0.313
	3	0.059	0.063	0.142
	4	0.088	0.112	0.138
	5	0.101	0.079	0.099
	6	0.105	0.097	0.038
	7	0.120	0.094	0.030
	8	0.140	0.075	0.010
	9	0.084	0.080	0.015
	10	0.170	0.143	0.012

Table D.4: Average weight of each category of a variable for the objects belonging to a cluster - 2017

		Cluster		
		1	2	3
TELEARB	Yes	0.291	0.228	0.127
	No	0.709	0.772	0.873
ISCO4D	White-collar high-skilled professions	0.845	0.725	0.254
	White-collar low-skilled professions	0.075	0.155	0.216
	Blue-collar high-skilled professions	0.021	0.040	0.160
	Blue-collar low-skilled professions	0.060	0.081	0.369
HATLEVEL	High School	0.195	0.368	0.776
	Bachelor	0.283	0.259	0.100
	Master/PhD	0.522	0.373	0.124
NACE3D	Agriculture	0.004	0.002	0.004
	Industry	0.020	0.064	0.067
	Construction	0.016	0.033	0.176
	Trade, transport and hospitality	0.122	0.136	0.244
	Information and communication	0.060	0.066	0.023
	Financial and insurance activities	0.197	0.196	0.078
	Real state activities	0.000	0.009	0.012
	Business services	0.136	0.128	0.156
	Public administration,...	0.062	0.226	0.148
	Arts, entertainment and recreation,...	0.384	0.141	0.093
DEGURBA	Cities	0.504	0.304	0.176
	Towns and Suburbs	0.258	0.359	0.526
	Rural Areas	0.239	0.337	0.297
EARNINGDEC	1	0.047	0.068	0.165
	2	0.060	0.083	0.256
	3	0.039	0.076	0.125
	4	0.088	0.108	0.162
	5	0.048	0.109	0.080
	6	0.085	0.095	0.056
	7	0.128	0.116	0.054
	8	0.129	0.135	0.035
	9	0.130	0.095	0.026
	10	0.246	0.116	0.039

Table D.5: Average weight of each category of a variable for the objects belonging to a cluster - 2018

		Cluster			
		1	2	3	4
TELEARB	Yes	0.287	0.238	0.152	0.148
	No	0.713	0.762	0.848	0.852
ISCO4D	White-collar high-skilled professions	0.842	0.767	0.602	0.293
	White-collar low-skilled professions	0.095	0.119	0.221	0.199
	Blue-collar high-skilled professions	0.029	0.027	0.044	0.152
	Blue-collar low-skilled professions	0.034	0.087	0.133	0.356
HATLEVEL	High School	0.190	0.253	0.524	0.719
	Bachelor	0.275	0.274	0.240	0.122
	Master/PhD	0.536	0.473	0.236	0.158
NACE3D	Agriculture	0.007	0.000	0.000	0.002
	Industry	0.005	0.037	0.063	0.058
	Construction	0.014	0.030	0.049	0.163
	Trade, transport and hospitality	0.124	0.176	0.142	0.243
	Information and communication	0.064	0.058	0.027	0.028
	Financial and insurance activities	0.166	0.232	0.105	0.074
	Real state activities	0.005	0.011	0.001	0.003
	Business services	0.145	0.201	0.083	0.153
	Public administration,...	0.089	0.142	0.375	0.166
	Arts, entertainment and recreation,...	0.382	0.114	0.155	0.110
DEGURBA	Cities	0.474	0.286	0.146	0.164
	Towns and Suburbs	0.270	0.350	0.410	0.521
	Rural Areas	0.256	0.365	0.444	0.315
EARNINGDEC	1	0.061	0.098	0.080	0.168
	2	0.080	0.098	0.084	0.260
	3	0.076	0.079	0.078	0.173
	4	0.083	0.091	0.093	0.127
	5	0.058	0.119	0.073	0.081
	6	0.118	0.096	0.104	0.048
	7	0.081	0.102	0.118	0.040
	8	0.131	0.095	0.104	0.037
	9	0.108	0.092	0.111	0.032
	10	0.204	0.131	0.154	0.036

Table D.6: Average weight of each category of a variable for the objects belonging to a cluster - 2019

		Cluster			
		1	2	3	4
TELEARB	Yes	0.569	0.516	0.332	0.155
	No	0.431	0.484	0.668	0.845
ISCO4D	White-collar high-skilled professions	0.807	0.685	0.553	0.223
	White-collar low-skilled professions	0.110	0.150	0.232	0.154
	Blue-collar high-skilled professions	0.024	0.042	0.064	0.202
	Blue-collar low-skilled professions	0.060	0.123	0.150	0.421
HATLEVEL	High School	0.214	0.400	0.589	0.837
	Bachelor	0.260	0.204	0.205	0.039
	Master/PhD	0.526	0.396	0.207	0.125
NACE3D	Agriculture	0.003	0.006	0.005	0.000
	Industry	0.044	0.037	0.054	0.045
	Construction	0.033	0.048	0.034	0.214
	Trade, transport and hospitality	0.183	0.143	0.198	0.269
	Information and communication	0.061	0.023	0.024	0.020
	Financial and insurance activities	0.216	0.178	0.086	0.063
	Real state activities	0.004	0.006	0.010	0.006
	Business services	0.190	0.093	0.126	0.141
	Public administration,...	0.134	0.172	0.370	0.115
	Arts, entertainment and recreation,...	0.131	0.294	0.092	0.126
DEGURBA	Cities	0.435	0.344	0.154	0.134
	Towns and Suburbs	0.299	0.336	0.414	0.557
	Rural Areas	0.267	0.320	0.432	0.309
EARNINGDEC	1	0.083	0.098	0.121	0.232
	2	0.072	0.125	0.127	0.243
	3	0.098	0.110	0.123	0.168
	4	0.118	0.052	0.088	0.137
	5	0.100	0.054	0.124	0.077
	6	0.096	0.066	0.077	0.055
	7	0.090	0.048	0.088	0.028
	8	0.102	0.123	0.073	0.023
	9	0.115	0.118	0.093	0.014
	10	0.123	0.208	0.084	0.022

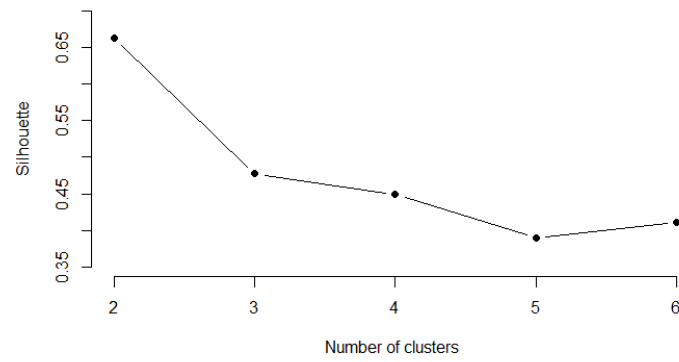
Table D.7: Average weight of each category of a variable for the objects belonging to a cluster - 2020

		Cluster					
		1	2	3	4	5	6
TELEARB	Yes	0.693	0.643	0.379	0.495	0.248	0.102
	No	0.307	0.356	0.621	0.505	0.751	0.898
ISCO4D	White-collar high-skilled professions	0.881	0.822	0.694	0.690	0.411	0.176
	White-collar low-skilled professions	0.068	0.112	0.219	0.107	0.195	0.174
	Blue-collar high-skilled professions	0.021	0.024	0.032	0.046	0.117	0.215
	Blue-collar low-skilled professions	0.030	0.042	0.055	0.158	0.277	0.435
HATLEVEL	High School	0.159	0.261	0.500	0.287	0.638	0.854
	Bachelor	0.200	0.238	0.250	0.215	0.154	0.082
	Master/PhD	0.641	0.501	0.249	0.498	0.208	0.064
NACE3D	Agriculture	0.003	0.002	0.005	0.000	0.000	0.006
	Industry	0.035	0.038	0.040	0.032	0.056	0.053
	Construction	0.024	0.030	0.029	0.045	0.113	0.259
	Trade, transport and hospitality	0.130	0.099	0.148	0.215	0.258	0.228
	Information and communication	0.071	0.040	0.027	0.078	0.034	0.022
	Financial and insurance activities	0.242	0.164	0.142	0.154	0.090	0.041
	Real state activities	0.010	0.006	0.016	0.006	0.002	0.003
	Business services	0.213	0.090	0.073	0.193	0.125	0.155
	Public administration,...	0.126	0.150	0.454	0.156	0.237	0.141
	Arts, entertainment and recreation,...	0.145	0.381	0.071	0.124	0.086	0.092
DEGURBA	Cities	0.420	0.313	0.128	0.326	0.176	0.105
	Towns and Suburbs	0.323	0.365	0.372	0.401	0.459	0.600
	Rural Areas	0.257	0.322	0.499	0.273	0.365	0.295
EARNINGDEC	1	0.048	0.074	0.106	0.111	0.130	0.170
	2	0.060	0.072	0.050	0.176	0.199	0.330
	3	0.087	0.067	0.087	0.093	0.146	0.214
	4	0.100	0.060	0.096	0.096	0.145	0.140
	5	0.119	0.070	0.130	0.076	0.110	0.060
	6	0.131	0.075	0.110	0.118	0.072	0.022
	7	0.120	0.117	0.112	0.059	0.064	0.013
	8	0.129	0.107	0.104	0.067	0.040	0.011
	9	0.113	0.150	0.106	0.083	0.037	0.018
	10	0.093	0.208	0.100	0.122	0.058	0.019

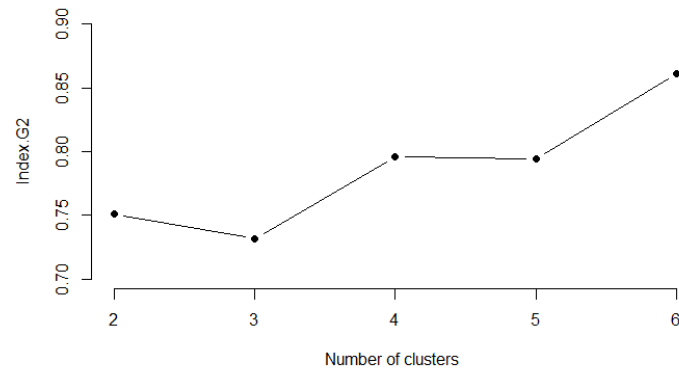
Table D.8: Average weight of each category of a variable for the objects belonging to a cluster - 2021

		Cluster				
		1	2	3	4	5
TELEARB	Yes	0.612	0.653	0.490	0.264	0.131
	No	0.388	0.347	0.511	0.736	0.869
ISCO4D	White-collar high-skilled professions	0.858	0.830	0.735	0.573	0.267
	White-collar low-skilled professions	0.084	0.096	0.126	0.178	0.169
	Blue-collar high-skilled professions	0.016	0.018	0.027	0.078	0.161
	Blue-collar low-skilled professions	0.041	0.055	0.112	0.171	0.403
HATLEVEL	High School	0.161	0.151	0.370	0.541	0.786
	Bachelor	0.224	0.225	0.224	0.201	0.096
	Master/PhD	0.615	0.624	0.407	0.258	0.119
NACE3D	Agriculture	0.000	0.000	0.000	0.003	0.003
	Industry	0.03	0.018	0.038	0.053	0.047
	Construction	0.014	0.018	0.008	0.059	0.204
	Trade, transport and hospitality	0.173	0.104	0.192	0.174	0.228
	Information and communication	0.078	0.052	0.110	0.040	0.039
	Financial and insurance activities	0.246	0.166	0.134	0.120	0.027
	Real state activities	0.007	0.000	0.000	0.007	0.008
	Business services	0.170	0.081	0.132	0.102	0.179
	Public administration,...	0.126	0.055	0.167	0.334	0.181
	Arts, entertainment and recreation,...	0.155	0.506	0.219	0.108	0.084
DEGURBA	Cities	0.433	0.406	0.332	0.144	0.117
	Towns and Suburbs	0.305	0.306	0.36	0.463	0.577
	Rural Areas	0.262	0.288	0.307	0.393	0.306
EARNINGDEC	1	0.052	0.066	0.156	0.141	0.188
	2	0.068	0.055	0.082	0.130	0.237
	3	0.073	0.081	0.036	0.102	0.259
	4	0.089	0.071	0.062	0.106	0.135
	5	0.096	0.063	0.085	0.095	0.072
	6	0.116	0.085	0.100	0.098	0.018
	7	0.139	0.081	0.114	0.079	0.024
	8	0.141	0.210	0.117	0.072	0.018
	9	0.133	0.126	0.148	0.089	0.022
	10	0.092	0.162	0.100	0.088	0.026

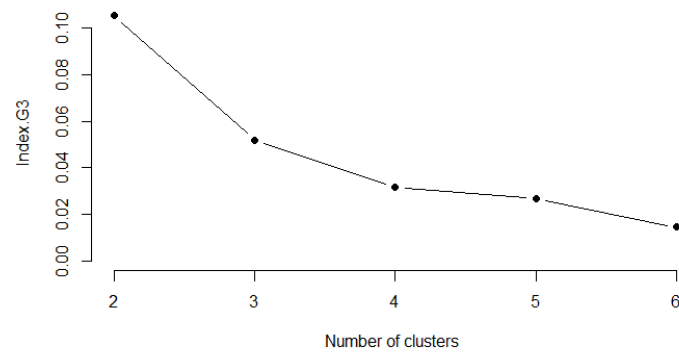
Table D.9: Average weight of each category of a variable for the objects belonging to a cluster - 2022



(a) Silhouette

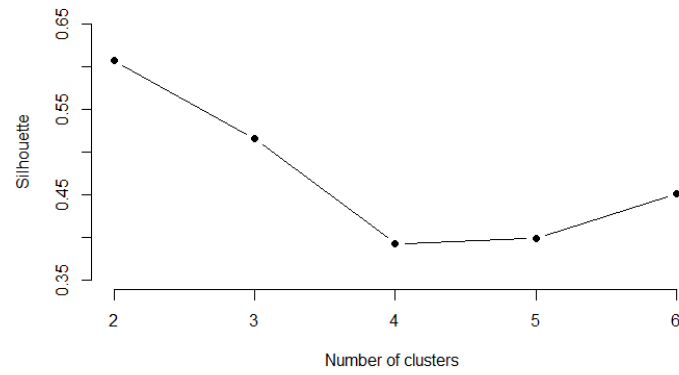


(b) Index G2

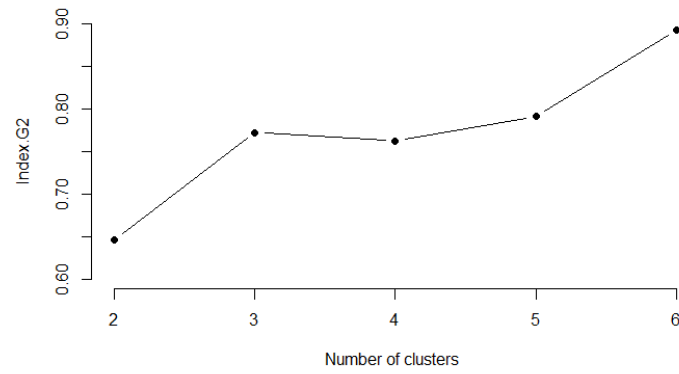


(c) Index G3

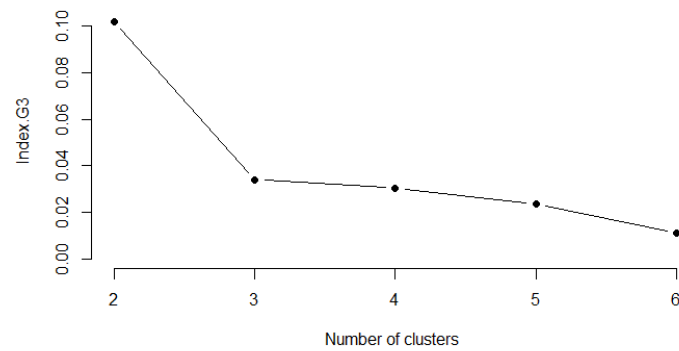
Figure D.1: Validity Indices - 2014



(a) Silhouette

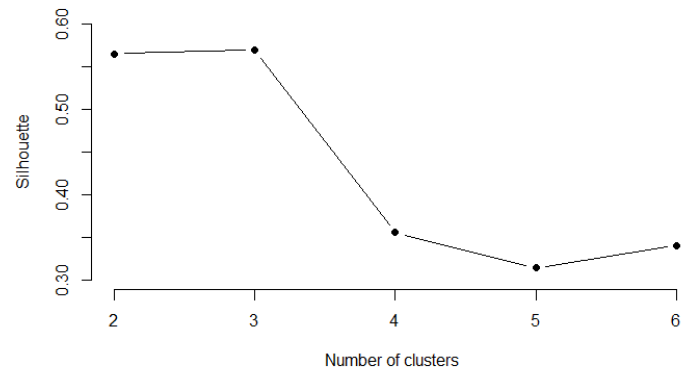


(b) Index G2

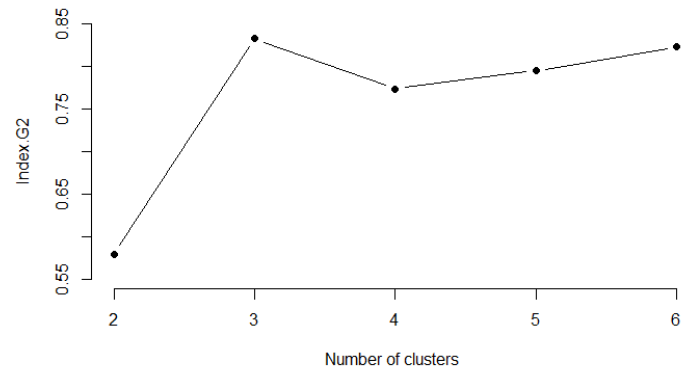


(c) Index G3

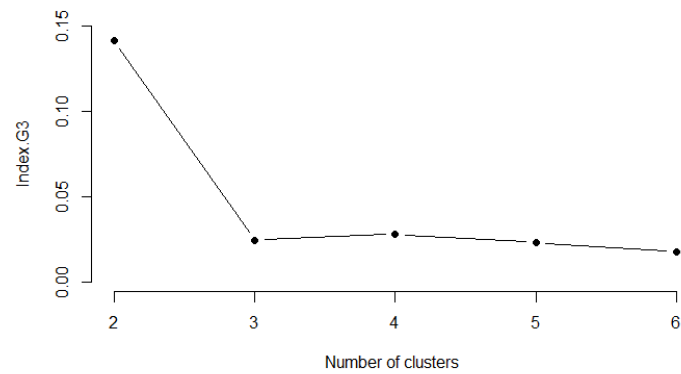
Figure D.2: Validity Indices - 2015



(a) Silhouette

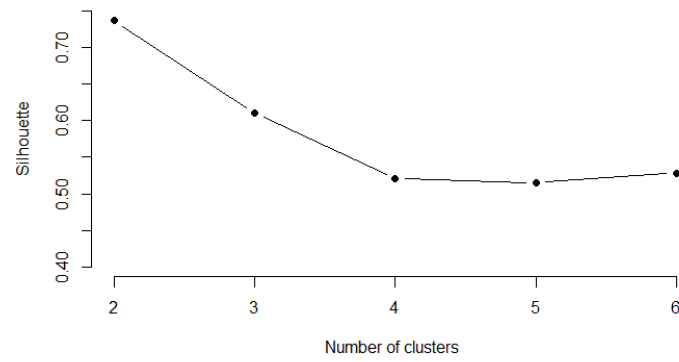


(b) Index G2

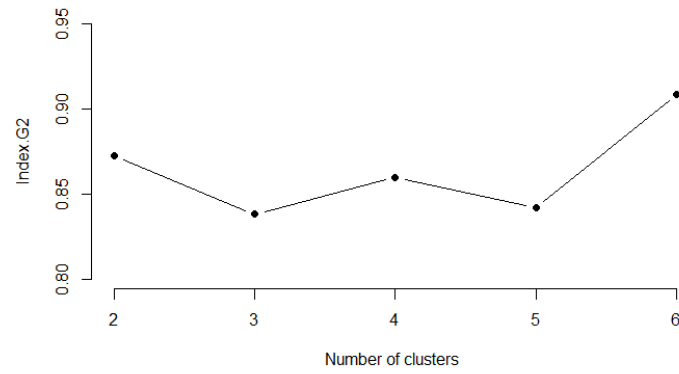


(c) Index G3

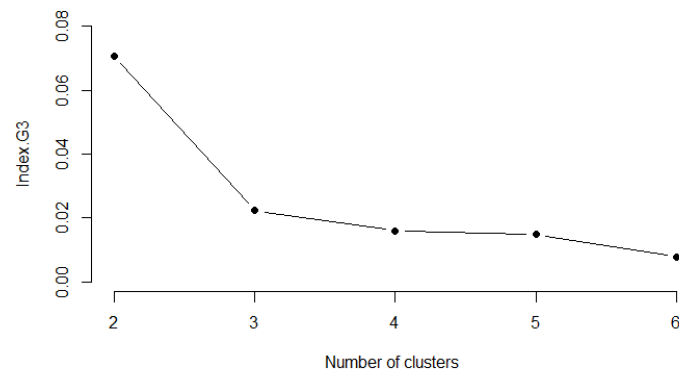
Figure D.3: Validity Indices - 2016



(a) Silhouette

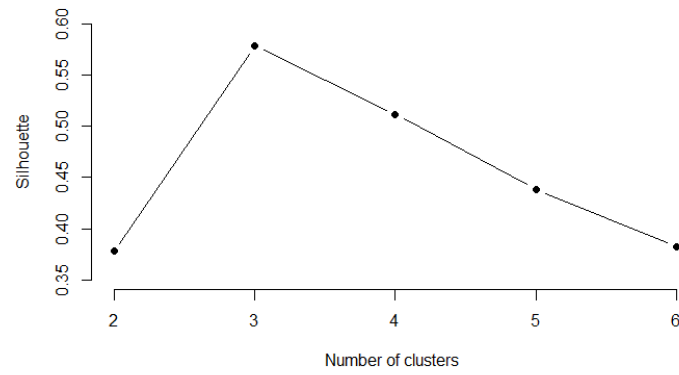


(b) Index G2

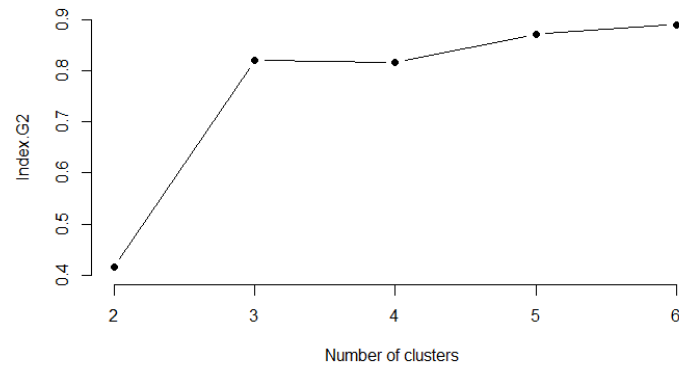


(c) Index G3

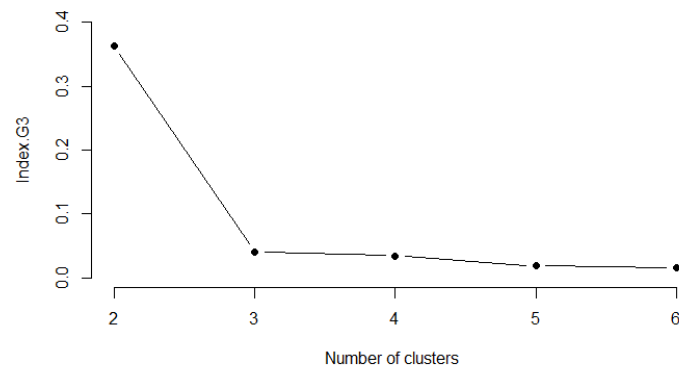
Figure D.4: Validity Indices - 2017



(a) Silhouette

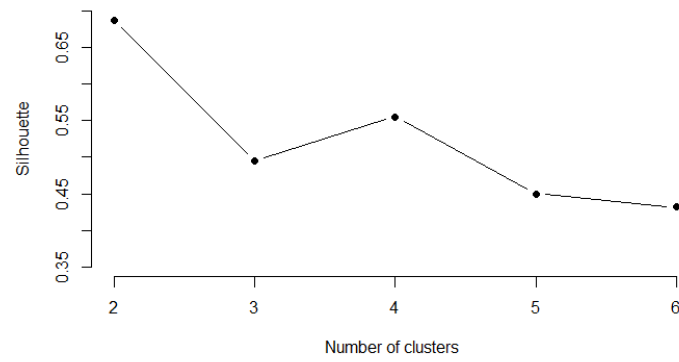


(b) Index G2

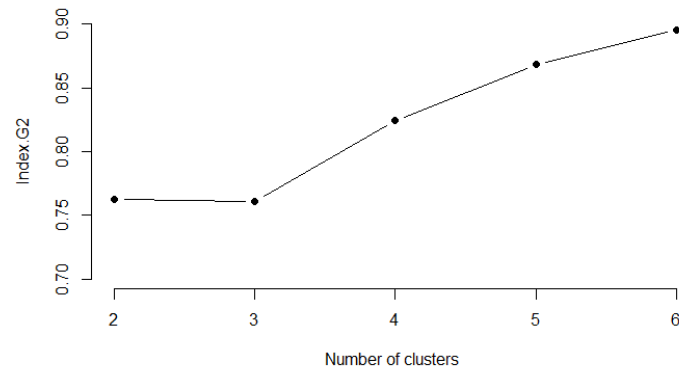


(c) Index G3

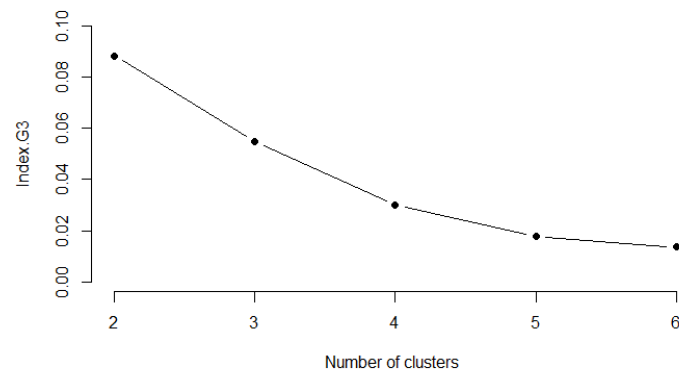
Figure D.5: Validity Indices - 2018



(a) Silhouette

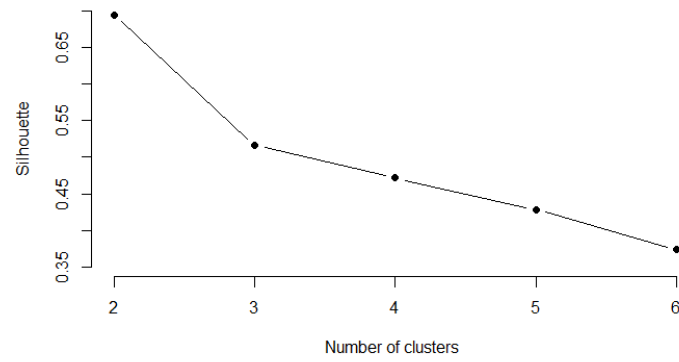


(b) Index G2

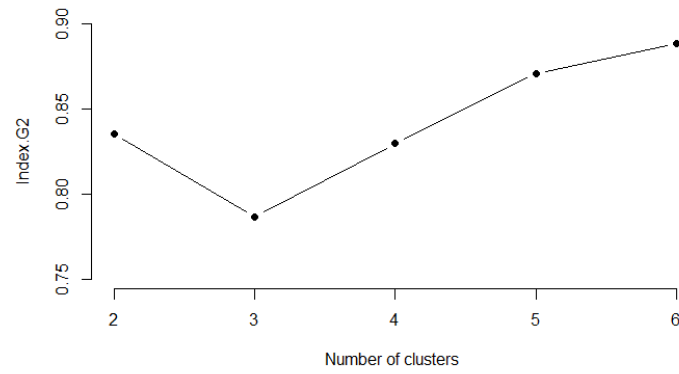


(c) Index G3

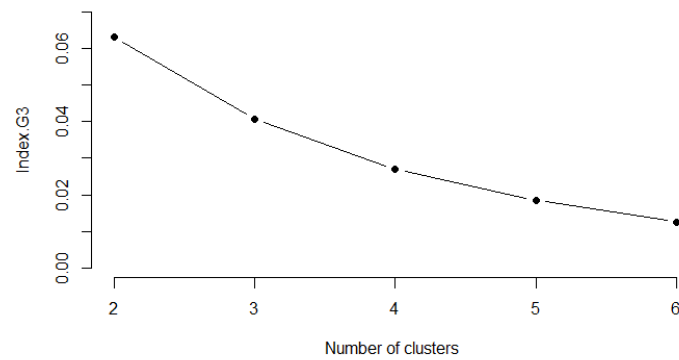
Figure D.6: Validity Indices - 2019



(a) Silhouette

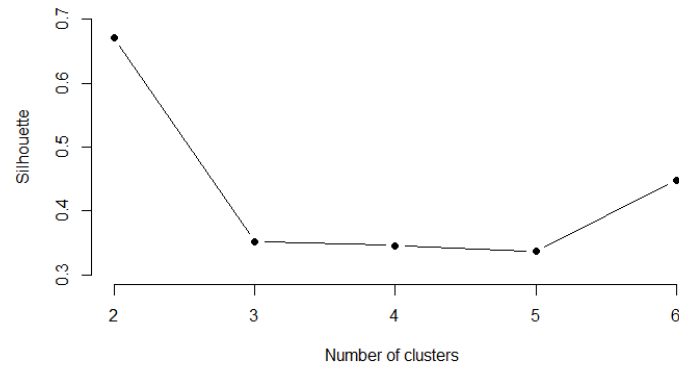


(b) Index G2

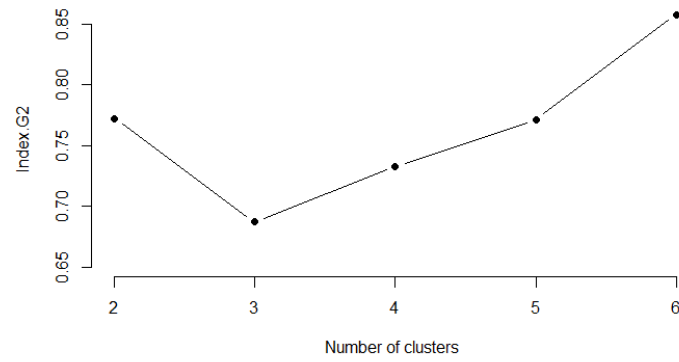


(c) Index G3

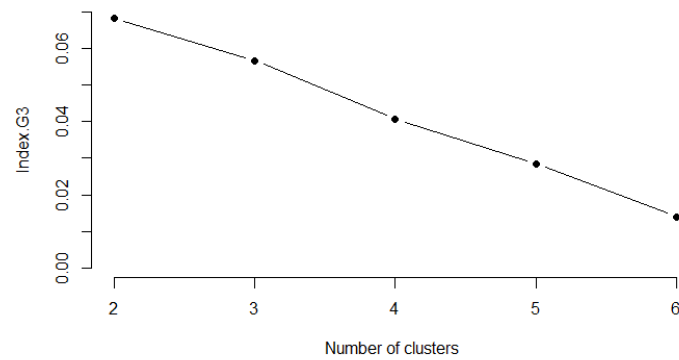
Figure D.7: Validity Indices - 2020



(a) Silhouette

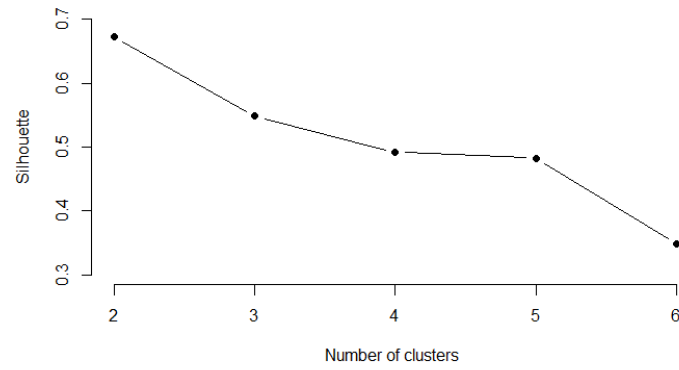


(b) Index G2

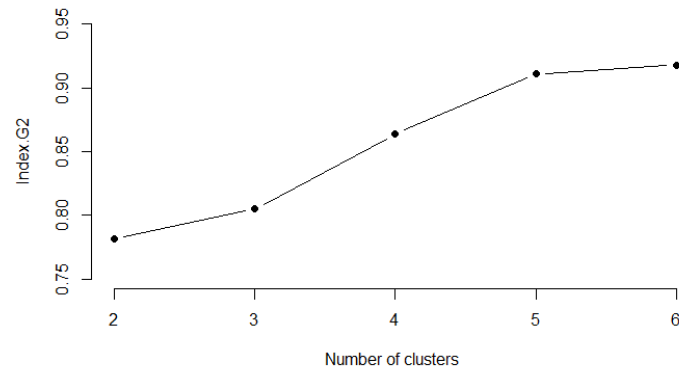


(c) Index G3

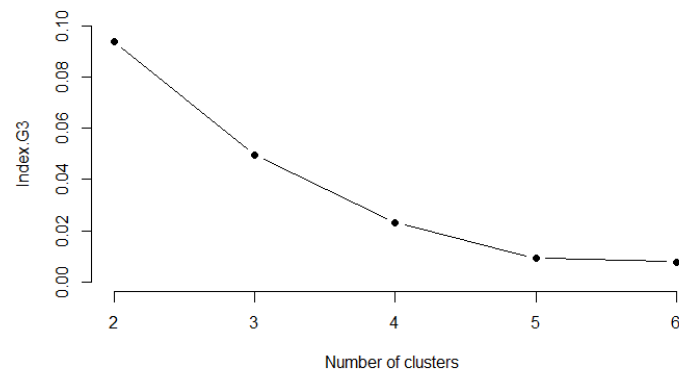
Figure D.8: Validity Indices - 2021



(a) Silhouette



(b) Index G2



(c) Index G3

Figure D.9: Validity Indices - 2022

Cluster	Objects	Description
1	1.EU 1.NC 1.OUTEU 2.EU 2.NC 3.EU 3.NC	• White-collar high-skilled professions • Higher education (around 80% bachelor and over) • Financial and insurance activities • Business Services • Arts, entertainment and recreation and other service activities • Cities • Middle/Upper-income earners³
2	2.OUTEU 3.OUTEU 4.NC 4.OUTEU 5.EU 5.NC 5.OUTEU 6.LU	• White-collar professions • Around 50% have Low and medium education levels⁴ • Public administration, defence, education, health and social work
3	4.EU	• White-collar high-skilled professions • Higher education (around 80% bachelor and over) • Financial and insurance Activities • Arts, entertainment and recreation and other service activities • Upper-income earners (43% belong to the top 10% earners)
4	1.PT 2.PT 3.PT 4.PT 5.PT	• Blue-collar low-skilled professions • Low and medium education levels (85%) • Trade, transport and hospitality • Business Services • Lower-income earners (around 75% earn the equivalent to the 3rd income decile or less)

Table D.10: Cluster Description - 2014

³EARNINGDEC between:
1-3: Lower-income earners;
4-7: Middle income earners;
8-10: Upper-income earners.

⁴Until High School

Cluster	Objects	Description
1	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.EU 4.NC 5.EU	• White-collar high-skilled professions • Higher education (75% bachelor and over) • Financial and insurance activities • Arts, entertainment and recreation and other service activities • Cities • Middle-upper/ Upper-income earners
2	1.OUTEU 2.OUTEU 3.OUTEU 4.OUTEU 5.NC 5.OUTEU 6.LU	• White-collar professions • Around 60% have low and medium education levels • Trade, transport and hospitality • Public administration, defence, education, health and social work
3	1.PT 2.PT 3.PT 4.PT 5.PT	• Blue-collar professions • Low and medium education levels (around 90%) • Construction • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (63% earn the equivalent to the 3rd income decile or less)

Table D.11: Cluster Description - 2015

Cluster	Objects	Description
1	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.EU 4.NC 5.EU	• White-collar high-skilled professions • Higher education (around 75% bachelor and over) • Financial and insurance activities • Arts, entertainment and recreation and other service activities • Upper-income earners (45% among the 30% top earners)
2	1.OUTEU 2.OUTEU 3.OUTEU 4.OUTEU 5.NC 5.OUTEU 6.LU	• White-collar professions standing out by their percentage in the low-skilled ones compared to the other clusters • 55% have low and medium education levels • Trade, transport and hospitality • Public administration, defence, education, health and social work
3	1.PT 2.PT 3.PT 4.PT 5.PT	• Blue-collar professions • Low and medium education levels (around 90%) • Construction • Trade, transport and hospitality • Lower-income earners (58% earn the equivalent to the 3rd income decile or less)

Table D.12: Cluster Description - 2016

Cluster	Objects	Description
1	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.NC	• White-collar high-skilled professions • Master/ PhD (around 60%) • Financial and insurance activities • Business services • Arts, entertainment and recreation and other service activities • Cities • Middle-upper/ Upper-income earners
2	1.OUTEU 2.OUTEU 3.OUTEU 4.EU 4.OUTEU 5.EU 5.NC 5.OUTEU 6.LU	• White-collar professions • Low and medium education levels (around 50 %) • Trade, transport and hospitality • Public administration, defence, education, health and social work
3	1.PT 2.PT 3.PT 4.PT 5.PT	• Blue-collar professions • Low and medium education levels (around 90%) • Construction • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (around 65% earn the equivalent to the 3rd income decile or less)

Table D.13: Cluster Description - 2017

Cluster	Objects	Description
1	1.EU 2.EU 3.EU 4.EU	• White-collar high-skilled professions • Higher education (80% bachelor and over) • Financial and insurance activities • Arts, entertainment and recreation and other service activities • Cities • Upper-income earners (38% among the 20% top earners)
2	1.NC 1.OUTEU 2.NC 3.NC 4.NC 5.EU 5.NC 6.LU	• White-collar high-skilled professions • Financial and insurance activities • Public administration, defence, education, health and social work
3	1.PT 2.OUTEU 2.PT 3.OUTEU 3.PT 4.OUTEU 4.PT 5.OUTEU 5.PT	• Blue-collar professions • Low and medium education levels (around 80%) • Construction • Business Services • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (55% earn the equivalent to the 3rd income decile or less)

Table D.14: Cluster Description - 2018

Cluster	Objects	Description
1	1.EU 2.EU 3.EU 4.EU	• White-collar high-skilled professions • Higher education (around 80% bachelor and over) • Arts, entertainment and recreation and other service activities • Cities • Upper-income earners
2	1.NC 1.OUTEU 2.NC 3.NC 4.NC	• White-collar high-skilled professions • Higher education (75% bachelor and over) • Financial and insurance activities • Business Services
3	5.EU 5.NC 5.OUTEU 6.LU	• White-collar professions and it stands out by its percentage of low skilled ones compared to the other clusters • Low and medium education levels (52%) • Public administration, defence, education, health and social work • Rural Areas
4	1.PT 2.OUTEU 2.PT 3.OUTEU 3.PT 4.OUTEU 4.PT 5.PT	• Blue-collar professions • Low and medium education levels (around 70%) • Construction • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (60% among the 30% that earn the least)

Table D.15: Cluster Description - 2019

Cluster	Objects	Description
1	1.EU 1.NC 1.OUTEU 2.EU 2.NC 2.OUTEU 3.NC 4.NC	• Telework • White-collar high-skilled professions • Higher education (around 80% bachelor and over) • Financial and insurance activities • Business and Services • Cities
2	3.EU 3.OUTEU 4.EU 5.EU	• Telework • White-collar high-skilled professions • Financial and insurance activities • Arts, entertainment and recreation and other service activities • Upper-income earners (45% among the 30% top earners)
3	4.OUTEU 5.NC 5.OUTEU 6.LU	• No telework • White-collar professions and stands out by their percentage in the low-skilled ones compared to the other clusters • Low and medium education (around 60%) • Public administration, defence, education, health and social work • Rural Areas
4	1.PT 2.PT 3.PT 4.PT 5.PT	• No telework • Blue-collar professions • Low and medium education (around 85%) • Construction • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (65% earn the equivalent to the 3rd income decile or less)

Table D.16: Cluster Description - 2020

Cluster	Objects	Description
1	1.EU 1.NC 2.EU 2.NC 3.NC	• Telework • White-collar high-skilled professions • Master/ PhD • Financial and insurance activities • Business Services • Cities • Middle-upper/ Upper-income earners
2	3.EU 4.EU 4.NC 5.EU	• Telework • White-collar high-skilled professions • Higher education (around 75% bachelor and over) • Arts, entertainment and recreation and other service activities • Upper-income earners (around 35% among the 20% top earners)
3	5.NC 6.LU	• No telework • White-collar professions (low-skilled percentage stands out against other clusters) • Low and medium education levels (50%) • Public administration, defence, education, health and social work • Rural areas
4	1.OUT 2.OUT 3.OUT	• White-collar professions • Higher education (70% bachelor and over) • Trade, transport and hospitality • Business Services
5	1.PT 4.OUT 5.OUT 5.PT	• No telework • Low and medium education (around 65%) • Trade, transport and hospitality • Public administration, defence, education, health and social work • Towns and Suburbs • Lower/ Middle-lower income earners (62% earn the equivalent to the 4th income decile or less)
6	2.PT 3.PT 4.PT	• No telework • Blue-collar professions • Low and medium education (85%) • Construction • Trade, transport and hospitality • Towns and Suburbs • Lower-income earners (70% earn the equivalent to the 3rd income decile or less)

Table D.17: Cluster Description - 2021

Cluster	Objects	Description
1	1.EU 1.NC 1.OUTEU 2.EU 2.NC 3.NC 4.NC	• Telework • White-collar high-skilled professions • Master/ PhD • Financial and insurance activities • Business Services • Cities • Middle-upper/ Upper-income earners
2	3.EU 4.EU	• Telework • White-collar high-skilled professions • Master/ PhD • Arts, entertainment and recreation and other service activities • Cities • Upper-income earners (50% among the top 30% earners)
3	2.OUTEU 5.EU	• White-collar high-skilled professions • Trade, transport and hospitality • Arts, entertainment and recreation and other service activities
4	3.OUT 4.OUT 5.NC 5.OUTEU 6.LU	• No telework • White-collar professions • Low and medium education (around 55%) • Public administration, defence, education, health and social work
5	1.PT 2.PT 3.PT 4.PT 5.PT	• No telework • Blue-collar professions • Low and medium education (around 80%) • Construction • Trade, transport and hospitality • Business services • Towns and Suburbs • Lower-income earners (68% earn the equivalent to the 3rd income decile or less)

Table D.18: Cluster Description - 2022

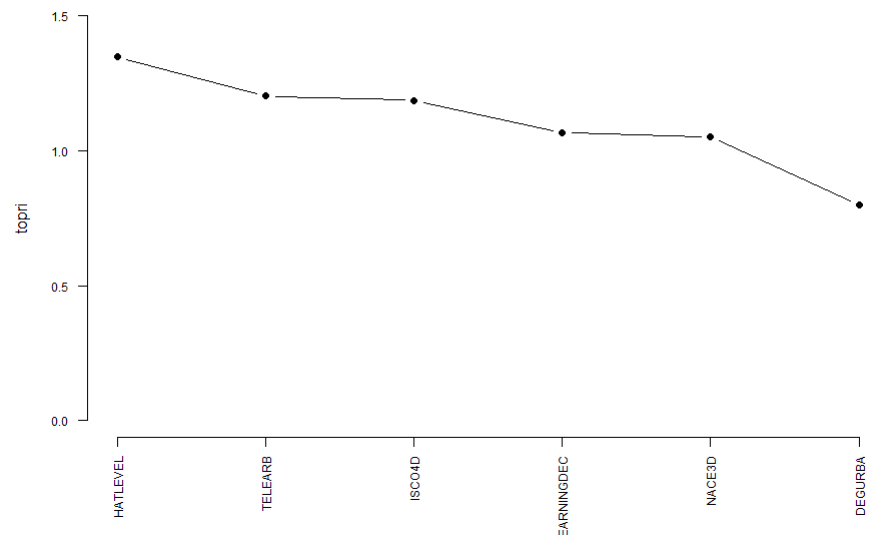


Figure D.10: TOPRI plot 2020

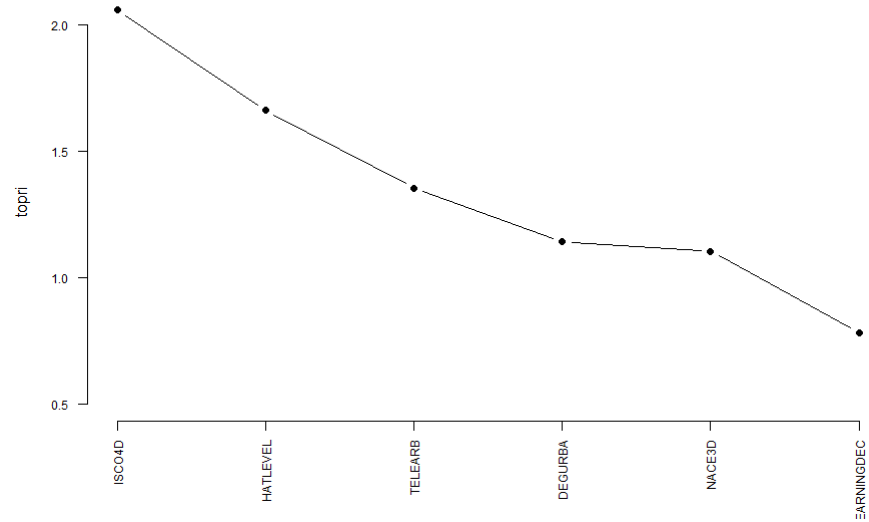


Figure D.11: TOPRI plot 2021

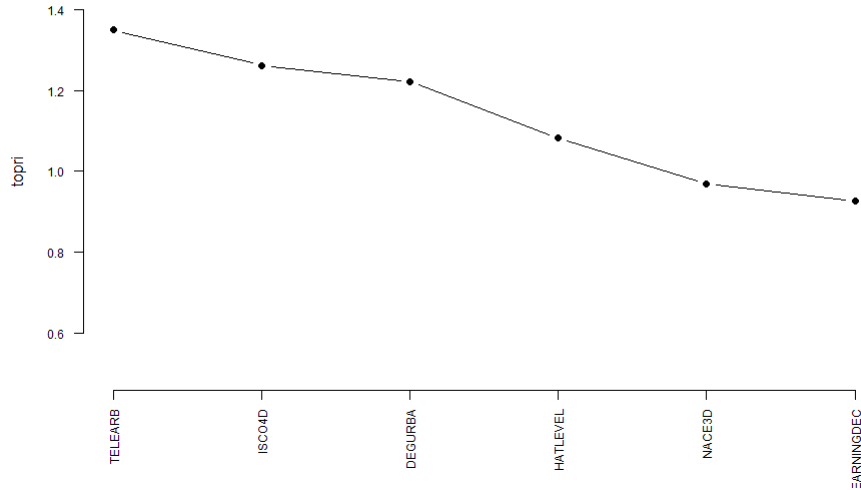


Figure D.12: TOPRI plot 2022

Sample 1 - Sample 2	95 % Confidence Interval	p-value
1.EU (2019) - 1.EU (2020)	]-0.43; -0.21[	4.164×10^{-08}
2.EU (2019) - 2.EU (2020)	]-0.55; -0.31[	1.814×10^{-11}
3.EU (2019) - 3.EU (2020)	]-0.34; -0.09[	0.0005495
4.EU (2019) - 4.EU (2020)	]-0.41; -0.06[	0.007807
5.EU (2019) - 5.EU (2020)	]-0.53; -0.18[	9.946×10^{-05}
1.OUTEU (2019) - 1.OUTEU (2020)	]-0.42; -0.20[	5.902×10^{-08}
2.OUTEU (2019) - 2.OUTEU (2020)	]-0.47; -0.15[	0.000139
3.OUTEU (2019) - 3.OUTEU (2020)	]-0.43; -0.14[	0.0001371
4.OUTEU (2019) - 4.OUTEU (2020)	]-0.24; 0.03[	0.1472
5.OUTEU (2019) - 5.OUTEU (2020)	]-0.34; -0.05[	0.0108
1.NC (2019) - 1.NC (2020)	]-0.39; -0.19[	1.714×10^{-08}
2.NC (2019) - 2.NC (2020)	]-0.49; -0.24[	2.975×10^{-08}
3.NC (2019) - 3.NC (2020)	]-0.38; -0.18[	4.166×10^{-08}
4.NC (2019) - 4.NC (2020)	]-0.46; -0.23[	2.218×10^{-08}
5.NC (2019) - 5.NC (2020)	]-0.36; -0.14[	1.785×10^{-08}
1.PT (2019) - 1.PT (2020)	]-0.37; 0.01[	0.07101
2.PT (2019) - 2.PT (2020)	]-0.19; 0.06[	0.3935
3.PT (2019) - 3.PT (2020)	]-0.08; 0.08[	1.00
4.PT (2019) - 4.PT (2020)	]0.02; 0.20[	0.01084
5.PT (2019) - 5.PT (2020)	]-0.09; 0.06[	0.7688
6.LU (2019) - 6.LU (2020)	]-0.21; -0.16[	$< 2.2 \times 10^{-16}$

Table D.19: Test for equality of proportions of people doing telework in 2019 and 2020

Sample 1 - Sample 2	95 % Confidence Interval	p-value
1.EU (2020) - 1.EU (2022)	]-0.20; 0.08[	0.4626
2.EU (2020) - 2.EU (2022)	]-0.21; 0.04[	0.2062
3.EU (2020) - 3.EU (2022)	]-0.22; 0.03[	0.1371
4.EU (2020) - 4.EU (2022)	]-0.32; 0.04[	0.1409
5.EU (2020) - 5.EU (2022)	]-0.24; 0.19[	0.9709
1.OUTEU (2020) - 1.OUTEU (2022)	]-0.22; 0.04[	0.1874
2.OUTEU (2020) - 2.OUTEU (2022)	]-0.12; -0.20[	0.7051
3.OUTEU (2020) - 3.OUTEU (2022)	]0.07; 0.35[	0.003881
4.OUTEU (2020) - 4.OUTEU (2022)	]-0.12; 0.13[	1.00
5.OUTEU (2020) - 5.OUTEU (2022)	]-0.03; 0.27[	0.1247
1.NC (2020) - 1.NC (2022)	]-0.14; 0.12[	0.9539
2.NC (2020) - 2.NC (2022)	]-0.10; 0.17[	0.687
3.NC (2020) - 3.NC (2022)	]-0.10; 0.13[	0.9139
4.NC (2020) - 4.NC (2022)	]-0.15; 0.11[	0.8592
5.NC (2020) - 5.NC (2022)	]-0.05; 0.19[	0.2546
1.PT (2020) - 1.PT (2022)	]-0.08; 0.37[	0.2565
2.PT (2020) - 2.PT (2022)	]-0.29; 0.11[	0.4456
3.PT (2020) - 3.PT (2022)	]-0.07; 0.10[	0.8797
4.PT (2020) - 4.PT (2022)	]-0.08; 0.06[	0.9797
5.PT (2020) - 5.PT (2022)	]-0.02; 0.12[	0.1948
6.LU (2020) - 6.LU (2022)	]0.05; 0.11[	5.179×10^{09}

Table D.20: Test for equality of proportions of people doing telework in 2020 and 2022

Year	C_1	C_2	C_3	C_4	C_5	C_6
2014	1.EU 1.NC 1.OUTEU 2.EU 2.NC 3.EU 3.NC	2.OUTEU 3.OUTEU 4.NC 4.OUTEU 5.EU 5.NC 5.OUTEU 6.LU	4.EU	1.PT 2.PT 3.PT 4.PT 5.PT	-	-

2015	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.EU 4.NC 5.EU	1.OU TEU 2.OU TEU 3.OU TEU 4.OU TEU 5.NC 5.OU TEU 6.LU	1.PT 2.PT 3.PT 4.PT 5.PT	-	-	-
2016	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.EU 4.NC 5.EU	1.OU TEU 2.OU TEU 3.OU TEU 4.OU TEU 5.NC 5.OU TEU 6.LU	1.PT 2.PT 3.PT 4.PT 5.PT	-	-	-
2017	1.EU 1.NC 2.EU 2.NC 3.EU 3.NC 4.NC	1.OU TEU 2.OU TEU 3.OU TEU 4.EU 4.OU TEU 5.EU 5.NC 5.OU TEU 6.LU	1.PT 2.PT 3.PT 4.PT 5.PT	-	-	-
2018	1.EU 2.EU 3.EU 4.EU	1.NC 1.OU TEU 2.NC 3.NC 4.NC 5.EU 5.NC 6.LU	1.PT 2.OU TEU 2.PT 3.OU TEU 3.PT 4.OU TEU 4.PT 5.OU TEU 5.PT	-	-	-

2019	1.EU 2.EU 3.EU 4.EU	1.NC 1.OUTEU 2.NC 3.NC 4.NC	5.EU 5.NC 5.OUTEU 6.LU	1.PT 2.OUTEU 2.PT 3.OUTEU 3.PT 4.OUTEU 4.PT 5.PT	-	-
2020	1.EU 1.NC 1.OUTEU 2.EU 2.NC 2.OUTEU 3.NC 4.NC	3.EU 3.OUTEU 4.EU 5.EU	4.OUTEU 5.NC 5.OUTEU 6.LU	1.PT 2.PT 3.PT 4.PT 5.PT	-	-
2021	1.EU 1.NC 2.EU 2.NC 3.NC	3.EU 4.EU 4.NC 5.EU	5.NC 6.LU	1.OUTEU 2.OUTEU 3.OUTEU	1.PT 4.OUTEU 5.OUTEU 5.PT	2.PT 3.PT 4.PT
2022	1.EU 1.NC 1.OUTEU 2.EU 2.NC 3.NC 4.NC	3.EU 4.EU	2.OUTEU 5.EU	3.OUTEU 4.OUTEU 5.NC 5.OUTEU 6.LU	1.PT 2.PT 3.PT 4.PT 5.PT	-

Table D.21: Clusters Composition 2014-2022

Bibliography

- Ackerman, M., Ben-David, S., Brânzei, S., & Loker, D. (2012). Weighted clustering. In *Proc. of the AAAI Conference on Artificial Intelligence* (Vol. 26, p. 858-863).
- Ametepe, S. F., Daudbasic, R., Ferring, M., Franziskus, A., Frising, A., Grad, L., ... Tran, H. (2022). *Analyse 05-22 - D'une Crise à L'Autre: La Cohésion Sociale sous Pression* (Report). Luxembourg, Luxembourg: Stateg.
- Bao, C., Peng, H., He, D., & Wang, J. (2018). Adaptive fuzzy c-means clustering algorithm for interval data type based on interval-dividing technique. *Pattern Analysis and Applications*, 21, 803-812.
- Barnard, J., & Meng, X.-L. (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research*, 8(1), 17-36.
- Beine, M., & Souy, B. (2016). The evolution of immigration and asylum policy in Luxembourg: insights from IMPALA. *Kritische Vierteljahresschrift für Gesetzgebung und Rechtswissenschaft*, 132-154.
- Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462), 470-487.
- Billard, L., & Diday, E. (2007). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Billard, L., & Kim, J. (2017). Hierarchical clustering for histogram data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(5), e1405.
- Bock, H.-H., & Diday, E. (2000). *Analysis of Symbolic Data*. Springer Berlin, Heidelberg.
- Boubacar, A., & Niu, Z. (2014). Conceptual clustering. In *Future Information Technology: FutureTech 2014* (p. 1-8). Springer Berlin Heidelberg.
- Brito, P. (1991). *Analyse de Données Symboliques: Pyramides d'Héritage* (PhD thesis). Thèse, Mathématiques de la Decision. University Paris-IX Dauphine.
- Brito, P. (1994). Use of Pyramids in Symbolic Data Analysis. In *New Approaches in Classification and Data Analysis* (p. 378-386). Springer.
- Brito, P. (1998). Symbolic clustering of probabilistic data. In *Advances in Data Science and Classification: Proc. of the 6th Conference of the International Federation of Classification Societies (IFCS-98)* (p. 385-390). Università "La Sapienza", Rome.
- Brito, P. (2002). Hierarchical and pyramidal clustering for symbolic data. *Journal of the Japanese Society of Computational Statistics*, 15(2), 231-244.
- Brito, P. (2014). Symbolic Data Analysis: Another look at the interaction of Data Mining and Statistics. *WTREs Data Mining and Knowledge Discovery*, 4(4).

- Brito, P. (2015). Clustering of symbolic data. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of Cluster Analysis*. Chapman and Hall/CRC.
- Brito, P., & Chavent, M. (2012). Divisive monothetic clustering for interval and histogram-valued data. In *Proc. ICPRAM 2012 - 1st International Conference on Pattern Recognition Applications and Methods* (p. 229-234). Portugal.
- Brito, P., Duarte Silva, A. P., & Dias, J. G. (2015). Probabilistic clustering of interval data. *Intelligent Data Analysis*, 19(2), 293-313.
- Brito, P., & Polaillon, G. (2011). Homogeneity and stability in conceptual analysis. In Nancy, France, A. Napoli, A.V. and Vychodil (Ed.), *Proc. of the 8th International Conference on Concept Lattices and Their Applications*. (p. 251-263). Nancy: INRIA.
- Campos, V. (2007). *Análise Simbólica de Dados e a sua Aplicação na Extração de Informação de Estatísticas Oficiais: Análise do Inquérito à Ocupação do Tempo* (Master's thesis). Instituto Superior de Estatística e Gestão da Informação da Universidade Nova de Lisboa.
- Cangiano, A. (2012). *Immigration Policy and Migrant Labour Market Outcomes in the European Union: New Evidence from the EU Labour Force Survey* (Tech. Rep.). Lab-Mig-Gov working paper, Turin: FIERI.
- Carmone Jr, F. J., Kara, A., & Maxwell, S. (1999). HINoV: A new model to improve market segment definition by identifying noisy variables. *Journal of Marketing Research*, 36(4), 501-509.
- Castles, S., Haas, H. d., & Miller, M. J. (2014). *The Age of Migration: International Population Movements in the Modern World* (5th ed.). Palgrave Macmillan.
- Chavent, M. (1998). A monothetic clustering method. *Pattern Recognition Letters*, 19(11), 989-996.
- Chavent, M., & Bock, H.-H. (2000). Clustering methods for symbolic objects. In *Analysis of Symbolic Data* (p. 294-341). Springer.
- Chavent, M., de Carvalho, F., Lechevallier, Y., & Verde, R. (2006). New clustering methods for interval data. *Computational Statistics*, 21(2), 211-229.
- Choe, C., & Van Kerm, P. (2018). Foreign workers and the wage distribution: what does the influence function reveal? *Econometrics*, 6(3), 41.
- Commission, E., Eurostat, Kraszewska, K., Juchno, P., & Todorova, A. (2020). *Migrant Integration Statistics – 2020 edition*. Publications Office of the European Union.
- Csernel, M., & de Carvalho, F. (1999). Usual operations with symbolic data under normal symbolic form. *Applied Stochastic Models in Business and Industry*, 15(4), 241-257.
- Dantas, A. B. d. S., & de Carvalho, F. (2011). Adaptive batch SOM for multiple dissimilarity data tables. In *Proc. IEEE 23rd International Conference on Tools with Artificial Intelligence* (p. 575-578).
- de Carvalho, F. (n.d.). Extension based proximities between constrained boolean symbolic objects. In *Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96)*. Kobe, Japan.
- de Carvalho, F. (1994). Proximity coefficients between boolean symbolic objects. In *New Approaches in Classification and Data Analysis* (p. 387-394). Springer.
- de Carvalho, F., Brito, P., & Bock, H.-H. (2006). Dynamic clustering for interval data based on L2 distance. *Computational Statistics*, 21(2), 231-250.
- de Carvalho, F., de Souza, R. M., Chavent, M., & Lechevallier, Y. (2006). Adaptive Hausdorff

- distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, 27(3), 167-179.
- de Carvalho, F., Irpino, A., & Verde, R. (2015). Fuzzy clustering of distribution-valued data using an adaptive L2 Wasserstein distance. In *Proc. IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (p. 1-8).
- de Carvalho, F., & Lechevallier, Y. (2009). Partitional clustering algorithms for symbolic interval data based on single adaptive distances. *Pattern Recognition*, 42(7), 1223-1236.
- de Carvalho, F., Lechevallier, Y., & Verde, R. (2008). Clustering methods in symbolic data analysis. In E. Diday & M. Noirhomme-Fraiture (Eds.), *Symbolic Data Analysis and the SODAS Software* (p. 181-204). Wiley.
- de Carvalho, F., & Simões, E. C. (2017). Fuzzy clustering of interval-valued data with City-Block and Hausdorff distances. *Neurocomputing*, 266, 659-673.
- de Carvalho, F., & Tenório, C. P. (2010). Fuzzy K-means clustering algorithms for interval-valued data based on adaptive quadratic distances. *Fuzzy Sets and Systems*, 161(23), 2978-2999.
- de Haas, H., & Vezzoli, S. (2018). *European Migrations – Dynamics, Drivers, and the Role of Policies*. Publications Office of the European Union.
- de Souza, R. M., & de Carvalho, F. (2004a). Clustering of interval data based on City-Block distances. *Pattern Recognition Letters*, 25(3), 353-365.
- de Souza, R. M., & de Carvalho, F. (2004b). Dynamic clustering of interval data based on adaptive Chebyshev distances. *Electronics Letters*, 40(11), 1.
- de Souza, R. M., de Carvalho, F., & Silva, F. C. (2004). Clustering of interval-valued data using adaptive squared Euclidean distances. In *Proc. International Conference on Neural Information Processing* (p. 775-780).
- de Souza, R. M., de Carvalho, F., & Tenorio, C. P. (2004). Two partitional methods for interval-valued data using Mahalanobis distances. In *Proc. Ibero-American Conference on Artificial Intelligence* (p. 454-463).
- Diday, E. (2000). Symbolic data analysis and the SODAS project: purpose, history, perspective. In *Analysis of Symbolic Data* (p. 1-23). Springer Berlin Heidelberg.
- Diday, E., & Noirhomme-Fraiture, M. (Eds.). (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.
- Division sociale. (2023). *Regards sur la Vie des Femmes au Luxembourg* (Report). Luxembourg, Luxembourg: Statec.
- Dudek, A. (2007). Cluster quality indexes for symbolic classification — an examination. In *Advances in Data Analysis: Proc. of the 30th Annual Conference of GfKl* (p. 31-38). Freie Universität Berlin.
- Dudek, A., Pelka, M., Wilk, J., & Walesiak, M. (2022). *symbolicDA: Analysis of Symbolic Data*. Retrieved from <https://CRAN.R-project.org/package=symbolicDA> (R package version 0.6-3)
- D’Urso, P., & Leski, J. M. (2016). Fuzzy c-ordered medoids clustering for interval-valued data. *Pattern Recognition*, 58, 49-67.
- D’Urso, P., & Giordani, P. (2006). A weighted fuzzy c-means clustering model for fuzzy data. *Computational Statistics & Data Analysis*, 50(6), 1496-1523.
- El-Sonbaty, Y., & Ismail, M. A. (1998). Fuzzy clustering for symbolic data. *IEEE Transactions on*

- Fuzzy Systems*, 6(2), 195-204.
- European Commission. (2008). *Statistical Classification of Economic Activities in the European Community* (Working paper). Luxembourg, Luxembourg: Eurostat.
- European Commission. (2009). *Commission Recommendation of 29 October 2009 on the use of the International Standard Classification of Occupations (ISCO-08) (Text with EEA relevance)* (No. 292).
- European Commission and Eurostat. (2011). *Migrants in Europe – A Statistical Portrait of the First and Second Generation*. Publications Office of the European Union.
- European Commission and Eurostat. (2013). *EU Labour Force Survey Explanatory Notes (to be applied from 2014Q1 onwards)* (Working paper). Luxembourg, Luxembourg: Eurostat.
- European Commission and Eurostat. (2015). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2014 – 2015 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2016a). *EU Labour Force Survey Explanatory Notes (to be applied from 2016Q1 onwards)* (Working paper). Luxembourg, Luxembourg: Eurostat.
- European Commission and Eurostat. (2016b). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2015 – 2016 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2018a). *EU Labour Force Survey Explanatory Notes (to be applied from 2017Q1 onwards)* (Working paper). Luxembourg, Luxembourg: Eurostat.
- European Commission and Eurostat. (2018b). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2016 – 2018 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2018c). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2017 – 2018 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2019). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2018 – 2019 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2021a). *EU Labour Force Survey Explanatory Notes (to be applied from 2021Q1 onwards)* (Working paper). Luxembourg, Luxembourg: Eurostat.
- European Commission and Eurostat. (2021b). *EU Labour Force Survey Transcoding Tables*. Retrieved from <https://ec.europa.eu/eurostat/documents/1978984/6037342/Transcoding+tables+-+version+8+July+2021.pdf>
- European Commission and Eurostat. (2021c). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2019 – 2021 edition*. Publications Office of the European Union.
- European Commission and Eurostat. (2022). *Labour Force Survey in the EU, Candidate and EFTA Countries – Main Characteristics of National Surveys, 2020 – 2022 edition*. Publications Office of the European Union.
- Eurostat. (2017). *Beginners: Statistical Concept - What are Official Statistics?* Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Statistical_concept_-_What_are_official_statistics? (Accessed: 2023-01-15)
- Eurostat. (2018). *Archive: First and Second-Generation Immigrants - Statistics on Employ-*

- ment Conditions*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=First_and_second-generation_immigrants_-_statistics_on_employment_conditions&oldid=387944 (Accessed: 2023-05-09)
- Eurostat. (2020a). *Beginners: Labour Market - Basic Definitions*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Beginners:Labour_market_-_basic_definitions#What_is_the_labour_force_survey.3F (Accessed: 2023-01-15)
- Eurostat. (2020b). *Glossary: People Outside the Labour Force*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:People_outside_the_labour_force (Accessed: 2022-12-11)
- Eurostat. (2020c). *Labour force survey (LFS) - Sampling Design, Sample Size and Sampling Errors*. Retrieved from [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Labour_force_survey_\(LFS\)_-_sampling_design_sample_size_and_sampling_errors#Which_sample_design_is_used_in_each_country_.3F](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Labour_force_survey_(LFS)_-_sampling_design_sample_size_and_sampling_errors#Which_sample_design_is_used_in_each_country_.3F) (Accessed: 2022-12-12)
- Eurostat. (2020d). *Work Organisation and Working Time Arrangements in 2019*. Retrieved from <https://ec.europa.eu/eurostat/documents/2995521/11384710/3-29092020-AP-EN.pdf/281a8d79-ad6a-a12d-40bd-bfa2dd781157> (Accessed: 2022-11-21)
- Eurostat. (2021a). *EU Labour Force Survey – Development and History*. Retrieved from <https://ec.europa.eu/eurostat/statistics-explained/index.php?> (Accessed: 2022-12-11)
- Eurostat. (2021b). *Glossary: Employed Person - LFS*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Employed_person_-_LFS (Accessed: 2022-12-11)
- Eurostat. (2021c). *Glossary: Labour Force Survey (LFS)*. Retrieved from [https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Labour_force_survey_\(LFS\)](https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Labour_force_survey_(LFS)) (Accessed: 2022-12-15)
- Eurostat. (2021d). *Migrants more Likely Over-Qualified than Nationals*. Retrieved from <https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210721-1> (Accessed: 2022-12-12)
- Eurostat. (2022a). *Employment and Unemployment (LFS) – Information on Data*. Retrieved from <https://ec.europa.eu/eurostat/web/lfs/information-data> (Accessed: 2022-12-11)
- Eurostat. (2022b). *EU Labour Force Survey – Main Features and Legal Basis*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?EU_labour_force_survey_%E2%80%93_main_features_and_legal_basis#Main_features_of_the_EU-LFS (Accessed: 2022-12-11)
- Eurostat. (2022c). *Glossary: Unemployment*. Retrieved from <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Unemployment> (Accessed: 2022-12-11)
- Eurostat. (2022d). *Main Characteristics of Foreign-Born People on the Labour Market*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_characteristics_of_foreign-born_people_on_the_labour_market#Reasons_for_migrating (Accessed: 2022-11-21)
- Eurostat. (2022e). *Migrant*. Retrieved from https://home-affairs.ec.europa.eu/pages/glossary/migrant_en (Accessed: 2022-12-28)

- Eurostat. (2023). *Coding Lists, Explanatory Notes and Classifications used over Time*. Retrieved from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=EU_labour_force_survey_-_documentation#Explanatory_notes_and_user_guide_for_the_core_variables (Accessed: 2023-08-28)
- Galdino, S., & Maciel, P. (2019). Hierarchical cluster analysis of interval-valued data using width of range Euclidean distance. In *Proc. IEEE Latin American Conference on Computational Intelligence (LA-CCI)* (p. 1-6).
- Gama, J., Carvalho, A. d. L., Faceli, K., Lorena, A. C., & Oliveira, M. (2017). *Extração de Conhecimento de Dados*. Sílabo, Lisboa.
- Gerstnerova, A. (2014). Migrant associations' dynamics in Luxembourg. *Ethnicities*, 16(3), 418-431.
- Glorius, B., & Nienaber, B. (2022). Locational choice and secondary movements from the perspective of forced migrants: A comparison of the destinations Luxembourg and Germany. *Comparative Population Studies*, 47.
- Gordon, A. (1999). *Classification*. Chapman & Hall/CRC.
- Gowda, K. C., & Diday, E. (1991). Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, 24(6), 567-578.
- Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857-871.
- Grilo, F. (2012). *Análise Simbólica de Dados* (Report). Porto, Portugal: Laboratory of Artificial Intelligence and Decision Support.
- Gutfleisch, T. (2022). Hiring discrimination against foreigners in multi-ethnic labour markets: Does recruiter nationality matter? Evidence from a factorial survey experiment in Luxembourg. *Research in Social Stratification and Mobility*, 77.
- Hacibeyoglu, M., & Ibrahim, M. H. (2018). Ef_unique: an improved version of unsupervised equal frequency discretization method. *Arabian Journal for Science and Engineering*, 43, 7695-7704.
- Hajjar, C., & Hamdan, H. (2011). Self-organizing map based on Hausdorff distance for interval-valued data. In *2011 Proc. IEEE International Conference on Systems, Man, and Cybernetics* (p. 1747-1752).
- Hardy, A. (2007). Validation of clustering structure: Determination of the number of clusters. In E. Diday & M. Noihomme-Fraiture (Eds.), *Symbolic Data Analysis and the SODAS Software* (p. 235-262). Wiley.
- Hardy, A., & Kasoro, N. (2009). Une nouvelle méthode de classification pour des données intervalles. *Mathématiques et Sciences Humaines. Mathematics and Social Sciences*(187), 79-91.
- Hardy, A., & Lallemand, P. (2004). Clustering of symbolic objects described by multi-valued and modal variables. In *Classification, Clustering, and Data Mining Applications* (p. 325-332). Springer.
- Hartung, A. (2010). *Structural Integration of Immigrants and the Second Generation in Europe: a Study of Unemployment Durations and Job Destinations in Luxembourg, Belgium and Germany* (PhD thesis). KU Leuven, Belgium.
- Hennig, C. (2023). fpc: Flexible Procedures for Clustering [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=fpc> (R package version 2.2-10)

- Hildebrand, V. A., Pi Alperin, M. N., & Van Kerm, P. (2017). Measuring and accounting for the deprivation gap of Portuguese immigrants in Luxembourg. *Review of Income and Wealth*, 63(2), 288-309.
- Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.
- Ichino, M., & Brito, P. (2022). The quantile methods to analyze distributional data. In P. Brito & S. Dias (Eds.), *Analysis of Distributional Data* (p. 81-102). Chapman and Hall/CRC.
- Ichino, M., & Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4), 698-708.
- International Labour Organization. (2013). *19th ICLS Resolution Differentiating Between Employed Persons, Unemployed Persons and Those Outside the Labour Force* (Report). Geneva, Switzerland: International Labour Office.
- Irpino, A., Verde, R., & de Carvalho, F. (2014). Dynamic clustering of histogram data based on adaptive squared Wasserstein distances. *Expert Systems with Applications*, 41(7), 3351-3366.
- Irpino, A., Verde, R., & de Carvalho, F. (2017). Fuzzy clustering of distributional data with automatic weighting of variable components. *Information Sciences*, 406, 248-268.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651-666.
- Jeng, J.-T., Chen, C.-M., Chang, S.-C., & Chuang, C.-C. (2019). IPFCM clustering algorithm under Euclidean and Hausdorff distance measure for symbolic interval data. *International Journal of Fuzzy Systems*, 21(7), 2102-2119.
- Jeng, J.-T., Chuang, C.-C., Tseng, C.-C., & Juan, C.-J. (2010). Robust interval competitive agglomeration clustering algorithm with outliers. *International Journal of Fuzzy Systems*, 12(3), 227-236.
- Kankaraš, M., & Moors, G. (2012). Cross-national and cross-ethnic differences in attitudes: A case of Luxembourg. *Cross-Cultural Research*, 46(3), 224-254.
- Kejžar, N., Korenjak-Černe, S., & Batagelj, V. (2021). Clustering of modal-valued symbolic data. *Advances in Data Analysis and Classification*, 15(2), 513-541.
- Kim, J. (2009). *Dissimilarity Measures for Histogram-Valued Data and Divisive Clustering of Symbolic Objects* (PhD thesis). University of Georgia Athens, GA, USA.
- Kim, J., & Billard, L. (2012). Dissimilarity measures and divisive clustering for symbolic multimodal-valued data. *Computational Statistics & Data Analysis*, 56(9), 2795-2808.
- Kim, J., & Billard, L. (2018). Double monothetic clustering for histogram-valued data. *Communications for Statistical Applications and Methods*, 25(3), 263-274.
- Korenjak-Černe, S., Batagelj, V., & Japelj Pavešić, B. (2011). Clustering large data sets described with discrete distributions and its application on TIMSS data set. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(2), 199-215.
- Kowarik, A., & Templ, M. (2016). Imputation with the R Package VIM. *Journal of Statistical Software*, 74, 1-16.
- Liu, F. (2016). *Cluster Analysis for Symbolic Interval Data using Linear Regression Method* (PhD thesis). University of Georgia.
- Malerba, D., Esposito, F., & Monopoli, M. (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *WIT Transactions on Information and Communication Technologies*, 28.
- Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the

- number of clusters in a data set. *Psychometrika*, 50, 159-179.
- Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4), 441-458.
- Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical clustering: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6), e1219.
- Noirhomme-Fraiture, M., & Brito, P. (2011). Far beyond the classical data models: Symbolic Data Analysis. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, 4(2), 157-170.
- Ntoutsi, I., Spiliopoulou, M., & Theodoridis, Y. (2009). Tracing cluster transitions for different cluster types. *Control and Cybernetics*, 38(1), 239-259.
- Ogasawara, Y., & Kon, M. (2021). Two clustering methods based on the Ward's method and dendrograms with interval-valued dissimilarities for interval-valued data. *International Journal of Approximate Reasoning*, 129, 103-121.
- Oliveira, M. (2010). *Monitorizando a Evolução de Clusters* (Master's thesis). Faculdade de Economia da Universidade do Porto.
- Oliveira, M., & Gama, J. (2010a). Bipartite graphs for monitoring clusters transitions. In *In Proc. Advances in Intelligent Data Analysis IX: 9th International Symposium, IDA 2010* (p. 114-124). Tucson, AZ, USA: Springer Berlin Heidelberg.
- Oliveira, M., & Gama, J. (2010b). Understanding clusters evolution. In *Proc. Workshop on Ubiquitous Data Mining* (Vol. 500, p. 16-20).
- Pacifico, L. D., & de Carvalho, F. (2011). A batch self-organizing maps algorithm based on adaptive distances. In *Proc. 2011 International Joint Conference on Neural Networks* (p. 2297-2304).
- Park, C., Choi, H., Delcher, C., Wang, Y., & Yoon, Y. J. (2019). Convex clustering analysis for histogram-valued data. *Biometrics*, 75(2), 603-612.
- Partners ASSO Developers. (2004). User Manual for SODAS 2 Software [Computer software manual].
- Peltier, F., & Klein, C. (2020). *Le Luxembourg, Terre d'Immigration* (Report). Luxembourg, Luxembourg: Statec.
- Portail de l'emploi. (2023). *Overview of the Labour Market*. Retrieved from <https://adem.public.lu/en/marche-emploi-luxembourg/panorama-marche-emploi.html> (Accessed: 2023-09-09)
- R Core Team. (2022). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rasson, J.-P., Pirçon, J.-Y., Lallemand, P., & Adans, S. (2008). Unsupervised divisive classification. In E. Diday & M. Noirhomme-Fraiture (Eds.), *Symbolic Data Analysis and the SODAS Software* (p. 149-156). Wiley.
- Reiff, P. (2019). *On Travaille plus d'Heures au Luxembourg que dans les Pays Voisins* (Report). Luxembourg, Luxembourg: Statec.
- Reiff, P. (2021a). *Les Principaux Déterminants des Salaires au Luxembourg* (Report). Luxembourg, Luxembourg: Statec.
- Reiff, P. (2021b). *Les Salaires au Luxembourg, plus Attractifs que ceux des Voisins ? Un Constat à Nuancer* (Report). Luxembourg, Luxembourg: Statec.
- Rizo Rodríguez, S. I., & de Carvalho, F. (2019). Clustering interval-valued data with automatic variables weighting. In *Proc. International Joint Conference on Neural Networks (IJCNN)* (p. 1-8).

- Rizo Rodríguez, S. I., & de Carvalho, F. (2022). Clustering interval-valued data with adaptive Euclidean and City-Block distances. *Expert Systems with Applications*, 198, 116774.
- Rodríguez, O., Calderon, O., & Zuniga, R. (2014). RSDA: RSDA - R to Symbolic Data Analysis [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=RSDA> (R package version 1.2)
- Rota, L., & Larue, B. (2021). *Evolution de l'Emploi en 2020 : le Luxembourg très bien Positionné en Europe* (Report). Luxembourg, Luxembourg: Statec.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.
- Schockmel, M., Hartung, A., Reiff, P., & Neumayr, J. (2022). *Panorama sur le Monde du Travail Luxembourgeois à l'Occasion du 1er Mai* (Report). Luxembourg, Luxembourg: Statec.
- Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., & Schult, R. (2006). Monic: modeling and monitoring cluster transitions. In *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 706-711).
- Statec. (2016). *Luxembourg en Chiffres 2016* (Report). Luxembourg, Luxembourg: Author.
- Statec. (2017). *Will there be more than One Million Inhabitants in 2080?* Retrieved from <https://statistiques.public.lu/en/actualites/population/population/2017/03/20170328.html> (Accessed: 2022-12-11)
- Statec. (2022a). *Labour Force Survey (LFS)*. Retrieved from <https://statistiques.public.lu/en/enquetes/enquetes-particuliers/eft-travail.html> (Accessed: 2022-12-11)
- Statec. (2022b). *Statec* (Vol. 2022). Retrieved from <https://statistiques.public.lu/en/statistique-publique/Statec.html> (Accessed: 2022-12-30)
- Thill, G., & Zahlen, P. (2012). *Regards sur les Flux Migratoires* (Report). Luxembourg, Luxembourg: Statec.
- Thomas, A. (2020). Cross-border labour markets and the role of trade unions in representing migrant workers' interests. *Journal of Industrial Relations*, 62(2), 235-255.
- Van Kerm, P., Yu, S., & Choe, C. (2016). Decomposing quantile wage gaps: a conditional likelihood approach. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 65(4), 507-527.
- Verde, R., & Irpino, A. (2007). Dynamic clustering of histogram data: using the right metric. In *Selected Contributions in Data Analysis and Classification* (p. 123-134). Springer.
- Verde, R., & Irpino, A. (2008). Comparing histogram data using a Mahalanobis-Wasserstein distance. In *Proc. COMPSTAT 2008* (p. 77-89). Springer.
- Walesiak, M., & Dudek, A. (2008). Identification of noisy variables for nonmetric and symbolic data in cluster analysis. In *Data Analysis, Machine Learning and Applications* (p. 85-92).
- Xu, X. (2016). A novel hierarchical clustering framework for complex symbolic data exploration. In *2016 Proc. IEEE 32nd International Conference on Data Engineering Workshops (ICDEW)* (p. 189-192).
- Yang, M.-S., Hung, W.-L., & Chen, D.-H. (2012). Self-organizing map for symbolic data. *Fuzzy Sets and Systems*, 203, 49-73.
- Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology*, 11(1), 8-21.

Zahlen, P. (2012). *Le Luxembourg 1960-2010* (Report). Luxembourg, Luxembourg: Statec.