Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

APPLIED ECONOMETRICS AND FORECASTING

MASTER´S FINAL WORK

INTERNSHIP REPORT

POTENTIAL MARKET VALUE FOR PORTUGUESE TERRITORY: DEMOGRAPHICS AS MOBILITY PROXY

JOANA ISABEL PINTO TOMÁS

OCTOBER – 2023

MASTER
APPLIED ECONOMETRICS AND FORECASTING

MASTER´S FINAL WORK
INTERNSHIP REPORT

POTENTIAL MARKET VALUE FOR PORTUGUESE TERRITORY: DEMOGRAPHICS AS MOBILITY PROXY

JOANA ISABEL PINTO TOMÁS

SUPERVISION:
DR. NUNO GOMES
PROFESSOR ISABEL PROENÇA

OCTOBER – 2023

Glossary

AI – Artificial Intelligence

CRISP-DM – Cross Industry Standard Process for Data Mining

DM – Data Mining

GRP – Gross Rating Point

LS – Local Search

MFW – Master's Final Work

OOH – Out of Home

OTS – Opportunity To See

POI – Point of Interest

POS – Point of Sale

PMV – Potential Market Value

PSE – Produtos e Serviços de Estatística

TSA – Tabu Search Algorithm

ABSTRACT

In retail a business location is of extreme importance as it can directly affect profits. Therefore, it is of the utmost importance to be accurate and critical when allocating products and sales personal to a location. This report is the result of the work performed at PSE[1] during a curricular internship. PSE is a data science company that offers data science services and market research. They are currently developing a model that segments a region according to its Potential Market Value. This model relies on multiple factors, including human mobility data. The challenge lies in implementing the model to new locations for which they do not yet have accurate human mobility data. This was the focus of the work developed, finding a proxy for human mobility by using demographic data pared with information regarding different types of Points of Interest.

To tackle the issue at hand data mining techniques were used following the CRISP-DM methodology as well as the Tabu Search Algorithm, culminating in the creation of a Potential Market Segmentation Model that can be reproduced for multiple locations and even other companies and retail industries.

The results of the new model exceeded the expectations, and this new approach seems to be quite promising. Although there is a loss of information regarding mobility, there is an increase in the knowledge regarding the consumer profile. It is important to notice that the success of the model is related to the type of product, since the consumption of the product used in this study is highly correlated with the residency area of the consumer which means that loosing information about mobility does not pose a major setback.

KEYWORDS: Mobility Proxy; Potential Market Value; Tabu Search

[1] PSE – Produtos e Serviços de Estatística, Company in Data Science and Research, provided the data used in this project, including their own data, the OOH mobility panel.

Table of Contents

LIST OF FIGURES

IN TEXT

APPENDIX

LIST OF TABLES

IN TEXT

APPENDIX

ACKNOWLEDGMENTS

First and foremost, I would like to thank PSE for the opportunity to be a part of their team during the duration of the curricular internship. It was a great professional experience from which I gathered a lot of knowledge.

A special thanks to Nuno Gomes, with whom I worked closely during my months at PSE and learned a lot from, leading to the creation of the project described in this Master's Final Work.

I would also like to express my gratitude to my supervisor Professor Isabel Proença for guiding me and helping me build this Master's Final work. Her guidance was fundamental in putting together all the work that was accomplished during the internship at PSE.

Last but not least, I would like to thank my family and friends, especially my mum, who has been nothing but supportive during my academic journey.

1. INTRODUCTION

Mobility data has become more common and available, making the demand for its analyses grow exponentially. There is a multitude of uses for this type of data, such as urban development or epidemic control (Zhao, et al., 2016). Combined with the identification of Points of Interest and demographic data, mobility data can be used to properly identify the most adequate points of sale for a business (Liu, et al., 2017). Proper identification of the market value, or market potential, of a certain location can be of vital importance to the survival of a retail business (Erbıyık, Özcan, & Karaboğa, 2012). And by ensuring proper territory segmentation and identifying and managing the territories according to their potential, a company can enhance their customer coverage as well as productivity, which will lead to an increase in profits, (Zoltners & Lorimer, 2000).

However, when dealing with mobility data, there are a few difficulties that need to be addressed. Not only the data can be extremely skewed between regions, but we can also have missing mobility data for certain regions, (Liu, et al., 2017). This leads to one of the crucial problems when implementing models with this type of data, which is the initial high costs of obtaining mobility data that are reliable and representative of the population and business in question. Therefore, finding a good proxy to replace these data during an initial part of an analysis can be a good solution to this issue. This was the focus of the internship at PSE that resulted in the present Master’s Final Work.

The internship resulted from a partnership between ISEG and PSE, under the Master’s in Applied Econometrics and Forecasting. PSE[2] is a Data Science company founded in 1994 that has always specialised in Advanced Analytics. They provide consultancy and data science services, implementing technology solutions and advanced market research. Their experience reaches a multitude of sectors (telecommunications, banking, insurance, retail, distribution, energy, mass consumption, government, health, etc.) and they have worked in very different functional areas (marketing, risk, production and operations, distribution, logistics, quality, financial, etc.). PSE is also at the cusp of mobility data collection and analysis. They have their own mobility panel that gathers data from a sample representative of the Portuguese population across the littoral regions of the country and the regions of Lisbon and Porto. The panel is comprised of a sample of about

[2] Produtos e Serviços de Estatística

3000 individuals aged 15 and older, residing in Greater Porto, Greater Lisbon, North Coast, Centre Coast and also in the Faro's District where the panel members have an APP on their smartphones collecting the data. The panel covers 149 municipalities and 7 million inhabitants of the entire population. This panel provides metrics for media audience, mainly the reach of outdoor advertisement.

PSE is presently undertaking a project for a company, that for confidentiality reasons will remain unnamed, that wishes to increase the sales on a specific range of their products. To achieve that, the company wants to know where they should allocate their sales personal and what are the best locations for potential points of sale. Therefore, it is important to consider what are the prime locations with higher Potential Market Value in order to allocate accordingly their sales personal to those locations. Knowing the value of the sales territory and incorporating it into the day-to-day operations of a business can be an extremely important asset in the marketing of products, which can lead to an increase in sales and profits. By having a value associated with the sales territory, a company can have a more effective control of the territory and improve their decision-making process when planning future territorial expansions; Schiff (1960).

Currently, PSE is working with a Predictive Mobility Sales Model that considers multiple factors, including their Mobility Panel Data, to provide a segmentation of the territory by Potential Market Value. But what happens when the company wants to apply the same strategy in a region that does not yet have the mobility panel implemented? This question led to the work developed during the internship.

The task at hand during the entirety of the internship was to analyse real industry data using Data Mining techniques through the CRISP-DM methodology, a standardised method for data mining projects. Culminating in the creation of a model using proxy variables that can replace the mobility data used by PSE in the services they provide to their clients. By using demographics and POI information, this model will allow us to segment the territory by Potential Market Value, identifying areas that can yield better sales. Through the development of this model, as a first approach, PSE can sell their services to a client at a lower cost than when using mobility data since it does not implicate the initial investment of creating a mobility panel for the region or country in question.

PSE is seeking to improve the services they provide to their clients and expanding the existing client network and projects in which they are involved in. This can be achieved by improving the system already in place, as well by creating a less costly service that produces equally satisfactory results. Therefore, the goal was the creation of a model that could be used and reproduced by various companies in retail industries, even when the mobility data is not yet available or easily accessible. The model was tested for the period of August 2022 to July 2023, the history of older years was not included due to the impact of covid. It is important to note that the work performed during this internship had the purpose of serving a real client interested in increasing their sales and profits, hence, some methods used, at times, may be essentially heuristics empirically driven.

This report is structured in the following manner: Chapter 1 provides a brief introduction about the company at which the internship was carried out, as well as the company's motivations and goals regarding the project that is the focus of this MFW. It also provides some context regarding the importance of the project that was executed, as well as the problems faced when undertaking projects such as the one discussed. Chapter 2 provides a literature review that dives deeper into the problems and motivations mentioned in Chapter 1, while also providing some theoretical concepts fundamental to understanding the worked conducted in the latter sections of this MFW. Chapter 3 introduces some methodologies, techniques and technologies used when undertaking Data Mining Projects such as the one in this report, briefly explaining what these methodologies are and focusing on the one that is the industry standard as well as the one used at PSE. Chapter 4 is structured into the different steps that make up a data mining project that follows the CRISP-DM methodology. These are practical steps that include all the actions necessary for the development of the project at hand. Lastly, Chapter 5 provides the main conclusions drawn from all the work developed and the results that were achieved.

2. LITERATURE REVIEW

2.1 Motivation and Framing

Location selection and territory segmentation are of vital importance for the success of retail businesses. In fact, location is one of the most relevant factors affecting profits and sales. The effects of location can be as relevant as other factors, such as pricing product range and customer service. Even small changes in location can have great impacts in performance and profitability; Formanék & Sokol (2022). Poor territory segmentation and alignment can result in workload imbalances for the sales force, which in turn will result in salespeople being unable to properly cover the territory, missing out on valuable customers and wasting time and resources than could be allocated differently in order to increase their customer base, sales and profits; Zoltners & Lorimer (2000). Incidentally, it is imperative to be careful and critical when making these decisions, since it is a crucial part of the success of any retail business (Erbıyık, Özcan, & Karaboğa, 2012).

It is, however, important to note that a perfectly balanced territory alignment should not be expected for any sales force or industry. Various factors, such as data imperfections and geographic restrictions, will influence the territories differently. Another important thing to note is that some studies indicate that, despite the positive relationship between sales and potential territory, this can lead to diminishing returns. Increasing potential can lead to excessive territory workload, which can exceed the available salespeople, resulting in missed sales opportunities. By properly managing the territories with high potential and low potential, productivity will be increased. Some case studies show that by using telemarketing and internet selling or other methods similar to these to reach customers in remote areas, despite being less effective that face to face, the reduction in travel time offsets lost sales. By reducing the area covered and focusing on the customers responsible for most of the sales, a company can reduce costs and increase productivity and profitability (Zoltners & Lorimer, 2000).

One of the challenges when choosing locations and segmenting the territory is often finding the proper data and identifying the factors that matter for the business in question and collecting the data regarding these. Numerous factors can influence location selection and territory alignment. Having the right data and the right attributes, which can and should vary depending on the business in question, will lead to better results.

Transportation infrastructure, public transportation and street centrality seem to be factors that highly affect the location, and in turn successes, of retail. Transportation networks influence greatly the location of retail stores. Another factor that also seems to highly influence the location of retail stores is centrality. Shopping malls are the ones that favour centrality the most, followed by most other types of stores and supermarkets (Lin, Chen, & Liang, 2018). It can be stressed that some retail gains from being more clustered to benefit customers from multi-store selective shopping behaviour and other types of businesses benefit from being more dispersed avoiding the competition of their rivals (Wang, Chen, Xiu, & Zhang, 2014).

Another determining factor in the success of a retail business is the habits of the population in that location. Taking these habits into account will also have a great impact in determining the success of a business. Hence, it is necessary to understand the characteristics of the population of a certain location. Mobility data reflects, in very much detail, the underlying dynamics of the residents of a specific location, and together with POI profiles and demographic information, can provide a very detailed picture of a region (Liu, et al., 2017).

Human mobility concerns how people move, characterising behaviour patterns such as driving to work, walking home or the use of public transportation. Understanding these patterns can be crucial for things such as epidemic control, urban planning, and traffic forecasting. Human mobility tends to exhibit strong temporal regularities, such as going to work or shopping, and these regularities can be used for predicting urban mobility through data mining methods (Zhao, et al., 2016).

Demographic information is complementary to human mobility. It can give us a big understanding of the type of population in an area and the type of markets and products appropriate for these areas. Variables such as population density, age and gender, household income, type of household and housing situation can be used to give a good depiction of the population in the area of study. However, unlike demographic data, it is likely that we find regions where no human mobility data has been collected and mobility data can often be highly heterogeneous between regions (Liu, et al., 2017). These are some challenges of dealing with human mobility data, availability and accuracy.

Hanson (1982) and Pappalardo et al. (2015) have delved into the relationship between demographic aspects and human mobility by looking at the relationship between sociodemographic variables and travel activity patterns while considering spatial constraints. In Hanson (1982) it was found that the sociodemographic factors outweighed the spatial factors when looking at trip frequency. Nonetheless, it is important to note that, when possible, spatial factors should not be disregarded. In Pappalardo et al. (2015) it was found that human mobility patterns are correlated with socio-economic indicators such as education, unemployment rate and income.

POI information is also directly linked to human activities, which can reflect human mobility and help create a representation of it. A business area is often located in a central part of a city and has a high count of POIs, such as offices, shopping malls and restaurants, residential areas tend to be the opposite. This pattern is usually the same across different cities (Jiang, et al., 2020). In Kang (2016) pedestrian volume was proxied by socioeconomic features, such as population density and employment density, and also retail type (type of POI), as well as other variables such as location and transportation attributes.

Taking all the factors discussed above as well as the findings from earlier studies and the relevance of sociodemographic factors in sales, using demographic data and POI information when mobility data is not yet available, in order to improve sales, seems to be an excellent alternative. This approach was used during the project developed at the internship. By using sociodemographic data as well as POI information, we are able to provide a business with a good territory segmentation that will help increase sales and productivity and, in turn, increase a company's profits.

2.2 Theoretical concepts

2.2.1 Data Mining

Data Mining (DM) is a powerful tool that combines statistical analysis, database management, pattern recognition and machine learning. Superficially, it seems to simply be a process of exploratory data analysis, but it is a process that is aimed at finding relationships of interest in large datasets, making it an inductive exercise rather than a hypothetic-deductive process. DM has the advantage of being able to handle extremely large datasets and due to the size of these datasets the impact of contaminated or invalid data can be ignored, it also allows for non-numeric data such as image data, text and geographical data (Hand, 1998).

Marketing, promotion, and sales are some applications of DM, as well as anomaly detection and diagnosis. Some examples of the uses of DM are the approval of loans by analysing individuals with similar incomes, credit and buying patterns and retail product placement-based purchasing patterns, and product association. DM relies on several machine-learning techniques such as decision trees, neural networks, and k-nearest-neighbour, allowing to extract relevant information from the data (Thuraisingham, 2000).

2.2.2 Clustering

Clustering is the process of allocating into different classes data that shows the most similarities in certain characteristics (Yaman, 2021). Clustering allows us to identify characteristics of current customers, potential new customers, and market segmentation. Knowing the characteristics of the costumers is critical for a business marketing strategy. Marketing segmentation allows businesses to customise according to their customers' characteristics, better targeting the customers. Clustering techniques often used in data mining are K-Means, Two-Step and Kohonen Network.

K-Means is a clustering algorithm that can be applied to multiple areas, such as image and speech data compression or even task decomposition in neural networks. This clustering method finds cluster centres by minimising a cost function of distance, or a dissimilarity measure. The measure often chosen is the Euclidean distance (Hammouda & Karray, 2000). The two-step clustering method has two stages. During the first stage, pre-clusters are formed with the goal of minimising the size of the matrix containing the distances for all possible cases. The second stage forms clusters using a hierarchical clustering algorithm with the pre-clusters defined in the previous stage, a range of solutions is then produced during this stage and posteriorly reduced to an ideal number of clusters based on an information criteria (Tkaczynski, 2017). A Kohonen Network is a data mining technique to determine the most appropriate number of clusters that can work with large data sets. This method is a self-organising network constructed for unsupervised learning. This means that it is designed to learn from the structure of the data, making it an extremely useful tool for exploratory data analysis (Yaman, 2021).

2.2.3 Tabu Search Algorithm

Finding a solution for optimisation problems can be quite difficult and these types of problems can be found in a multitude of different practical areas such as telecommunications, logistics and transportation, which motivated the creation of different optimisation techniques, one of which, Tabu Search (TS) (Glover, Taillard, & Werra, 1993). The Tabu Search Algorithm (TSA) is a meta heuristic algorithm, developed by Fred Glover [(1986); (1989)], that is used in combinatorial optimisation problems, finding an optimal solution for situations such as vehicle routing and scheduling, as well as container loading problems (Prajapati, Jain, & Chouhan, 2020). A combinatorial problem is the optimisation of a linear function over a finite set of possible solutions and the field of decision making when the set of solutions is discrete is called Combinatorial Optimisation (Pirim, Bayraktar, & Eksioglu, 2008). However, before delving deeper into the TSA, it is important to understand what heuristics and metaheuristics are (Figure 1 in the Appendix).

Following the definition achieved in Romanycia & Pelletier (1985), a heuristic is an AI tool that can be a program, a rule or even knowledge that one has reason to believe it will be useful and, that when added to a problem-solving system, will improve its performance. Meta-heuristics can be described as higher-level heuristics that are used when searching for a solution for optimisation problems. The term Heuristic is usually used when referring to a procedure that seeks to find an optimum solution, although, not guaranteeing that it will find one, if a solution even exists, while the term Meta-heuristic is used to mention the general frameworks of a heuristic (Pirim, Bayraktar, & Eksioglu, 2008).

TS builds on one of the first and most popular approaches: Local Search (LS). LS is an iterative process that starts with an initial feasible solution, and by applying successions of modifications, progressively improves the solution. The feature that distinguishes TS from LS is that TS has what are called *Tabu Lists*, which record the recent moves made, preventing the cycling to test previous solutions that have already been visited. When discussing TS or LS, it is also important to understand two basic elements, *search space* and *neighbourhood structure*. The *search space* of a TS can be defined as all the possible solutions that can be visited and the *neighbourhood structure* can be defined as the possible transformations, commonly denoted as *N(S)*, that can be applied to the current solution, *S* (Gendreau & Potvin, 2019).

Two other important notions are the concepts of *short-term* memory and *long-term* memory. *Short-term* memory has a limited capacity in time and storage, a *Tabu List* can be considered *short-term* memory. *Long-term* memory differs in storage and time, and it is through *long-term* memory that diversification can be achieved. A solution that has already been visited, with short-term memory, can be revisited within a different neighbourhood, whereas in the long-term memory the probability of a solution being revisited is quite small (Pirim, Bayraktar, & Eksioglu, 2008).

Another extremely important concept when talking about TS is the *Aspiration Criteria*. Despite of the benefits of Tabus, sometimes these may cause stagnation in the searching process or cause relevant moves to not happen. Therefore, it is often necessary to implement conditions to overcome these situations. A common *Aspiration Criteria* used in TS implementations is allowing a move if it produces a better solution than the current one, even if the move in question is Tabu. Finally, it is also important to consider a *Termination Criteria*. A TS implementation could, in theory, go on indefinitely when the optimal value is not known, hence it is necessary to eventually stop the search. Usual termination criteria are to stop after a certain number of iterations (or time), number of iterations with no improvement, or when a predetermined threshold value is attained (Gendreau & Potvin, 2019). In Figure 2 in the Appendix, we can see a generalised layout of implementing a TSA.

The benefit of using algorithms such as these lies in the fact that they can obtain an optimal solution for very sizable problems in short periods of time (Dokeroglua, et al., 2019). In certain optimisation problems, TS seems to outperform other meta-heuristics, and it also had the advantage that it can be stopped at any time, since its iterative nature allows for the algorithm to be stopped once a feasible solution is found. Nonetheless, the success of its implementation lies in correctly specifying it to the problem at hand (Pirim, Bayraktar, & Eksioglu, 2008).

3. METHODOLOGY

3.1 Data Mining Projects and the Main Methodologies Used

As we enter the Big Data realm, it is necessary to use Data Mining Techniques and follow a methodology. Data mining can be a creative and iterative process that does not always follow a standard framework. This poses a problem. The success or failure of a project might depend on the person who is developing it, and results may not be necessarily repeated within a team or company (Wirth & Hipp, 2000). Hence, in any data mining project, regardless of the problem at hand, it is necessary and advisable to follow a uniform methodology. The three main methodologies often considered are the following (Azevedo & Santos, 2008):

- Knowledge Discovery in Databases (KDD), (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).
- Sample, Explore, Modify, Model and Assess (SEMMA), which was developed by the SAS institute.
- Cross Industry Standard Process for Data Mining (CRISP-DM), which was developed by Daimler Chrysler, SPSS and NCR.

These three methodologies have various steps in common, but the CRISP-DM methodology stands out for its focus on starting with understanding the business and the problem at hand before conducting any further analysis. This is also the DM methodology used at the PSE, and it has also become the industry standard. Therefore, the methodology used in this MFW. It aims to make projects more reliable, repeatable, manageable, faster and less costly.

3.2 Cross Industry Standard Process for Data Mining (CRISP-DM)

This methodology can be described as a hierarchical process consisting of a set of tasks organised by four levels of abstraction: *phase*, *generic task*, *specialised task* and *process instance* (Figure 3 in the Appendix). A project consists of various stages and each one is made up of ge*neric tasks*, then followed by *specialised tasks* which replace *generic tasks* for specific situations and finally, the process instance step is a record of all the tasks and decisions performed during the data mining project (Wirth & Hipp, 2000).

In practical terms, we are interested in the CRISP-DM Reference Model, which provides the different stages of a DM project as well as tasks and outputs. This cycle of a DM project will consist of six stages. These stages are not rigid, and it is required to move back and forward between them, as what happens next depends on the outcome of a previous stage (Figure 4 in the Appendix).

Business understanding is the first stage, and its focus is on understanding the objectives and requirements of the project at hand from the perspective of the business, posteriorly converting the problem into a data mining problem. The second step is the *Data Understanding* stage, and it starts with an initial data collection moving forward with a multitude of tasks that allow you to get familiarised with the data, identify quality problems and gain some insights (Figure 5 in the Appendix).

Afterwards, we move on to *Data Preparation*. This stage produces the final dataset, the one that will be used in the modelling. The *Modelling* consists of several modelling techniques that are selected and applied to the problem at hand according to the data available and the aimed objectives.

Evaluation is the stage where the quality of the model is evaluated, seeking for a high-quality model that fulfils the business requirements. Lastly, there is the *Deployment* stage, which can consist of the production of a report or presentation as well as the review of the entirety of the project (Chapman, et al., 2000).

3.3 Technologies Used

The tool used to implement the methodology previously described was IBM SPSS Modeler software. SPSS Modeler is a tool for data science and machine learning that allows for data analysis using a visual drag-and-drop method, where the nodes, that perform the data analysis tasks, follow the concept of a stream, allowing the data scientist to visually implement the necessary steps that allow for data analysis, transformation, and eventual model building. This results in a process that can be easily understood by others that were not involved in the stream's creation (Wendler & Gröttrup, 2016), (Figure 6 in the Appendix).

4. IMPLEMENTATION OF THE CRISP-DM METHODOLOGY

4.1 Business Understanding

The focus of the project conducted during the internship, that resulted in this MFW, was to achieve a satisfactory segmentation of the Portuguese territory by the Potential Market Value it can have to a given business without using PSE's Mobility Panel. Instead, using a set of proxy variables that could approximate adequately the information of the mobility panel, understanding the characteristics of the population and replacing the mobility data.

Segmenting a territory is of utmost importance for companies to better understand and know where to allocate their products in order to increase sales and profits by targeting the right audience for a certain product or service, allowing to optimise the allocation of the resources available.

This project focused on the Predictive Mobility Sales Model that PSE currently has in place for one of their clients. The goal of the client by implementing the model is to know where to better allocate their sales personal as well as products in order to increase the sales of a specific group of products. The current model, aside from factors specific to the business in question, such as data on previous sales, it also uses PSE's Mobility Panel. The problem with relying on PSE's mobility panel lies in the expansion of the model to areas where there is no mobility data yet. PSE is trying to implement this project in other countries in which they do not have their mobility panel data implemented. Hence, the need for an alternative way to calculate this PMV that does not rely on the use of the Mobility Panel Data, which can be costly and hard to implement.

Figure 7 exemplifies the predictive procedure that is currently in place where, by using information regarding commercial history, demography, mobility data, Point of Interest (POI) count and performance of current points of sale (PoS), PSE is able to attribute a Potential Market (PM) value to each *spot*. The *spots* are then segmented according to the PM value, allowing the business to know which *spots* have a higher possibility of sales.

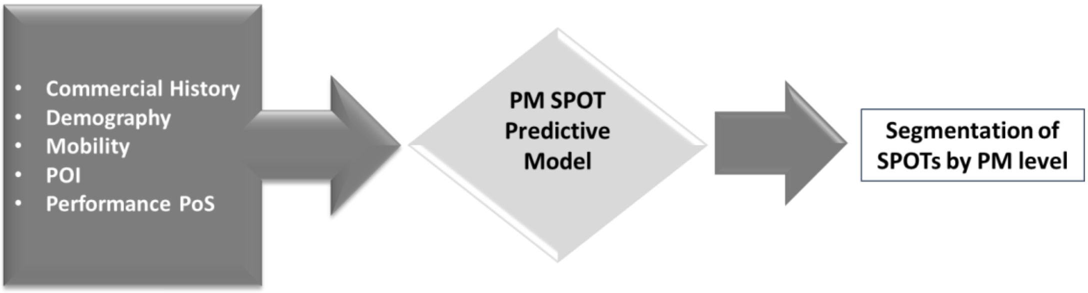


Figure 7 – Predictive Mobility Sales Model (Potential Market Predictive Model) (Source: Adapted from PSE)

PSE'S Out of Home (OOH) Mobility Panel is an official audience measure for outdoor advertising in Portugal. The records are provided from a representative sample of the population, about 3000 individuals, aged 15 years and above residing in Greater Porto, Greater Lisbon, North Coast, Centre Coast and also in the Faro's District, covering 149 municipalities and 7 million inhabitants of the entire population. The sample members have an APP installed on their smartphones that collects the information for the panel. Some of the metrics provided for the study of media audience are Reach (Coverage), Frequency (OTS) and GRPs (PSE PANEL_OOH, 2022).

4.2 Data Understanding

For this project, we worked with data provided by PSE as well as complementary data from PORDATA. PSE works with what they call *spots*. *Spots* are areas of 300 by 300 meters, forming squares that make up for the entire Portuguese territory (1 032 040 *spots*). Therefore, the data unit used in this project is the *spot* and consequently, all data and variables had to be transformed to represent the spot according to the weight of the spot population. The data used was from the latest Census available for the year 202l and the new model will be tested for a history of 12 months.

Regarding the data provided by PSE we have, for each *spot*, the number of residents, number of residents by age group[3], number of POIs by category[4], and mobility/traffic of people and vehicles. See Table I in the Appendix for the description of each type of POI.

[3] The age groups supplied were the following [0,4],[5,9],[10;13], [14,19],[20,24],[25,64], 65>.

[4] POIs are organized in 4 different segments, where 4 is the segmentation with Airports, train terminal stations, bus terminal stations, boat terminal stations, large shopping malls, which are the ones that can be of higher interest for the business. In segment 1 there are Beauty Salons, Workshops, Police Stations, Tennis Courts, Car Washes, Swimming Pools, which are the ones with the least interest for the business.

From PORDATA we gathered data regarding two main dimensions, relevant to characterise and understand the population: education, and economic. For education we collected data providing the number of individuals of the population for each level of schooling (No school, 1st, 2nd and 3rd cycles, high school, post high school, university education)[5]. Regarding the economic dimension we gathered the PPIndex[6], as well as information on salaries and the number of employed individuals of the population by industry sector (Primary, Secondary and Tertiary)[7]. See Table II in the Appendix with an initial list of variables and respective description.

4.2.1 Preliminary Data Transformations

Data preparation is often one of the most time-consuming steps of a data mining project alongside with the data understating step. The time spent on this step is due to its importance. When preparing the data, it is important to consider three key points: (1) raw data can be incomplete; (2) high-quality data will lead to high performing data mining processes, this can be achieved by cleaning and selecting relevant data in the dataset, producing a smaller dataset that will improve results; (3) in order to find high-quality patterns, it is necessary to have high-quality data (Zhang, Zhang, & Yang, 2010). Therefore, proper data preparation will lead to better results.

The data collected from PORDATA is given by county, therefore, in order to be used in this project, it had to be transformed appropriately according to the sample unit used by PSE which are the *spots*. *Spots*, as mentioned above, are squares that make up for the territory, so that, in order to transform the data accordingly, the following steps were taken:

- For the *spots* where there is no population all the variables were set to 0.
- All the variables that concerned number of individuals for a given attribute were transformed into rates per *spot*.
- In order to calculate these rates, the following calculation was performed:

$$Variable\ per\ spot_i = \frac{\left(\frac{Variable\ per\ county}{Nº\ of\ spots\ in\ the\ county}\right)}{Spot_i\ Population}$$

[5] Data from the 2021 Census

[6] Latest year available in PORDATA at the time the project was conducted was 2019.

[7] Both Salaries and Industry Sector are also from the 2021 Census.

- Where the $Spot_i\ Population$ is the resident population of each *spot*, and to determine it the resident population of each parish according to the 2021 census was considered. To determine which *spots* are allocated to each parish the QGIS software was used, and in the cases where a spot is located in two or more parishes its allocation was determined by the size of the area that belongs to each parish. The parish containing the largest area of the *spot* will be the parish allocated to the *spot*. Finally, to determine the population of the *spot*, the population of its parish is divided by the number of *spots* in it.[8]
- For example, in order to calculate the No School rate for a *spot*, the following transformation was done to the variable obtained from PORDATA,

$$No\_School_i = \frac{\left(\frac{No\_School}{Nº\ of\ spots\ in\ the\ county}\right)}{Spot_i\ Population},$$

where *No_School* is the number of individuals with no education in the county where the *spot* is located.

This transformation was not only necessary in order to have data representative of each *spot*, but also because *spots* are extremely small areas. Therefore, it was decided that proportions would be used instead of absolute values so that values would not be too small and easier to interpret.

After the preliminary data analysis (Section 4.2.1) that helped understand the behaviour of the variables as well as of the Portuguese population and territory, it was decided which variables should be included in the model. The percentiles of each variable (*per spot*) were calculated (Table III in the Appendix) and, according to the percentile, the following point system was attributed to each variable:

- $X_i \geq 99\%$ - 20 points
- $99\% < X_i \geq 95\%$ - 10 points
- $95\% < X_i \geq 75\%$ - 5 points
- $X_i < 75\%$ - 0 points

[8] The variables pertaining to the number of residents by age groups, that were provided by PSE, followed a similar approach. Since the population of each spot had already been calculated, its representation by age group is the results of the product of the population by the age distribution according to the municipality where the spot is located. Using the available data from the 2021 census.

This was done in order to follow the methodology already used by PSE in the model they currently have in place, (Table IV in the Appendix).

4.2.2 Data analysis

Before tackling the problem at hand, a preliminary data analysis was conducted. The CRISP-DM methodology emphasises data analysis in order to get familiar with the data. This is a vital and extensive step. After the preliminary data transformations, following the CRISP-DM methodology as well as PSE's working methods, we analysed how the variables behave across the Portuguese territory. This allowed us to get familiar with the data, a crucial aspect of a DM project, and to understand the demography of the *spots*.

For this we focused on the following dimensions: Population Density, Age Group, Level of Education, Economic Class, Total POI distribution[9], Total Traffic, Pedestrian Traffic and Vehicle Traffic. The analysis of these dimensions can be split into two approaches:

- For Population Density, Total POI distribution, Total Traffic, Pedestrian Traffic and Vehicle Traffic we transformed the variables into categorical ordinal variables that followed a Low to High classification.
- The remaining dimensions were also transformed into ordinal variables following clustering techniques considering the multiple variables that make up each dimension.
- It is important to note that in both cases it was taken into consideration the situation where a *spot* does not have a population. Aside from the Low to High classification, it was also considered a segment for No Population.

Population Density

For the analysis of this dimension, the variable pertaining to the number of residents per *spot* provided by PSE was used. The population density was calculated by dividing the number of residents by the area of a *spot*, and posteriorly transformed into a categorical variable, producing then the following map and pie chart, (Figure 8). It is possible to see that 21% of the *spots* have no resident population, 66% have low

[9] For this analysis it was calculated the total POI count per *spot*, this was done by summing the POI from the four different levels of POIs.

population density (less than 100 inhabitants per km2), 11% have medium levels of population density ([100,1000[inhabitants per km2), and only 2% are highly populated (more than 1000 inhabitants per km2).

Through visual inspection of the map, we can see that the Portuguese population seems to be concentrated in the North and Central coastal territories, with a greater concentration on large cities, specifically Lisbon and Porto. The Algarve region also shows highly populated areas, although with at a lower density than the regions of Lisbon and Porto. The remaining areas present mostly low population density and small pockets of medium population density scattered across the country.

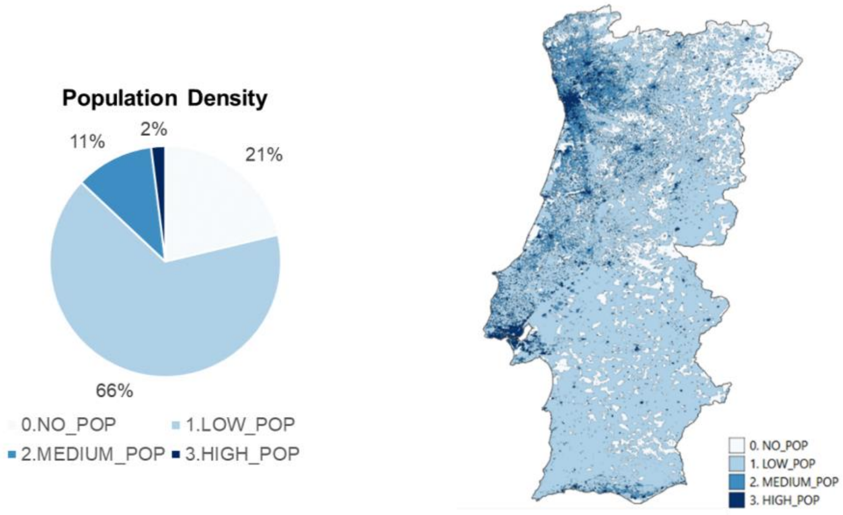


Figure 8 – Pie Chart - Population Density % (by category) (Left);
Map of Population Density by Category (Right)

Total POI distribution

In the case of the distribution of total POI count per *spot* across the country we can conclude that there is a large percentage of *spots* that do not have any POI (94.48%), which is expected when considering that many of the *spots* are located on roads, rural areas and forest areas, and these are squares of only 300x300m. Only 0.65% *spots* have over 9 POIs, 0.86% have between 4 and 9 POIs, and 4.01% have less than 4 POIs. The spots that present a higher POI count are in the big cities, Lisbon and Porto, and also in the Algarve region, followed by the regions that surround these centres, also the coastal region has a higher concentration of *spots* with more POIs than the interior. Most *spots* do not have POIs or a have a very low count of POIs (Figure 9 and Table V in the Appendix).

Total Traffic

Total traffic is the combination of pedestrian traffic and vehicle traffic and for the calculation of the percentages of the variables pertaining to mobility, only the *spots* that are a part of the regions covered by the PSE OOH panel were considered. The analysis showed us that 71% of *spots* have no traffic, 14% have low levels of traffic, 11% medium levels, 3% high levels and only 1% have dense traffic. Through visual analysis we can see that the areas of the country with the most total traffic are the North and Central Coast, with a focus on the big cities, especially Lisbon and Porto, following the Algarve region and several roads, the main arteries of the country, (Figure 10 in the Appendix).

Vehicle Traffic

Regarding the analysis of vehicle traffic separate from pedestrian, we can see that 72% of the *spots* have no vehicle traffic, 14% have little vehicle traffic. 11% have average traffic levels, 2% high levels and only 1% dense vehicle traffic. This follows a very similar distribution with Total Traffic since most of the total traffic registered is made up of vehicle traffic. Through visual inspection, it is clear that the areas of the country with the most vehicle traffic are the north and central coast, with a focus on large cities, especially Lisbon and Porto, especially the metropolitan centres, (Figure 11 in the Appendix).

Pedestrian Traffic

Pedestrian traffic pertains to the movements registered on foot, without the use of any vehicle. Our analysis showed that 88.67% of the *spots* in the area covered by the OOH panel have no pedestrian traffic, 9.77% have little pedestrian traffic, and only 1.56% have medium/high pedestrian traffic. Visual inspection shows that the areas of the country with the most pedestrian traffic are the metropolitan centres of Lisbon and Porto, the north and centre coast, followed by the Algarve region (Figure 12 in the Appendix).

Age Group

The following analysis is the result of clustering the different variables pertaining to age, obtaining a segmentation of the territory by age that considers the following categories: No Population, Young, Working Age and Elderly. The segmentation was done by performing Two-Step, K-means and Kohonen. The results of the three methods were compared, and the segmentation provided by K-means was the selected one. When comparing the three clustering techniques, Two-Step and Kohonen were producing too many segments that, for the purpose of this analysis, seemed unnecessary, hence why, ultimately, K-means was deemed as the one producing the most adequate results. The analysis shows that 21% of the *spots* have no resident population, 14% of the *spots* have a concentration of young individuals, 36% have a concentration of individuals of working age and 29% have a higher concentration of elderly people. The oldest population is found in the innermost regions of the country. The younger population is dispersed over small regions in various parts of the country and individuals of working age are concentrated throughout the coast and south (Figure 13 in the Appendix).

Education Level

The following analysis follows the same approach as the one in Age Group, but in this case, it is the result of the combination of all the variables pertaining to education level, obtaining a segmentation of the territory by education that considers the following categories: No Population, Low Education, Medium Education and High Education. For the same reasons as previously stated, K-Means was also selected as the one providing the most desirable segmentation. It was concluded that 21% of the *spots* have no resident population, 21% have low levels of education, 31% average/medium levels of education and 27% high levels of education. The population with more education is found in the region of Lisbon and Tagus Valley, followed by the Algarve region. There are also some centres in the Interior and in the North. The population with average levels of education is found along the coast of the country, as well as in the Southern Interior region. The population with less education is found in the Northern Interior and Central Interior of the country (Figure 14 in the Appendix).

Economic Class

This analysis follows the same approach as the previous two cases. In this case, combining Salaries, *PPIndex* and Sector (Primary, Secondary, Tertiary), achieving the segmentation: No Population, Low Class, Middle Class and Upper-Middle Class. And again, the results provided by K-Means were chosen for the same reasons as previously stated in the case of Age Group. The results show that 21% of the *spots* have no resident population, 21% have a predominance of individuals who belong to the lower class, 38% have a predominance of individuals who belong to the middle class and only 20% have a predominance of individuals who belong to the upper middle class. The Upper Middle Class is mostly found in the region of Lisbon and Tagus Valley, followed by the Algarve, Porto, and some regions of the Interior. The Middle Class is widely distributed throughout the country and the Lower Class is found mostly in the North Interior, South Interior and in the South Coast region (Figure 15 in the Appendix).

4.3 Modelling

It is important to note that for the analysis conducted in this section, only approximately 240 000 *spots* were considered since those are the *spots* that correspond to the areas where PSE has their mobility panel. This was done in order to compare the results of the model PSE currently has in place, using mobility, with the new results obtained with the use of TSA and Socio-Demographic data, allowing us to assess the performance of the new model.

4.3.1 Potential Market Value Segmentation

The current model that PSE has in place to calculate the PMV indicator for each *spot*, using mobility data, has the following objective function:

$$PMV_i = 1 \cdot \ P\_POI\_1_i) + 2 \cdot (P\_POI\_2X_i + P\_POI\_3_i) + 3 \cdot (P\_Total\_Traffic_i + P\_POI\_4_i) + 4 \cdot (P\_Residents_i + Pedestrian\_Traffic_i), \text{ i} = 1, \ldots., 240\ 000 \text{ SPOTS}$$

The weights that are currently in place are the result of the adjustments made over the course of the two years of working on this project. Since PSE initially did not have any measure of consistency/assertiveness for the model, during the beginning of its implementation, creating a history of successes (new sales generated using the model) served as a base metric for redefining not only the objective function but also the respective weights. The final objective function and its weights also had to respect a set of business rules imposed by PSE's client in order to suit their needs and commercial capabilities.

PMV_i gives the PMV of $spot_i$. Using the PMV generated by this function, *spots* are then segmented according to the percentiles of the PMV,

- Segment 0: if the *spot* has no information regarding mobility.
- Segment 1: percentile < 40%
- Segment 2: percentile ≥ 40% and percentile < 80%
- Segment 3: percentile ≥ 80% and percentile < 97%
- Segment 4: percentile ≥ 97%

The segments 1 through 4 are the four demand segments, meaning that the *spots* allocated to segment 1 are the ones that will be less prone to sales, whereas the *spots* allocated to segment 4 will be the ones with a higher possibility of sales rate. The focus of PSE is to accurately identify the *spots* that belong to each segment, especially the ones in Segment 3 and 4 since these are the *spots* responsible for higher level of sales. Now that the current model is defined, let us discuss the goal of implementing the new model that was the focus of this MFW.

The goal of the new *spot* segmentation model is to be able to achieve a satisfactory result where the loss of information regarding the mobility of the population is compensated by an increase in detailed demographic information regarding the resident population. It is expected that this increase in knowledge regarding the population characteristics may be extremely relevant since recorded sales performance of any product tends to be directly associated with the consumer profile and typology.

The new model, without the mobility data and using instead the sociodemographic variables, has the following objective function,

$$\begin{aligned} PMV_i = {} & P_1 \cdot P\_Residents_i + P_2 \cdot P\_No\_School_i + P_3 \cdot P\_Basic\_Education_i + P_4 \\ & \cdot P\_High\_School_i + P_5 \cdot P\_Post\_High\_School_i + P_6 \cdot P\_University_i \\ & + P_7 \cdot P\_PPIndex_i + P_8 \cdot P\_Salaries_i + P_9 \cdot P\_Primary_i + P_{10} \\ & \cdot P\_Secondary_i + P_{11} \cdot P\_Tertiary_i + P_{12} \cdot P\_Dependents_i + P_{13} \\ & \cdot P\_Young_i + P_{14} \cdot P\_Working\_Age_i + P_{15} \cdot P\_Elderly_i + P_{16} \\ & \cdot P\_POI\_1_i + P_{17} \cdot P\_POI\_2_i + P_{18} \cdot P\_POI\_3_i + P_{19} \cdot P\_POI\_4_i, \mathrm{i} \\ & = 1, \dots., 240\ 000\ \text{SPOTS} \end{aligned}$$

Where $P_j, j = 1, \dots, 19$ are the weights attributed to the variables. The next challenge was to find combinations of the weights associated with the variables, that gives rise to a final PMV that tends to equal in the higher percentiles (higher demand) the *spots* identified by the current model, so that, these can then be segmented according to the demand segments, following the same 1 through 4 levels of demand as previously described. With this goal in mind, the TSA was then applied.

4.3.2 Implementation of the Tabu Search Algorithm

Once the new model was defined, the next step was to attribute a value to each weight $P_i, i = 1, \ldots, 19$, so that the new model would *match* the results of the current one, specifically for the *spots* belonging to Segment 3 and 4. This means that the scores for the higher percentiles need to, on average, match in both models. For this, the TSA was implemented with the following structure,

- **Initial Solution**, S_0: $P_1 = 0, \ldots, P_{19} = 0$
- It was defined that the weights can vary between 1 and 19.
- **Next solution**: The next solution is given by incrementing by one unit the weight of the first variable. If the solution generated leads to a better *match*, the weight of this variable continues to increase. Otherwise, the increase in weight recorded is cancelled out and the next variable is advanced.
- **Termination criterion:** 3 hours. It is important to note that initially a match percentage was used as the Termination Criteria, but for the sake of improving optimisation, it was then decided to proceed with running time as the Termination Criteria.
- **Visited Solution Vector (Tabu List):** Each run of the algorithm generates a listing of all tested solutions so that these cannot be used in future interactions.
- **Randomness Criterion:** In order to introduce a jump criterion in the algorithm, with the goal of looking at new neighbourhood of solutions, it was defined that after 20 iterations in a row without any improvement, 3 variables are randomly selected, and their weights are also randomly assigned. This becomes the new starting point for the algorithm.
- **Temporal Space:** The current model is executed with a weekly periodicity. In order to better evaluate the quality of the new model, it was executed and tested for the history of the last 12 months (August 2022 to July 2023). It was decided not to use history from older years, due to the deviant impact that covid has had on mobility and consumption patterns.

For each solution tested using the Tabu Search algorithm, we have the following Confusion Matrix:

Table VI: Confusion Matrix

	Real (Top Spot)	Prediction (Top Spot)	
		YES	NO
SPOT 1	YES	True Positive (TP)	False Negative (FN)
	NO	False Positive (FT)	True Negative (TN)
SPOT 2	YES	True Positive (TP)	False Negative (FN)
	NO	False Positive (FT)	True Negative (TN)
SPOT 3	YES	True Positive (TP)	False Negative (FN)
	NO	False Positive (FT)	True Negative (TN)
…	…	…	…
SPOT N	YES	True Positive (TP)	False Negative (FN)
	NO	False Positive (FT)	True Negative (TN)

T – True; F – False

Source: PSE

Using the confusion matrix, we are able to determine an assertiveness score for each tested solution, where *Score = Sum of TP/Total Top Spots*. For each solution, there can only be classified as *Top Spots*, the same number of spots that already exist classified as such. By following this principle, the algorithm will look for a neighbouring solution from the one previously tested, advancing whenever the new solution obtains a better score.

The final solution given by the algorithm is the following, $P_1 = 18, P_2 = 1, P_3 = 1, P_4 = 3, P_5 = 1, P_6 = 8, P_7 = 12, P_8 = 13, P_9 = 1, P_{10} = 2, P_{11} = 1, P_{12} = 2, P_{13} = 4, P_{14} = 5, P_{15} = 1, P_{16} = 6, P_{17} = 6, P_{18} = 16, P_{19} = 17.$

4.4 Model Evaluation

To evaluate the performance of the new model, we looked at two different measures:

- *Match Top Spots* – How accurately the new model is *matching* the *spot* segmentation according to the results given by the model previously in place, focusing on Segment 3 and 4 since these are the segments where higher demand is expected.
- *Sales Coverage* – How well we are capturing the *spot* segmentation responsible for most of the sales.

The final solution given by the TSA, for the last 12 months, registered a correspondence of 78% for the two segments with the highest demand, segment 3 and 4.

This is an extremely positive result taking into account the loss of the mobility data. By analysing the weights, we can observe that the resident population and the number of POIs are still extremely relevant. Regarding the new variables, certain variables seem to stand out more, such as the purchasing power index, salary, and higher education, as well as *spots* that are associated with the Working age category.

When looking at how the new model is capturing the sales associated with the current segmentation model, we were able to assess that the new model is being able to capture, on average, 91% of the sales. This value exceeds the initial expectations, which is extremely positive and promising. In the months were there is typically an increase in the mobility of the population (vacation/summer months) we can see that the new model suffers a decrease in the *Match of Top Spots*. However, the variation is not as significant in terms of *Sales Coverage*. This means that, even though some *spots* are not being identified as high demand *spots*, the *spots* that are being identified are the ones responsible for most of the sales representation. As a global result, we have a *Match Top Spots* coverage of 78% and a *Sales Coverage* of 91% (Table VII).

Table VII: Evaluation Results for the last 12 months

Month	Match Top Spots	Sales Coverage
Aug/22	68%	83%
Set/22	73%	89%
Oct/22	82%	94%
Nov/22	84%	94%
Dec/22	72%	85%
Jan/23	80%	93%
Feb/23	82%	94%
Mar/23	81%	93%
Apr/23	84%	93%
May/23	80%	94%
Jun/23	78%	89%
Jul/23	72%	85%
Average	78%	91%

Source: PSE

It is important to note that the positive results of the model are most likely associated with the type of product under analysis in this study. The consumption of this product is highly correlated with the consumer's area of residence and therefore, the loss of mobility information does not seem to represent a major setback in the segmentation of the Potential Market of this product.

4.5 Deployment

The results obtained exceeded the initial expectation and appear to be quite promising, allowing not only to confidently advance to new markets but also the certainty of being able to adapt this approach to other types of products associated with new consumer profiles.

The next goal of PSE is to expand the uses of the model developed to other countries, as well as to apply the same methods to other industries with different product ranges. This will be done by testing similar models and approaches to the new locations and products and testing the performance of the models in a similar manner to what was done for the case of Portugal.

5. CONCLUSION

Location is one of the determining factors in the success of a retail business (Formanék & Sokol, 2022). Through the use of statistical methods and data mining, it is possible to provide businesses with indicators regarding possible locations that will lead to an increase in sales. When determining locations that have a high Potential Market Value, mobility can be taken into consideration as well as sociodemographic factors about the resident population (Liu, et al., 2017). However, whereas demographic data is easily accessible, accurate mobility data can be more difficult to obtain.

Achieving a model that would accurately produce a segmentation of the country by its Potential Market Value, without the use of mobility data, but proxying as best as possible this information, was the goal of the work developed during the internship that resulted in this MFW.

By following the CRISP-DM methodology the initial focus was *Business Understanding*. It is important to understand the goals of the business in question and why solving the problem is important. In this case, properly segmenting the territory according to its potential market value can help business understand where and how to allocate their products in order to better fit the consumer's needs and, in turn, increase their profits. Then, it is important to properly analyse the data we are using, *Data Understanding*. In this step it is crucial to become familiar with the data and understand the characteristics of the population we are dealing with, in order to then be able to prepare the data in a way that is suitable for the project and goals at hand. *Data preparation* is one of the most important steps of a Data Mining project. Data can often be heterogeneous, which will negatively impact the results, so, by properly prepare the data, we can improve the results of a data mining project. Once these steps were properly taken, it was then possible to proceed to the *Modelling* phase.

The model currently in use at PSE relies on POI information and mobility data. The new model replaces mobility data with demographic variables, and to each of these variables, a weight had to be attributed to. The challenge here was to find a combination of weights that would tend to match the results of the current model, especially by having both models match in the identification of the *spots* that will be in the higher percentiles of the Potential Market Value, i.e., the *spots* where the demand will be higher.

The final solution given by the TSA had a correspondence of 78% for the two *spot* segments with the highest demand, which is an extremely positive result and exceeded the initial expectations of the model. Looking at the weights obtained for each variable, it was possible to see that resident population and POI count, variables already present in the current model, remain extremely relevant (Jiang, et al., 2020). Resident population has the highest weight, 18, and the four types of POI (1,2,3,4) have respectively weights of 6, 6, 16 and 17. In the case of the new sociodemographic variables that were added to the new model, Purchasing Power Index, Salaries, University Education and Working age stand out, as they have respectively the following weights, 12, 13, 8 and 5 (Pappalardo, et al., 2015).

Regarding the sales coverage, the new model, despite only capturing 78% of the Top Stops, on average captures 91% of the sales. This means that new model, despite losing some of the Top Stops identified by the current model, it still captures extremely well the *spots* responsible for most of the sales. Although it is important to notice that the success of the new model is highly correlated with the nature of the product, since it is a product that is highly correlated with the area of residency of the consumer.

The work developed, and the results obtained during this MFW have permitted PSE to confidently move forward with new projects, allowing them to implement models such as the one developed in new projects for different locations and products. There are, however, some limitations that should be carefully analysed when implementing this model to other territories.

The results obtained with this model are linked to the socioeconomic data that is directly associated with the Portuguese reality. Purchasing Power, Education and Age Group distribution can form sets of the population with very different weights from other countries, and these same groups may also have a different behaviour regarding mobility/consumption. Additionally, we also need to consider that mobility data is high frequency data whereas demographic data is not, and therefore it will not capture the same variations as mobility data.

The Portuguese territory also registers a large concentration of its POIs. There tends to be a great centrality of services in Portugal. Whether it is in large cities or smaller ones, national urban planning has promoted an agglomeration of the main points of interest.

This agglomeration, which justifies the importance of these variables in the model for the national territory, will have to be analysed when applied to other countries that may not register the same agglomeration of POIs.

Lastly, we need to consider the size of the territory in question. The fact that Portugal is a rather small country made it easier for to address certain aspects of the model. Due to the size of the country, there are quite a few behavioural similarities across the whole population. This similarity may not exist for larger countries. It is also necessary to consider that, in the case of Portugal, it was feasible to divide the territory into squares of 300 by 300 meters. For bigger countries, this may not be as feasible and alternatives will have to be found, such as dividing the territory into larger blocks by considering provinces, for instance.

REFERENCES

An Integrative Approach for Measuring Semantic Similarities using Gene Ontology - Scientific Figure on ResearchGate. (n.d.). Retrieved July 9, 2022, from https://www.researchgate.net/figure/The-flowchart-of-tabu-search-process-The-tabu-search-process-is-shown-step-by-step-in_fig3_270274600

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS European Conference Data Mining,* (pp. 182-185). Amsterdam .

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS. Retrieved from https://api.semanticscholar.org/CorpusID:59777418

Dokeroglua, T., Sevinc, E., Kucukyilmaz, T., & Cosar, A. (2019). A survey on new generation metaheuristic algorithms. *Computers & Industrial Engineering, 137*. doi:https://doi.org/10.1016/j.cie.2019.106040.

Erbıyık, H., Özcan, S., & Karaboğa, K. (2012). Retail store location selection problem with multiple analytical hierarchy process of decision making application in Turkey. *Procedia - Social and Behavioral Sciences, 58*, 1405-1414. doi:https://doi.org/10.1016/j.sbspro.2012.09.1125

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining To Knowledge Discovery in Databases. *AI Magazine, 17*(3), 37-54. doi:https://doi.org/10.1609/aimag.v17i3.1230

Formanék, T., & Sokol, O. (2022). Location effects: Geo-spatial and socio-demographic determinants of sales. *Journal of Retailing and Consumer Services, 66*, 102902. doi:https://doi.org/10.1016/j.jretconser.2021.102902

Gendreau, M., & Potvin, J.-Y. (2019). Tabu Search. In M. Gendreau, & J.-Y. Potvin (Eds.)*, Handbook of Metaheuristics. International Series in Operations Research & Management Science* (Vol. 272, pp. 37-55). Springer, Cham. doi:https://doi.org/10.1007/978-3-319-91086-4_2

Glover, F. (1986). Future Paths for integer programming and links to artificial intelligence. *Computers & Operations Research, 13*(5), 533-549. doi:https://doi.org/10.1016/0305-0548(86)90048-1

Glover, F. (1989). Tabu Search - Part I. *ORSA Journal on Computing, 1*(3), 190-206. doi:https://doi.org/10.1287/ijoc.1.3.190

Glover, F., Taillard, E., & Werra, D. d. (1993). A user's guide to tabu search. *Annals of Operations Research, 41*, 3-28. doi:https://doi.org/10.1007/BF02078647

Hammouda, K., & Karray, F. (2000, 1). *A comparative study of data clustering techniques.* University of Waterloo, Ontario, Canada, Department of Systems Design Engineering.

Hand, D. J. (1998). Data Mining: Statistics and More? *The American Statistician, 52*(2), 112-118. doi:https://doi.org/10.1080/00031305.1998.10480549

Hanson, S. (1982). The Determinants of Daily Travel-Activity Patterns: Relative Location and Sociodemograpic Factors. *Urban Geography, 3*(3), 179-202. doi:https://doi.org/10.2747/0272-3638.3.3.179

Jiang, R., Song, X., Fan, Z., Xia, T., Wang, Z., Chen, Q., . . . Shibasaki, R. (2020). Transfer Urban Human Mobility via POI Embedding over Multiple Cities. *ACM/IMS Transactions on Data Science, 2*, 4-26. doi:https://doi.org/10.1145/3416914

Kang, C.-D. (2016). Spatial access to pedestrians and retail sales in Seoul, Korea. *Habitat International, 57*, 110-120. doi:https://doi.org/10.1016/j.habitatint.2016.07.006

Lin, G., Chen, X., & Liang, Y. (2018). The location of retail stores and street centrality in Guangzhou, China. *Applied Geography, 100*, 12-20. doi:https://doi.org/10.1016/j.apgeog.2018.08.007.

Liu, Y., Liu, C., Lu, X., Teng, M., Zhu, H., & Xiong, H. (2017). Point-of-Interest Demand Modeling with Human Mobility Patterns. *23rd ACM SIGKDD International Conference,* (pp. 13-17). Halifax, NS, Canada. doi:https://doi.org/10.1145/3097983.3098168

Pappalardo, L., Pedreschi, D., Smoreda, Z., & Giannotti, F. (2015). Using Big Data to study the link between human mobility and socio-economic development. *IEEE International Conference on Big Data*, (pp. 871-878). Santa Clara, CA, USA. doi:10.1109/BigData.2015.7363835

Pirim, H., Bayraktar, E., & Eksioglu, B. (2008). Tabu Search: A Comparative Study. In *Tabu Search.* doi:10.5772/5637

Prajapati, V. K., Jain, M., & Chouhan, L. (2020). Tabu Search Algorithm (TSA): A Comprehensive Survey. *3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)* (pp. 1-8). Jaipur, India: IEEE. doi:10.1109/ICETCE48199.2020.9091743

PSE PANEL_OOH. (2022, May 10). Retrieved from PSE: https://www.pse.pt/

Romanycia, M. H., & Pelletier, F. J. (1985). What is a heuristic? *Computational Intelligence, 1*(1), 47-58. doi: https://doi.org/10.1111/j.1467-8640.1985.tb00058.x

Schiff, M. (1960). The Sales Territory As a Fixed Asset. *Journal of Marketing, 25*(2), 51-53. doi:https://doi.org/10.1177/002224296002500209

Thuraisingham, B. (2000). A primer for understanding and applying data mining. *IT Professional, 2*(1), 28-31. doi:10.1109/6294.819936

Tkaczynski, A. (2017). Segmentation Using Two-Step Cluster Analysis. In T. Dietrich, S. Rundle-Thiele, & K. Kubacki (Eds.), *Segmentation in Social Marketing.* Singapore: Springer. doi:https://doi.org/10.1007/978-981-10-1835-0_8

Wang, F., Chen, C., Xiu, C., & Zhang, P. (2014). Location analysis of retail stores in Changchun, China: A street centrality. *Cities, 41*, 54-63. doi:https://doi.org/10.1016/j.cities.2014.05.005

Wendler, T., & Gröttrup, S. (2016). *Data Mining with SPSS Modeler.* Springer Cham. doi:https://doi.org/10.1007/978-3-319-28709-6

Wirth, R., & Hipp, J. (2000). Crisp-dm: towards a standard process model for data mining. *4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, (pp. 29-40).

Yaman, T. T. (2021). Segmenting Potential Customers with Kohonen Network: A Banking Sector Case Study. In A. M. Al-Sartawi, A. Razzaque, & M. M. Kamal (Eds.), *Artificial Intelligence Systems and the Internet of Things in the Digital Era - Proceedings of EAMMIS 2021* (Vol. 239, pp. 300-3012). Springer. doi:10.1007/978-3-030-77246-8_29

Zhang, S., Zhang, C., & Yang, Q. (2010). Data preparation for data mining. *Applied Artificial Intelligence. 17,* pp. 375-381. Taylor & Francis. doi:https://doi.org/10.1080/713827180

Zhao, K., Tarkoma, S., Liu, S., & Vo, H. (2016). Urban Human Mobility Data Mining: An Overview. *4th IEEE International Conference on Big Data* (pp. 1911-1920). Washington: Institute of Electrical and Electronics Engineers Inc. doi:10.1109/BigData.2016.7840811

Zoltners, A. A., & Lorimer, S. E. (2000). Sales Territory Alignment: An Overlooked. *Journal of Personal Selling & Sales Management, 20*(3), 139-150. doi:10.1080/08853134.2000.10754234

APPENDIX

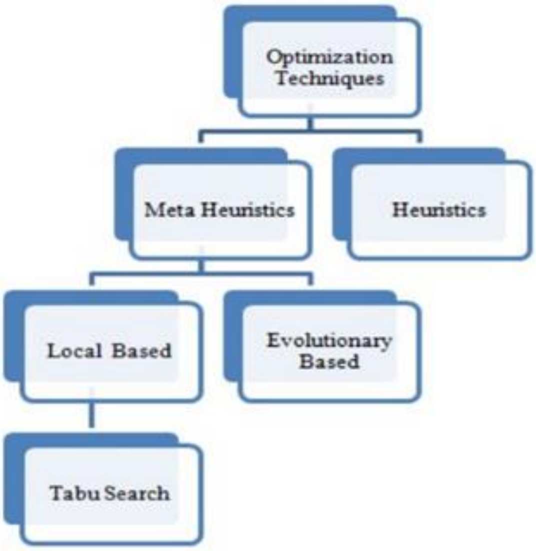


Figure 1 – Classification of Optimisation Techniques (Source: (Prajapati, Jain, & Chouhan, 2020))

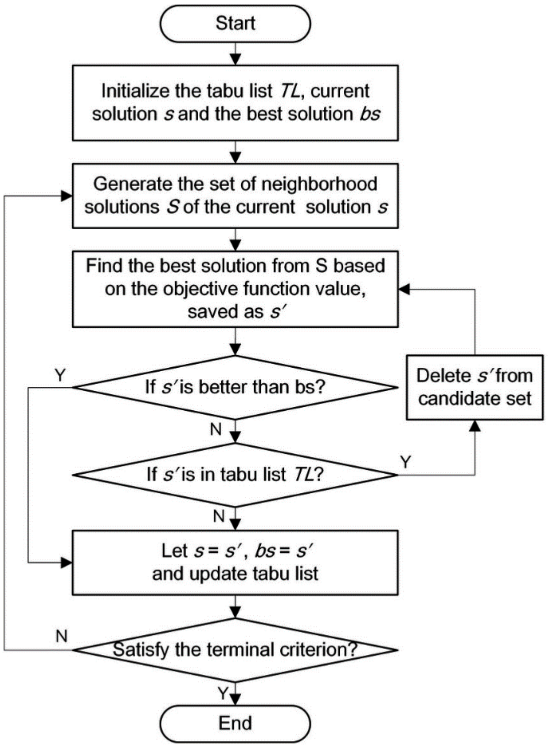

Start
Initialize the tabu list TL, current solution s and the best solution bs
Generate the set of neighborhood solutions S of the current solution s
Find the best solution from S based on the objective function value, saved as s′
If s′is better than bs?
Y
N
If s′is in tabu list TL?
Y
N
Delete s′from candidate set
Let s = s′, bs = s′ and update tabu list
Satisfy the terminal criterion?
N
Y
End

Figure 2 – Flowchart of the Tabu Search Process (Source: An Integrative Approach for Measuring Semantic Similarities using Gene Ontology - Scientific Figure on ResearchGate)

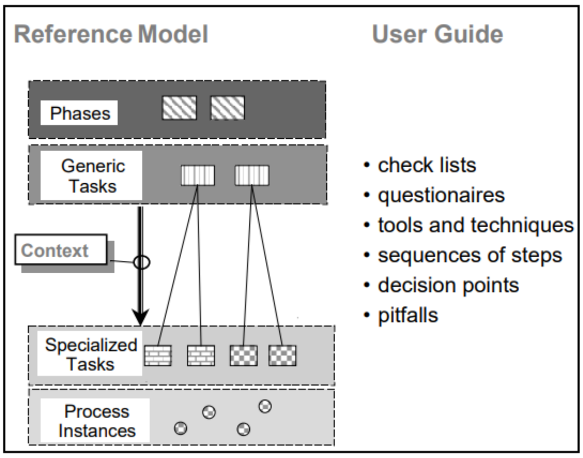


Figure 3 – Four Level Breakdown of CRISP-DM (Wirth & Hipp, 2000)

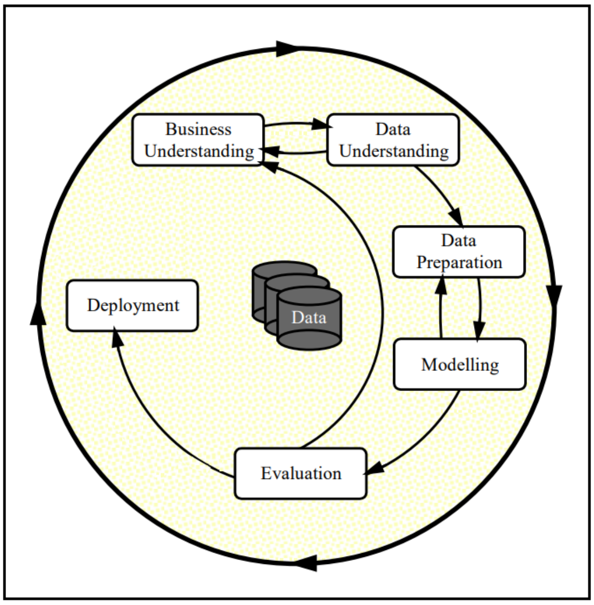
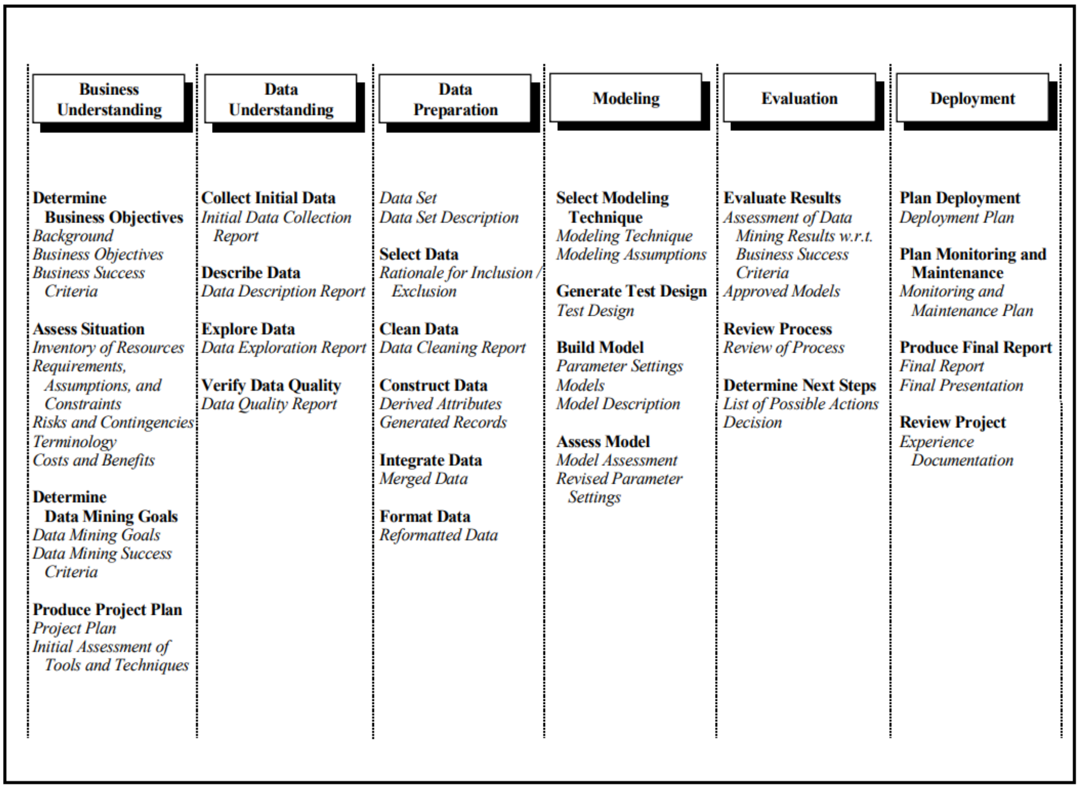


Figure 4 – Phases of the CRISP-DM Process Model (Wirth & Hipp, 2000)



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 5 - Overview of the CRISP-DM tasks and outputs (Source: PSE)

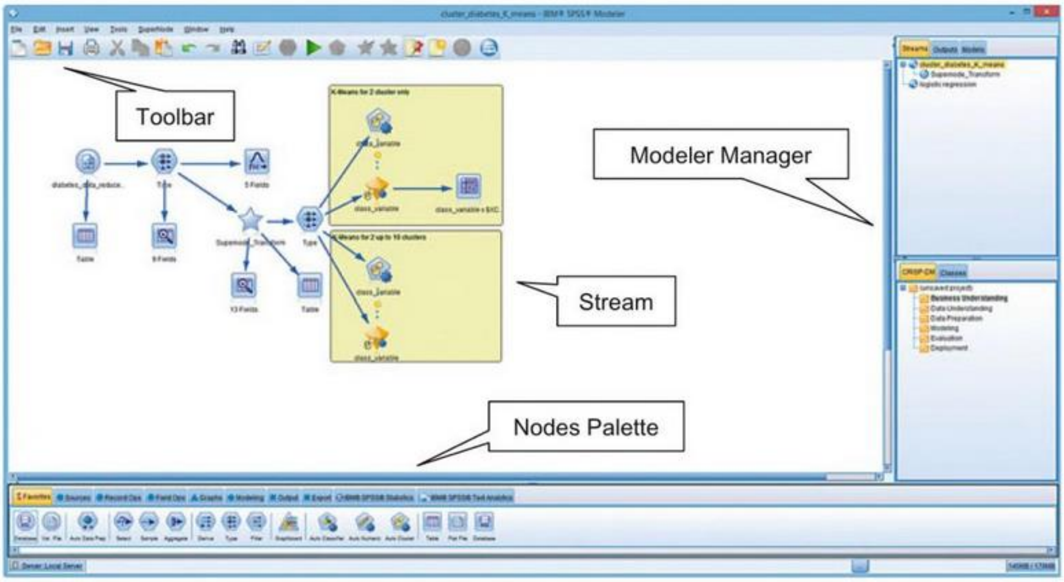
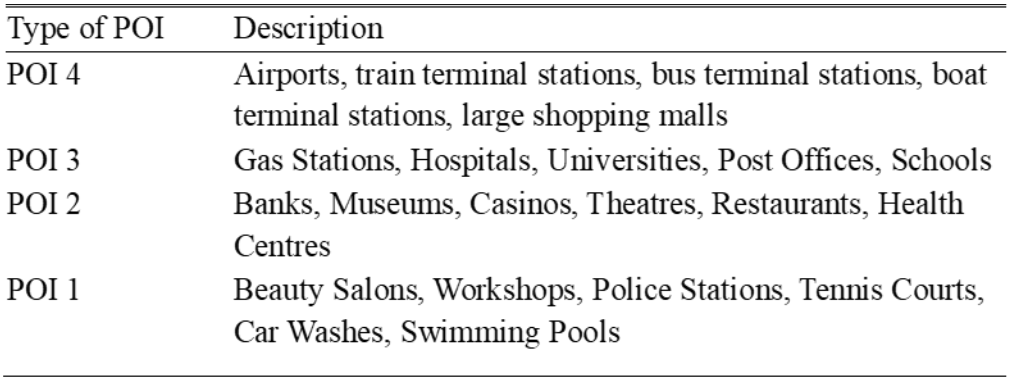


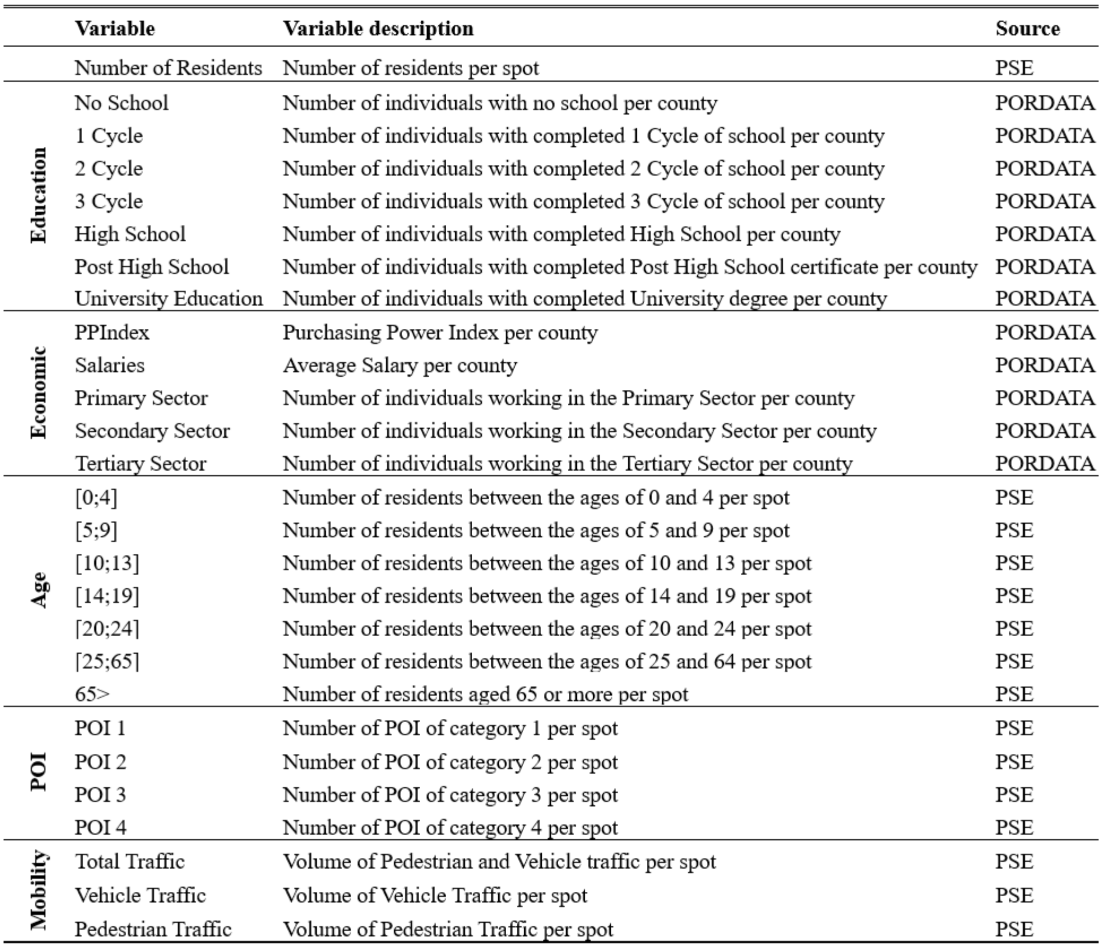
Figure 6 – IBM SPSS Modeler Interface (Wendler & Gröttrup, 2016)

Table I: Description of POI type



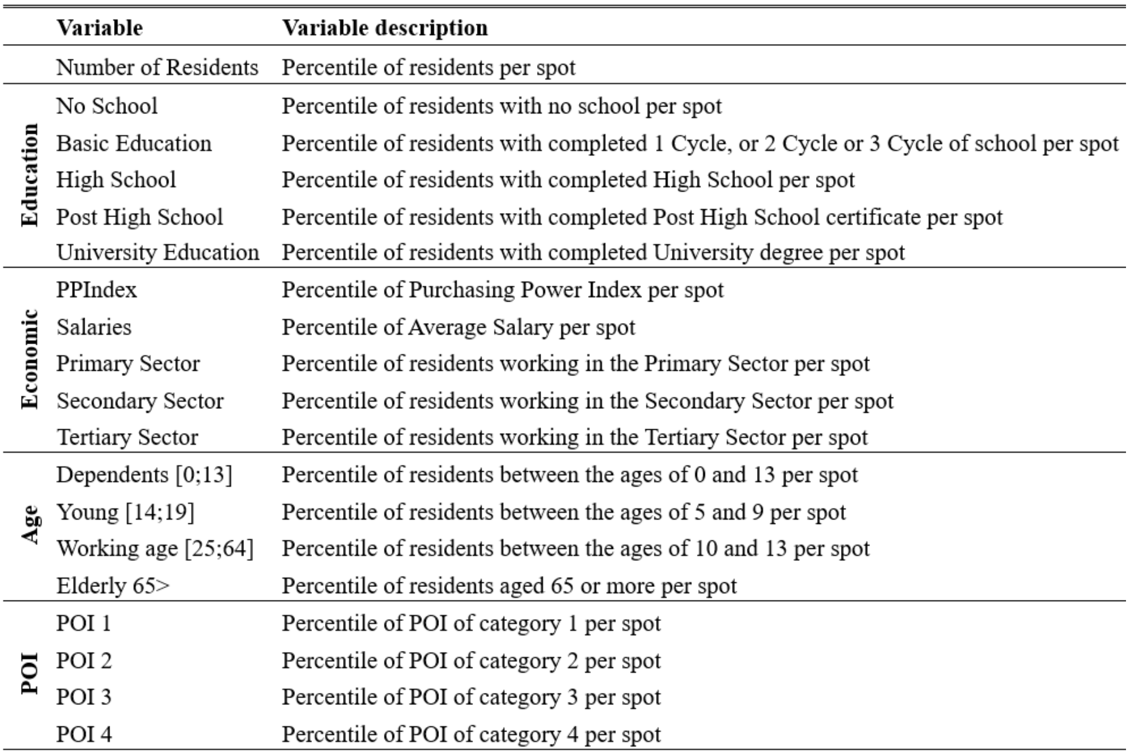
Type of POI	Description
POI 4	Airports, train terminal stations, bus terminal stations, boat terminal stations, large shopping malls
POI 3	Gas Stations, Hospitals, Universities, Post Offices, Schools
POI 2	Banks, Museums, Casinos, Theatres, Restaurants, Health Centres
POI 1	Beauty Salons, Workshops, Police Stations, Tennis Courts, Car Washes, Swimming Pools

Table II: Initial list of Variables and Description



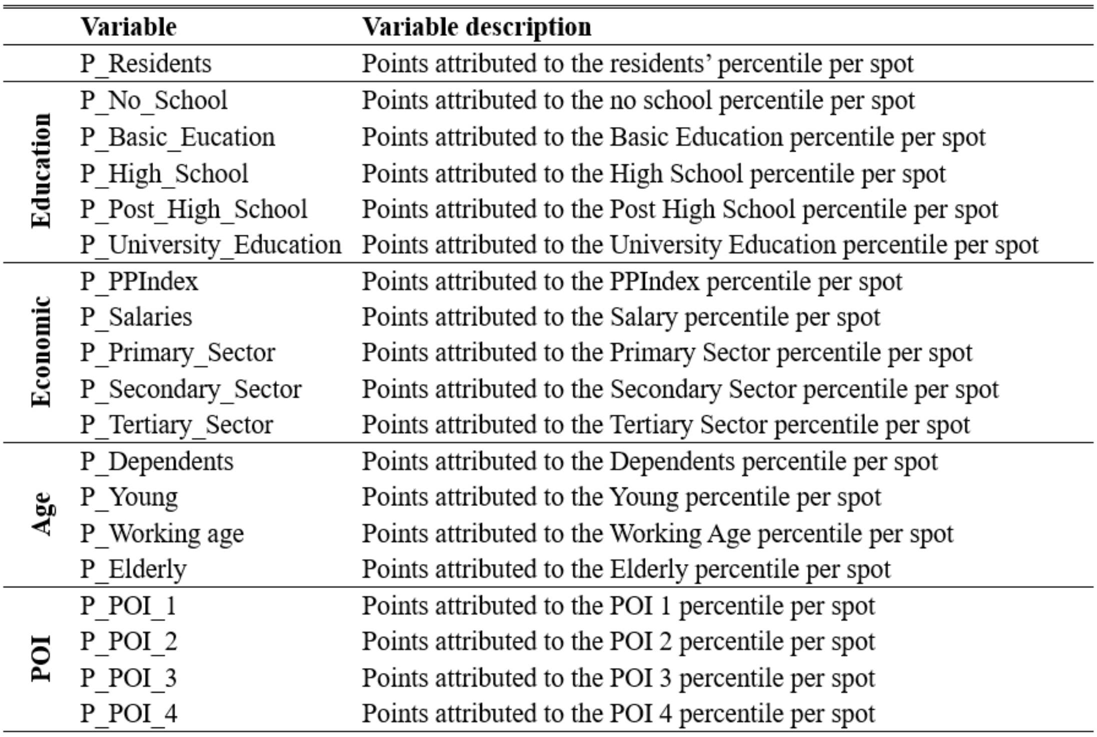
	Variable	Variable description	Source
	Number of Residents	Number of residents per spot	PSE
Education	No School	Number of individuals with no school per county	PORDATA
	1 Cycle	Number of individuals with completed 1 Cycle of school per county	PORDATA
	2 Cycle	Number of individuals with completed 2 Cycle of school per county	PORDATA
	3 Cycle	Number of individuals with completed 3 Cycle of school per county	PORDATA
	High School	Number of individuals with completed High School per county	PORDATA
	Post High School	Number of individuals with completed Post High School certificate per county	PORDATA
	University Education	Number of individuals with completed University degree per county	PORDATA
Economic	PPIndex	Purchasing Power Index per county	PORDATA
	Salaries	Average Salary per county	PORDATA
	Primary Sector	Number of individuals working in the Primary Sector per county	PORDATA
	Secondary Sector	Number of individuals working in the Secondary Sector per county	PORDATA
	Tertiary Sector	Number of individuals working in the Tertiary Sector per county	PORDATA
Age	[0;4]	Number of residents between the ages of 0 and 4 per spot	PSE
	[5;9]	Number of residents between the ages of 5 and 9 per spot	PSE
	[10;13]	Number of residents between the ages of 10 and 13 per spot	PSE
	[14;19]	Number of residents between the ages of 14 and 19 per spot	PSE
	[20;24]	Number of residents between the ages of 20 and 24 per spot	PSE
	[25;65]	Number of residents between the ages of 25 and 64 per spot	PSE
	65>	Number of residents aged 65 or more per spot	PSE
POI	POI 1	Number of POI of category 1 per spot	PSE
	POI 2	Number of POI of category 2 per spot	PSE
	POI 3	Number of POI of category 3 per spot	PSE
	POI 4	Number of POI of category 4 per spot	PSE
Mobility	Total Traffic	Volume of Pedestrian and Vehicle traffic per spot	PSE
	Vehicle Traffic	Volume of Vehicle Traffic per spot	PSE
	Pedestrian Traffic	Volume of Pedestrian Traffic per spot	PSE

Table III: List of Percentile Variables



	Variable	Variable description
	Number of Residents	Percentile of residents per spot
Education	No School	Percentile of residents with no school per spot
	Basic Education	Percentile of residents with completed 1 Cycle, or 2 Cycle or 3 Cycle of school per spot
	High School	Percentile of residents with completed High School per spot
	Post High School	Percentile of residents with completed Post High School certificate per spot
	University Education	Percentile of residents with completed University degree per spot
Economic	PPIndex	Percentile of Purchasing Power Index per spot
	Salaries	Percentile of Average Salary per spot
	Primary Sector	Percentile of residents working in the Primary Sector per spot
	Secondary Sector	Percentile of residents working in the Secondary Sector per spot
	Tertiary Sector	Percentile of residents working in the Tertiary Sector per spot
Age	Dependents [0;13]	Percentile of residents between the ages of 0 and 13 per spot
	Young [14;19]	Percentile of residents between the ages of 5 and 9 per spot
	Working age [25;64]	Percentile of residents between the ages of 10 and 13 per spot
	Elderly 65>	Percentile of residents aged 65 or more per spot
POI	POI 1	Percentile of POI of category 1 per spot
	POI 2	Percentile of POI of category 2 per spot
	POI 3	Percentile of POI of category 3 per spot
	POI 4	Percentile of POI of category 4 per spot

Table IV: List of Point Variables (Final Variables used)



	Variable	Variable description
	P_Residents	Points attributed to the residents' percentile per spot
Education	P_No_School	Points attributed to the no school percentile per spot
	P_Basic_Eucation	Points attributed to the Basic Education percentile per spot
	P_High_School	Points attributed to the High School percentile per spot
	P_Post_High_School	Points attributed to the Post High School percentile per spot
	P_University_Education	Points attributed to the University Education percentile per spot
Economic	P_PPIndex	Points attributed to the PPIndex percentile per spot
	P_Salaries	Points attributed to the Salary percentile per spot
	P_Primary_Sector	Points attributed to the Primary Sector percentile per spot
	P_Secondary_Sector	Points attributed to the Secondary Sector percentile per spot
	P_Tertiary_Sector	Points attributed to the Tertiary Sector percentile per spot
Age	P_Dependents	Points attributed to the Dependents percentile per spot
	P_Young	Points attributed to the Young percentile per spot
	P_Working age	Points attributed to the Working Age percentile per spot
	P_Elderly	Points attributed to the Elderly percentile per spot
POI	P_POI_1	Points attributed to the POI 1 percentile per spot
	P_POI_2	Points attributed to the POI 2 percentile per spot
	P_POI_3	Points attributed to the POI 3 percentile per spot
	P_POI_4	Points attributed to the POI 4 percentile per spot

	Spots	%
0. NO_POI	975059	94,48%
1. LOW_POI	41400	4,01%
2. MEDIUM_POI	8868	0,86%
3. HIGH_POI	6713	0,65%

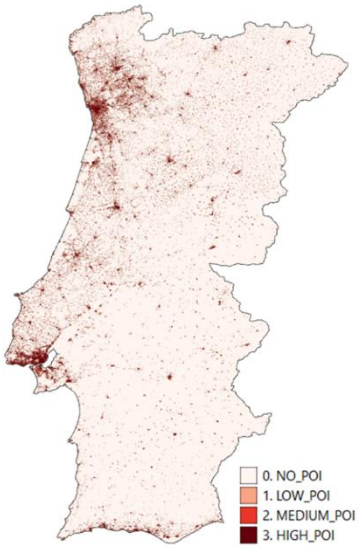


Table V: Total POI distribution (Left)

Figure 9– Map of the POI distribution (Right)

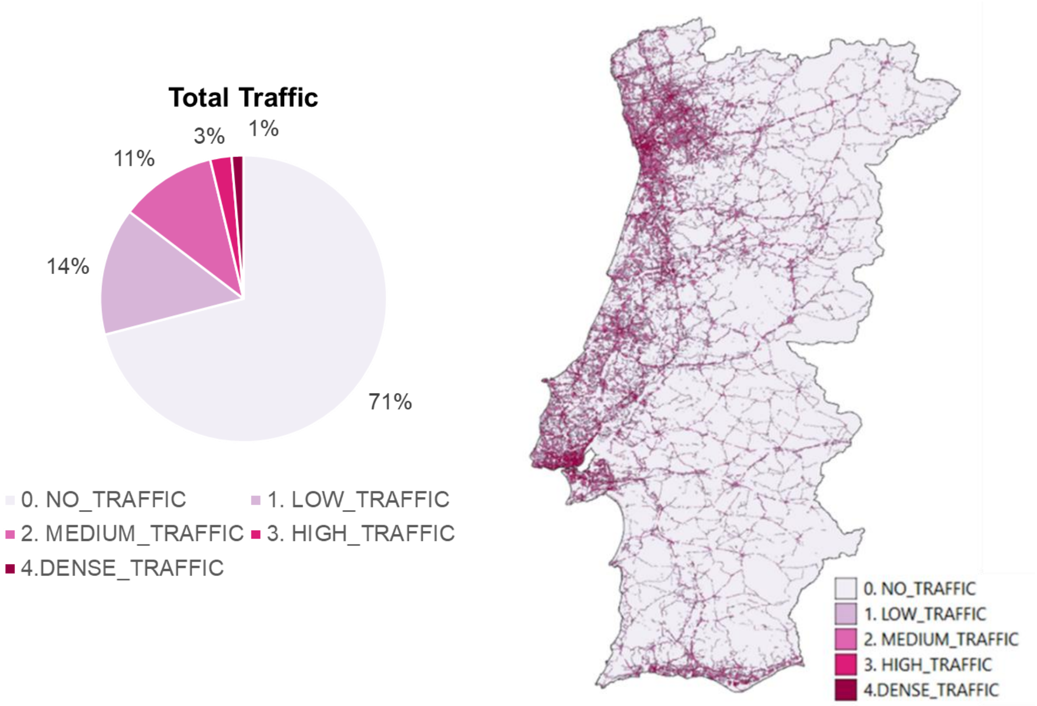


Figure 10 – Pie Chart - Total Traffic % (by category) (Left); Map of Total Traffic by Category (Right)

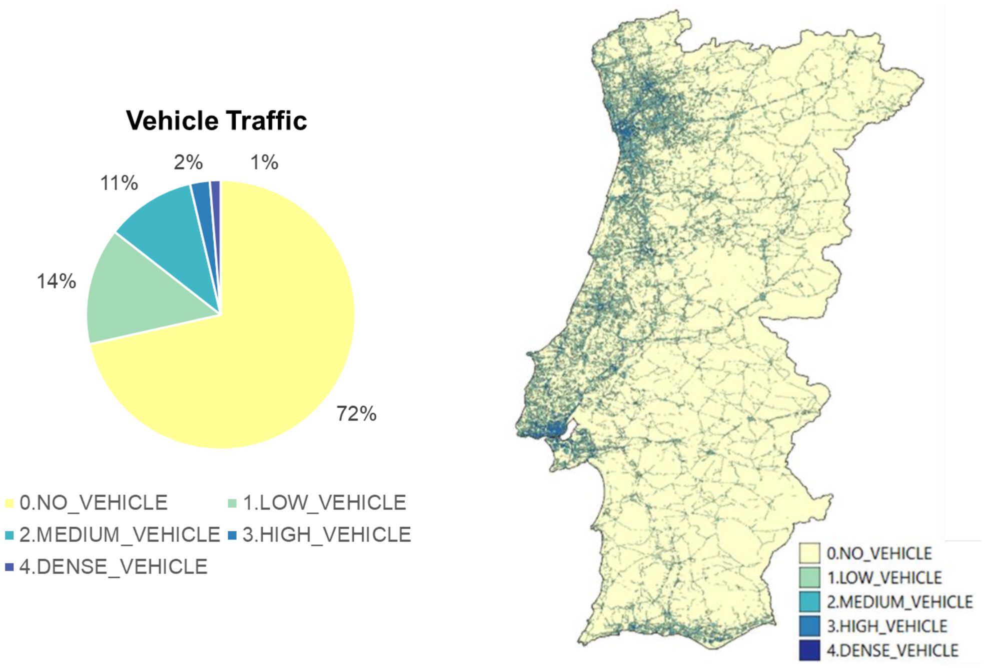


Figure 11 – Pie Chart – Vehicle Traffic % (by category) (Left); Map of Vehicle Traffic by Category (Right)

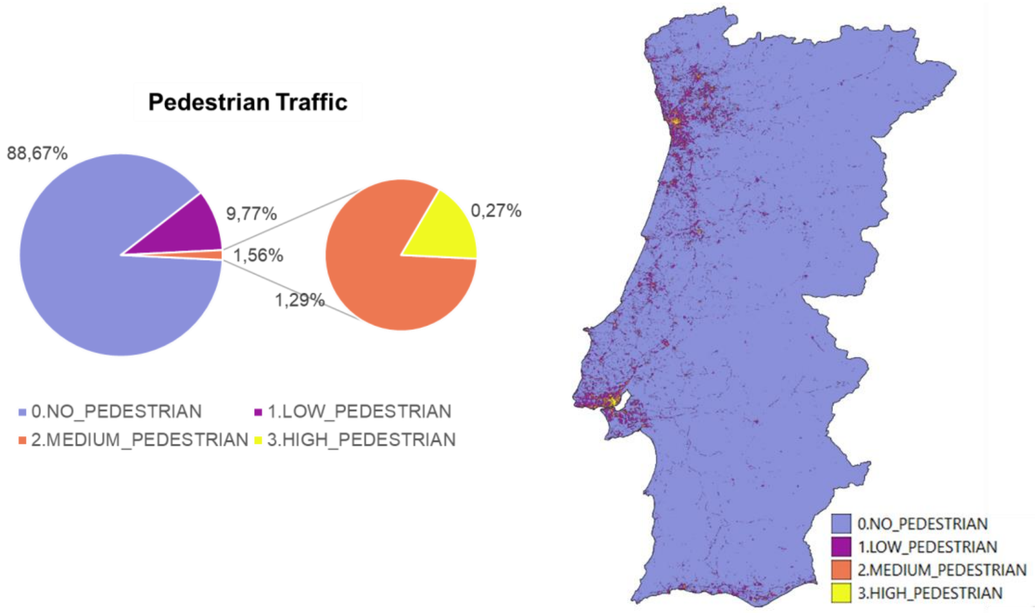


Figure 12 – Pie Chart – Pedestrian Traffic % (by category) (Left); Map of Pedestrian Traffic by Category (Right)

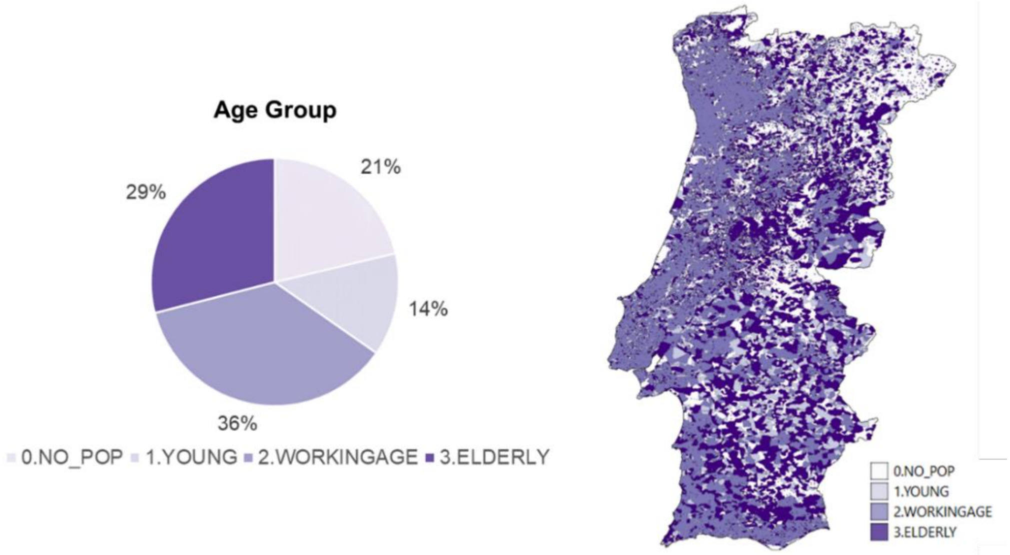


Figure 13 – Pie Chart – Age group % (by category) (Left); Map of Age Group by Category (Right)

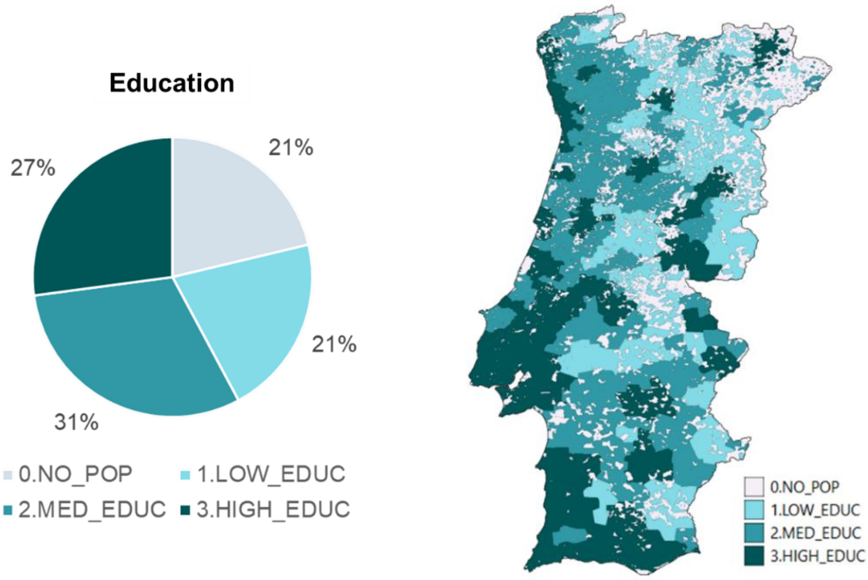


Figure 14 – Pie Chart – Education Level % (by category) (Left); Map of Education Level by category (Right)

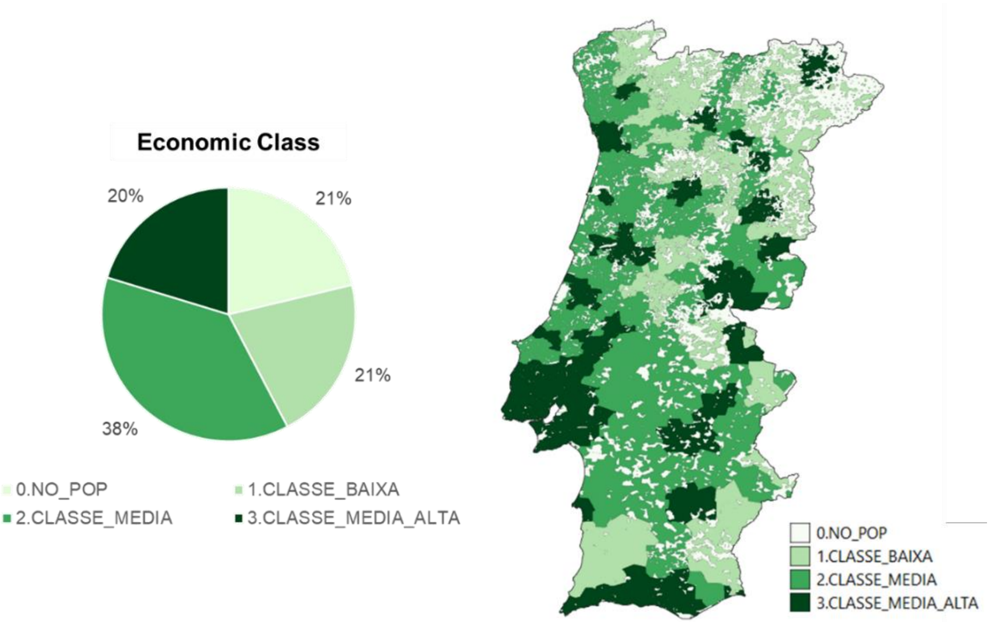


Figure 15 – Pie Chart – Economic Class % (by category) (Left); Map of Economic Class by category (Right)

Data Appendix

Resident population aged 15 years and over according to the Census: total and by highest level of complete education; https://www.pordata.pt/municipios/populacao+residente+com+15+e+mais+anos+segundo+os+censos+total+e+por+nivel+de+escolaridade+completo+mais+elevado-69, November 29, 2022.

Average monthly earnings of employees: total and by sex; https://www.pordata.pt/Municipios/Ganho+m%c3%a9dio+mensal+dos+trabalhadores+por+conta+de+outrem+total+e+por+sexo-282 , November 29, 2022.

Per capita purchasing power; https://www.pordata.pt/municipios/poder+de+compra+per+capita-118, November 29, 2022.

Employed population according to the Census: total and by sector of economic activity;https://www.pordata.pt/municipios/populacao+empregada+segundo+os+censos+total+e+por+sector+de+actividade+economica-145-604, November 29, 2022.