The Evolution of Immigrant Groups in Luxembourg

A Symbolic Data Analysis Approach

Catarina Campos Silva[1], Paula Brito[2], Pedro Campos[3]

[1]Faculty of Economics, University of Porto, Portugal

[2]Faculty of Economics, University of Porto & LIAAD INESC TEC, Portugal

[2]Faculty of Economics, University of Porto & LIAAD INESC TEC & Statistics Portugal, Portugal

Abstract

Luxembourg, known for its immigration tradition, attracts immigrants for work. This study examines different immigrant groups in the labour market from 2014 to 2022. The data source is the Luxembourgish Labour Force Survey (LFS), and Symbolic Data Analysis (SDA) and the Monitoring the Evolution of Clusters (MEC) frameworks were used for its analysis.

For each year, microdata of the LFS were aggregated and 21 symbolic objects were created based on birthplace and length of residence in Luxembourg, primarily described by the empirical distributions of 16 variables. Hierarchical clustering using complete linkage and the Chernoff's distance was applied, using the variables with maximal clustering information, identified by the Heuristic Identification of Noisy Variables (HINoV) algorithm. Finally, the MEC framework was used to monitor cluster transitions over time by identifying temporal relations between these structures.

Results show that six variables were enough to split the objects into groups with similar labour market profiles. Furthermore, it was demonstrated that workers from the European Union (EU) and Neighbouring countries (NC) have similar profiles while the Portuguese have opposite characteristics. The Luxembourgers are in between. Profiling people from non-EU countries was challenging. Lastly, the MEC framework revealed many movements of the non-EU objects between clusters.

One important outcome for public policies is that different labour market profiles of immigrants were identified, based on the split of the Luxembourgish population into clusters. The combination of the LFS, SDA and the MEC framework allows the replication of the work in countries that use the LFS, enabling comparison of results and future monitoring.

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

1. Introduction

Immigration to Luxembourg is not a recent phenomenon as it dates back to the 19th century. The fact that, nowadays, the percentage of people born in a foreign country surpasses 50%, combined with the fact that employment is the main reason for immigration, makes the country's labour market highly internationalised. Thus, it is quite relevant to carry out a study that makes a comparative analysis of the labour market outcomes across different immigrant

groups and natives, allowing the identification of disparities. This may help policymakers to identify areas where immigrants may require additional support. The time factor is also important to see how things are evolving over the years.

The main objective of this work was to split the Luxembourgish population into homogeneous groups, focussing on the phenomenon of immigration, and describing their profiles in the labour market through time. It is important to note that aggregated data, more concretely groups of immigrant population, and not individual observations, are the basis of this study.

Finally, the most significant contribution of this work is the use of a methodology that is not yet frequently applied in Official Statistics, Symbolic Data Analysis (SDA), but has some advantages which are described later on. Furthermore, accounting for time revealed that some changes may be taking place. To the best of our knowledge, no work has been published combining the data and methodologies used in this work.

2. Related Work

First, a distinction between the national Labour Force Surveys (LFSs) and the European Union Labour Force Survey (EU-LFS) is important. The former is the survey that is done in each country. The second is a cross-sectional and longitudinal sample survey, collated by Eurostat from the LFS data of the participating countries. The EU-LFS is a way of harmonising and comparing labour market data between various countries. There are several overlaps between the two, as the LFSs largely contribute to the EU-LFS.

There are several works about the labour market conditions in Luxembourg. Some use the LFS as a data source (Senyo Fofo et al., 2022), (Eurostat, 2021), (Eurostat, 2022), others use different data sources (Rota & Larue, 2021) and (Reiff, 2021). All of them allow the conclusion that, in general, the Luxembourgish labour market conditions are better than the conditions in other EU countries, making it an attractive country for both native-born and immigrant workers. However, most of the works are about the population as a whole; when immigrant groups are considered, they generally only take one year into account which does not allow to have an idea of the evolution of the outcomes.

Symbolic Data Analysis (Diday & Noirhomme-Fraiture, 2008), an approach allowing handling data with inherent variability, was selected to analyse the LFS data. SDA provides a framework to represent and analyse groups, and not individuals, gathered on the basis of some given criteria. SDA has an extremely important application in the field of Official Statistics since, for confidentiality reasons, the National Statistics Institutes are forbidden to share individual information, thus analysing aggregated data allows for freely sharing results. Microdata aggregation also allows different independent surveys made on the same population to be

combined, if the same variables are used to form groups. Its use in Official Statistics has been demonstrated by several authors (de Carvalho et al., 2008), (Campos, 2007), (Bisdorff, 2000), (Grilo, 2012) and (Brito et al., 2015). However, none of the mentioned works use immigrant population groups as the basis for the study.

3. Methodology

Given that the LFS is the primary European source of information on the labour market, the annual Luxembourgish LFS data from 2014 to 2022 were chosen as the data source for this work. The LFS in Luxembourg has its specificities. In 2014, the sampling units were the households and they were selected by the procedure of random digit dialling. However, from 2015 onwards 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 and 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. Weights are computed so that the sample is representative of the population. In a simplified way, 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 stratum have the same weight (European Commission; Eurostat, 2022). The majority of the variables in this survey are categorical and those that are not were discretized. The cohort of people under study were the employees since they make up around 90% of the labour force every year.

Microdata were pre-processed and two variables, YEARESID (the duration of stay in Luxembourg in completed years) and COUNTRYB (place of birth - Luxembourg, Neighbouring Country, Portugal, EU (excluding the previously mentioned countries) and non-Eu countries), were selected for aggregation.The cartesian product of the categories of these two variables originated 21 symbolic objects in each year which were primarily described by 16 modal variables (multi-valued variables with a frequency attached to each category). Table 1 provides an example for this type of variable. The variable MIGREAS describes the immigration reasons of people born in three different countries. From country A, 55% of people immigrated for reason 1, 15% for reason 2,... Thus MIGREAS is a categorical modal variable.

Table 1: Migration reasons distribution by country of birth

Country of Birth	MIGREAS
A	$\{1(0.55), 2(0.15), 3(0.20), 4(0.10)\}$
B	$\{1(0.45), 2(0.10), 3(0.25), 4(0.20)\}$
C	$\{1(0.50), 2(0.05), 3(0.15), 4(0.30)\}$

As mentioned above, the main objective of the work was to split the Luxembourgish population into homogeneous groups, considering labour market characteristics. For this purpose, cluster analysis was selected to explore the data. More concretly, the hierarchical clustering algorithm using complete linkage and the Chernoff's distance of order $\frac{1}{2}$, $d_{Cher}{}^{\frac{1}{2}}$, with a L_1-type aggregation function (Malerba et al., 2002) was selected as it demonstrated good results, i.e., a good separation between the different clusters and good homogeneity within each one.

Given two discrete probability distributions T and V on one variable Y, one may assess the dissimilarity between two symbolic objects t and v, using the the Chernoff's distance as follows:

$$d_{Cher}{}^{\frac{1}{2}}(T,V) = -\log\left(\sum_{m_j \in O}\left(t(m_j)\cdot v(m_j)\right)^{\frac{1}{2}}\right)$$

where $O = \{m_1, \ldots, m_c\}$ is the category set of the categorical modal variable Y.

When the symbolic objects t and v are described by more than one categorical modal variable, one needs to apply the aggregation function as follows:

$$d(t,v) = \sum_{j=1}^{p}\left(w_j d_{Cher}{}^{\frac{1}{2}}(T,V)\right)$$

where, $\forall j \in 1, \ldots, p, w_j > 0$ are weights with $\sum_{j=1\ldots p} w_j = 1$ and p is the number of variables.

To avoid using variables with minimal cluster information, an extended version, adapted to symbolic variables, of the algorithm Heuristic Identification of Noisy Variables (HINoV) proposed in (Walesiak & Dudek, 2008) was used. The Adjusted Rand Index (ARI) (Hubert & Arabie, 1985), an index of agreement between pairs of partitions, is the basis of this technique. At the end, one gets a plot which ranks the variables from the more meaningful to the least meaningful. The variables chosen were those that over the nine years under analysis were at the top seven times or more. This allowed reducing the set of 16 variables to just six – DEGURBA (degree of urbanisation of the area where the person has his/her usual residence),

TELEARB (ability to do remote work), NACE3D (economic activity of the local unit for the main job), ISCO4D (occupation in the main job), HATLEVEL (highest level of education completed) and EARNINGDEC (decile into which the monthly pay from the main job falls).

Lastly, as it is relevant to monitor the clusters' evolution through time, the MEC framework, with representation by enumeration, proposed in (Oliveira & Gama, 2010), was employed. This framework aims at monitoring cluster transitions over time, through the identification of temporal relations between these structures. In the enumeration representation, a cluster is characterized by the set of its elements and the method to detect transitions is based on the use of graphs and conditional probabilities. This framework introduced the concepts of survival threshold and splitting thresholds which allow for the definition of the different exogenous transitions: birth, death, split, merge, and survival. The main contribution of this methodology is that it allows observing whether there were many or few movements of objects between clusters from one year to the next.

All the work was done using the R software. Although some changes have been made in the sources, the main packages used were symbolicDA, VIM, RSDA, and fpc.

4. Results

To obtain the results, the following steps were considered. For each year, the hierarchical clustering algorithm was applied to the set of 21 symbolic objects described by the distributions of six modal variables. Note that these 21 objects correspond to groups in the population and not to individuals. Based on the dendrograms and three validity indices – Silhouette (Rousseeuw, 1987), Index G2 (Milligan & Cooper, 1985) and Index G3 (Gordon, 1999) - a partition into k clusters was selected. Cluster descriptions were obtained based on the average frequency of each category of each variable for objects belonging to the cluster. The frequencies that stood out indicate that there is some relation between a specific cluster and a particular category of a categorical modal variable. Finally, the MEC framework was applied to monitor evolution through time.

Analyzing the maximum height of each of the dendrograms (3.26, 2.21, 2.11, 2.09, 1.71, 1.53, 1.95, 2.03, 2.31, from 2014 to 2022, respectively) allowed concluding that the objects were getting closer until the Brexit and the start of the Pandemic. This could mean that the disparities between groups were decreasing, but these two events helped to reverse what was happening.

The Portuguese groups are together in a cluster in seven out of the nine years. These clusters are mainly characterized by little teleworking, blue-collar jobs, low levels of education, low wages, and work in construction, trade, transport, and hospitality. However, things may be

changing as the proportion of recently arrived immigrants (group 1.PT) that hold more advanced degrees is increasing. Furthermore, there is a relation between being Portuguese and living in the suburbs.

The 1.EU, 2.EU, 3.EU, 4.EU, 1.NC, 2.NC, 3.NC, 4.NC groups tend to be together in the same cluster. These clusters are usually characterized by work in white-collar high-skilled professions, more concretely, in Financial and insurance activities and in Arts, entertainment and recreation and other service activities, living in cities, having high wages and by the ability to do telework. Having a high education level (Bachelor's or higher) is also a characteristic of these clusters.

The 5.NC and 6.LU groups are always together in a cluster. These clusters are mainly characterized by work in white-collar professions (high and low-skilled), medium education levels, no ability to do telework, working in Public administration, defence, education, health and social work activities, and living in rural areas. The 5.EU group is also in the same cluster as these two groups in 7 out of 9 years.

The OUTEU groups are the ones that move more between clusters. Consequently, it was not possible to find any kind of pattern in data. This phenomenon might be due to the fact that the OUTEU groups include people born in countries with vastly different levels of development.

Lastly, it was possible to conclude that, regardless of the place of birth, the Business service sector is common among recent arrived immigrants (YEARESID=1), while Public administration, defence, education, health and social work activities are common among the longer-established people (YEARESID=5) and the natives. Moreover, the percentage of individuals living in cities declines as YEARESID rises.

The MEC framework allowed verifying that there were a lot of movements throughout the years. Among the clusters that are mostly made up of PT groups all the transitions that took place between them were survival transitions, demonstrating some consistency in the similarity of the working conditions between the PT groups over time. Similarly, transitions between clusters that are mostly made up of EU and NC groups were either survival or merges where all the elements were part of it. This also demonstrates some consistency in the similarity of the working conditions of most EU and NC groups. On the other hand, a lot of movements that are not identifiable as any of the transitions foreseen in the literature and merges often include OUTEU groups. As mentioned above, the place of birth of the people in the population sample that makes up the OUTEU groups is quite distinct and varies over the years. As a result, in some years it is common for OUTEU groups to have more educated people with better working conditions and in other years not so much, giving rise to movements towards groups with more similar characteristics.

5. Final Considerations and Future Work

In this work, we used complementary approaches to study the labour force in Luxembourg, with a special focus on immigrants. With just six variables, the Luxembourgish population was split and different labour market profiles were identified. The Symbolic Data Analysis approach, the MEC framework, and the Luxembourgish LFS data were crucial for achieving the results. Some of the conclusions are in line with existing research. For example, the fact that non-Luxembourg nationals are generally more likely to have a higher level of education than Luxembourgers which is not the case for Portuguese nationals, who have significantly fewer university degrees (Senyo Fofo et al., 2022). Other results were new, for example, the importance of the degree of urbanisation to split the population into homogeneous groups.

It would be interesting to carry out this analysis in the coming years to monitor the changes that might occur. Additionally, given the replicability of this analysis in countries where the LFS is performed, it would be relevant to do a similar research in countries where immigration is also a large phenomenon and see if the patterns are close to those of Luxembourg.

The limitations of this work lie in the inherent constraints of the data and techniques used. One of the major constraints lies in the tools that are currently available to aggregate microdata and analyse symbolic data. Thus, it is suggested the exploration and use of other clustering algorithms that have already been described in the literature.

Finally, the greatest contribution of this work lies in the use of a methodology that is not yet widely used in Official Statistics but has an important application in this field. To the best of our knowledge, no work combines data from the Luxembourgish LFS with SDA. Furthermore, accounting for time, revealed that some changes might be occurring, for example, the level of education of the recently arrived Portuguese.

Acknowledgement

This work was done through a partnership between the Faculty of Economics of the University of Porto and the National Institute of Statistics and Economic Studies of the Grand Duchy of Luxembourg (STATEC) under the European Master in Official Statistics Program (EMOS). We would like to thank Statec for providing the data and giving support.

References

Bisdorff, R. (2000). Illustrative Benchmark Analyses. In H.-H. Bock, & E. Diday (Eds.), *Analysis of Symbolic Data* (pp. 355-385). Berlin, Heidelberg": Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-57155-8_13

Brito, P., Silva, D., & Dias, J. G. (2015). Probabilistic clustering of interval data. *Intelligent Data Analysis, 19*, pp. 293-313. https://doi.org/10.3233/IDA-150718

Campos, V. (2007). Análise Simbólica de Dados e a sua Aplicação na Extracção de Informação de Estatísticas Oficiais: Análise do Inquérito à Ocupação do Tempo.

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* (pp. 181-204). Wiley. https://doi.org/10.1002/9780470723562.ch11

Diday, E., & Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software.* Wiley. https://doi.org/10.1002/9780470723562

European Commission; 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. (2021). *Migrants more Likely Over-Qualified than Nationals*. Retrieved 12 12, 2022, from https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20210721-1

Eurostat. (2022). *Main Characteristics of Foreign-Born People on the Labour Market*. Retrieved 11 21, 2022, from https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Main_characteristics_of_foreign-born_people_on_the_labour_market#Reasons_for_migrating

Gordon, A. (1999). *Classification.* Chapman & Hall/CRC. https://doi.org/10.1201/9781584888536

Grilo, F. (2012). *Análise Simbólica de Dados.* Laboratory of Artificial Intelligence and Decision Support, Porto, Portugal.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, pp. 193-218.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31*(8), pp. 651-666. https://doi.org/10.1007/978-3-540-87479-9_3

Malerba, D., Esposito, F., & Monopoli, M. (2002). Comparing dissimilarity measures for probabilistic symbolic objects. *WIT Transactions on Information and Communication Technologies, 28*. https://doi.org/10.2495/DATA020041

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

Oliveira, M., & Gama, J. (2010). Understanding clusters evolution. *Proc. Workshop on Ubiquitous Data Mining*, *500*, pp. 16-20.

Reiff, P. (2021). *Les Salaires au Luxembourg, plus Attractifs que ceux des Voisins ? Un Constat à Nuancer.* Statec, Luxembourg, Luxembourg.

Rota, L., & Larue, B. (2021). *Evolution de l'Emploi en 2020 : le Luxembourg très bien positionné en Europe.* Statec, Luxembourg, Luxembourg.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, pp. 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Senyo Fofo, A., 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.* Luxembourg, Luxembourg: Statec.

Walesiak, M., & Dudek, A. (2008). Identification of noisy variables for nonmetric and symbolic data in cluster analysis. *Data Analysis, Machine Learning and Applications*, (pp. 85-92). https://doi.org/10.1007/978-3-540-78246-9_11