FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Inferring Origin-Destination Matrices to Support Smart Mobility Solutions in Urban Networks

Fábio Caramelo

U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: José Pedro Tavares, PhD

Co-supervisor: Rosaldo J. F. Rossetti, PhD

July 25, 2019

Inferring Origin-Destination Matrices to Support Smart Mobility Solutions in Urban Networks

Fábio Caramelo

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Ana Paula Cunha da Rocha, PhD

External Examiner: Pedro José Ramos Moreira de Campos, PhD

Supervisor: José Pedro Maia Pimentel Tavares, PhD

July 25, 2019

Abstract

Through the last years, a shift of population towards urban areas can be observed, resulting in urban areas more densely populated. This increase in population density causes problems in areas like mobility. Every day the population has to commute, affecting the mobility of the areas due to the home and no home based trips. Therefore, it is imperative to counter the effects caused by this increase in population density in urban areas and consequent negative impacts.

Transportation networks can be modeled by a set of nodes and links that make the connection between the nodes, representing a transportation network. In this way, it is possible to apply transport models that allow planning transportation systems able to respond to the desired demand, usually represented by a static Origin-Destination Matrix. This matrix contains information about the trips between the areas of the network and by assigning these trips to the network, the flows on the network are obtained. On the other hand, dynamic O-D Matrices - matrices that vary over time - are used for traffic management and control and provide a valuable source of information for this purpose. Several methods to estimate these matrices have been proposed, as explained in this report, briefly explaining their approaches as well as their advantages and disadvantages.

Despite the wide variety of methods available for this estimation, none of these is able to estimate the evolution of these matrices throughout the years. To do this estimation, an artificial population will be created using data from the demographic census and other data sources. From this population, its' evolution will be simulated. By combining this evolution of the population with data collected from mobility surveys, it is expected to obtain an estimation for the evolution of those Origin-Destination Matrices.

By being able to predict the demand of the population, changes to the transport policies can be made in advance to match the needs of the transport needs. This possibility allows for the waste of resources to be avoided (by allowing to make alterations that meet the future demand), and, with good planning of these transport networks, the mobility of the population should not be affected, also reducing the economic cost and environmental impact of those trips.

Keywords - Smart Mobility; Smart Cities; Origin-Destination Matrices; Transportation Planning; Multi-Agent; Artificial Society;

Resumo

Nos últimos anos tem-se assistido a uma migração da população em direcção às áreas urbanas, resultando em áreas urbanas com uma forte densidade populacional. Este aumento de densidade tem os seus impactos como, por exemplo, dificuldades nas deslocações quer pendulares quer não pendulares, ou seja, afecta a mobilidade da população que se tem de deslocar todos os dias, seja para ir para o trabalho ou outras actividades quotidianas. É por isso imperativo que sejam tomadas iniciativas para diminuir os efeitos causados por este aumento de densidade populacional e consequentes impactos negativos.

A nível da oferta, as redes de transportes podem ser modeladas por um conjunto de nós e arcos, que fazem a ligação entre os mesmos, permitindo assim representar um sistema viário através de uma rede. Através desta representação podem ser aplicados os modelos de transporte que permitem o planeamento de sistema de transportes de modo a responder à procura esperada representada normalmente por uma matriz Origem-Destino estática. Esta matriz contém informação sobre o número de viagens entre as áreas da rede em questão, que sendo afectada à rede permite obter os fluxos de viagens da mesma. Por outro lado, as matrizes O-D dinâmicas - matrizes que variam ao longo do tempo - são utilizadas para controlo e gestão do tráfego e são uma fonte de informação valiosa neste âmbito. Existem vários métodos para estimar estas matrizes, conforme analisados neste trabalho, explicando brevemente as suas abordagens assim como as suas vantagens e desvantagens.

Apesar da variedade existente de métodos para estimar estas matrizes, nenhum destes tem a capacidade de fazer a previsão da evolução das matrizes ao longo dos anos. Para fazer esta estimativa, uma população artificial vai ser gerada usando dados dos censos demográficos, assim como de outras fontes de informação. A partir desta população, a sua evolução será então simulada. Combinando a evolução desta população com os dados recolhidos através de inquéritos de mobilidade, é esperado que seja possível estimar a evolução temporal destas matrizes Origem-Destino.

Ao ser possível prever as necessidades da população, podem ser efectuadas antecipadamente alterações ao sistema de transportes de modo a responder a esta procura, permitindo assim a avaliação de diferentes políticas estratégicas no âmbito dos transportes. Esta possibilidade permite evitar o desperdício de recursos (ao permitir fazer mudanças que vão de acordo às futuras necessidades), e, com um bom planeamento destas redes de transporte, melhorar a mobilidade da população, também reduzindo quer os custos económicos, quer o impacto ambiental destas viagens.

Palavras chave - Mobilidade inteligente; Cidades inteligentes; Matrizes Origem-Destino; Planeamento de Transportes; Multi-Agente; Sociedade Artificial

Acknowledgements

First, I would like to thank my family for their support throughout these years. This wouldn't be possible without them, to them I give my biggest gratitude for never denying me anything so that I could be who I am.

I also thank both my supervisors, Prof. Rosaldo Rossetti and Prof. José Pedro Tavares for all their help and guidance throughout this project. In particular, I would like to especially thank Dr. Nuno Biga, for his key insights into the topic and for sharing his practical experience in the application domain that eventually culminated in the definition of my dissertation topic. Their suggestions, comments, and invaluable inputs were ultimately important to the course of this work.

At last, I would like to thank my friends, especially from **random.org*, for all their help during these years and for all the incredible memories and moments.

Fábio Caramelo

“Wisdom comes from experience.
Experience is often a result of lack of wisdom.”

Terry Pratchett

Contents

List of Figures

LIST OF FIGURES

List of Tables

Abbreviations

AVI	Automatic Vehicle Identification
GPS	Global Positioning System
GSM	Global System for Mobile communications
LPR	License Plate Recognition
O-D	Origin-Destination
ODM	Origin-Destination Matrices
RFID	Radio-frequency identification

Chapter 1

Introduction

In this Chapter, the scope of this project will be introduced as well as the problems this project will try to solve. In the first Section (1.1), the area where this project is inserted will be described, Section 1.2 describes the motivation and goals for this project, and finally in Section 1.3 the way this report is structured and what each chapter contains is explained.

1.1 Context

Urban areas have become more densely populated throughout the years. In 1950 about 50% of the population on the European Union was living in urban areas but currently, there are more than 75% living in these areas. In 2050 it is expected that this number will increase to more than 80% [CDBN11]. Therefore the number of overcrowded locations increases, making it more difficult to move from one place to another. Because of this, it is fundamental to minimize the effects caused by this growth of population in cities.

Through the years, the concept of smart cities has been emerging. However, there is still no clear definition of this concept but [HBB^{+}00] defines one as "a city that monitors and integrates conditions of all of its critical infrastructures, including roads, bridges, tunnels, rails, subways, airports, seaports, communications, water, power, even major buildings, can better optimize its resources, plan its preventive maintenance activities, and monitor security aspects while maximizing services to its citizens.". A smart city can be seen as a city that monitors various services and uses the acquired information to improve its own services, as in this case, the mobility services. In order to minimize the impacts of location that gets denser each time, it is imperative to develop smart mobility solutions.

Every day people travel from one place to another and these trips can be by car, public transports or other means of transportation possible in a transportation network. This network is defined by a set of nodes and links that make the connection between the nodes. This representation is the starting point to transportation planning and forecasting and one of its objectives is to estimate the flow in the transportation links (resulting from the assignment of trips). Examples of transportation forecasting methods are reported elsewhere [BAR15, SRM+16, AAA+17, AFR19]. Such methods are greatly used to support transportation analysis approaches, which traditionally follow the well-known Four-Step Model [McN00]; this method is divided into four steps: trip generation, trip distribution, mode split, and route assignment. This project aims to contribute to the first two steps of this model: trip generation, where the number of trips to and for each zone is calculated, as well as trip distribution, where the matching of the number of trips between each pair of origins and destinations is calculated.

This planning and forecasting are of great importance due to the impact that mobility has on peoples' lives. A poorly designed network can, for instance, lead to delays in commutes if the current network has no capacity to meet the demand and ultimately increase the pollution caused by these trips. A fundamental step to provide an adequate network is to understand the actual and future population' mobility patterns and making these patterns clear and easy to understand to the policymakers who have the opportunity to respond to the needs.

1.2 Motivation and Goals

The main goal of this project is to infer Origin-Destination Matrices and their evolution over time by combining information collected using surveys regarding the populations' mobility patterns and population social-economic data. An artificial population will be initialized using data from the demographic census and its evolution simulated. Consequently, the evolution of the Origin-Destination Matrices will also be simulated. This should allow decision-makers to know in advance for a medium-term the needs of the population. For this project, the population of the parish of *Pombal* will be used as a case study.

Additionally, this simulation can be used to test potential changes on the territory (such as the changes in land uses) or other events that can impact the mobility patterns.

1.3 Report Structure

Chapter 2 contains the state of the art divided into three sections: the first section explores and analyzes the various methods to collect traffic data, the second section will describe the several proposed methods to infer origin-destination matrices and finally, the third illustrates the concept of artificial societies as well as some practical applications. Chapter 3 describes the approach that will be taken for the development of this project, and Chapter 4 discusses the case study for this method as well as the results of the performed tests. At last, Chapter 5 presents the conclusions of the present research work: these will make an overview and the main contributions of this dissertation are presented, as well as some ideas for further development.

Chapter 2

Literature Review

This chapter explores the state of the art in the context of this project. Section 2.1 will discuss the first step, the data collection by enumerating and explaining the main techniques used to collect data from the traffic as well as the metrics collected by each one. Section 2.2 starts by giving an overview of origin-destination matrices and the following subsections explore some methods of estimation for static and dynamic matrices, respectively. Section 2.3 informs about artificial societies and some of its practical applications, and finally, Section 2.4 summarizes the previous sections and identifies the problem to be solved.

2.1 Traffic Data Collection

Through the years, different methods and approaches of acquiring traffic data have been tested. It is fundamental to collect accurate data in order to apply it to transportation studies.

These techniques can be separated into two groups: one where only information about the flow is captured (number of vehicles) and on the second group information about which vehicles cross each point along a cordon or screen line is recorded. This second group allows to identify the vehicles and is entitled Automatic Vehicle Identification (AVI). These techniques usually provide more information because if enough sensors are deployed through the network, it is possible to reconstruct the path followed by each vehicle allowing a better perception of the mobility patterns [SRM+16].

In the last years, technology has propagated into a wider variety of devices: cars have Bluetooth, GPS and some have an internet connection; smartphones are everywhere and also include these technologies. This quick spread of this wide availability provides a good opportunity to explore these methods; however, privacy needs to be taken into account since it should not be possible to identify one person based on the collected data.

2.1.1 Manual Traffic Counts

Traffic surveys are conducted by people that stand by the side of a road and record information about the traffic situation in that road, like the number of cars in each way. These surveys have some clear disadvantages such as the time to complete and the high cost to perform. Since this technique requires at least one person to be standing on each link to analyze for a period of time big enough to capture information that results in reliable information about a given road. Furthermore, since it is a task that is not automated it requires manpower. With the ability to capture and collecting images of these networks, these problems can be mitigated but this method is not cost effective neither quick.

Because of the high cost and the high time interval required to obtain information, the periodicity of these surveys tends to be low which may result in outdated information.

2.1.2 Automatic Traffic Counts

A common way to collect traffic data is by placing induction loops on the links to analyze, such as the one represented in Figure 2.1. These loops are deployed on the road and are able to detect vehicles that cross that section by the change produced by the vehicle in the loop. This system is able to count the number of vehicles that cross a given point as well as the speed (aggregate average speed for a given time interval) of those vehicles. Furthermore, some loops are also able to classify the type of vehicle.

However, these loops are known for not being reliable, especially if the correct maintenance is not provided. According to [HWH+10], every day in California, 30% of the existing sensors in the network do not work properly. This is something to take into consideration when evaluating the data because the process to extract information from this data should be able to minimize the effects provoked from this missing data.

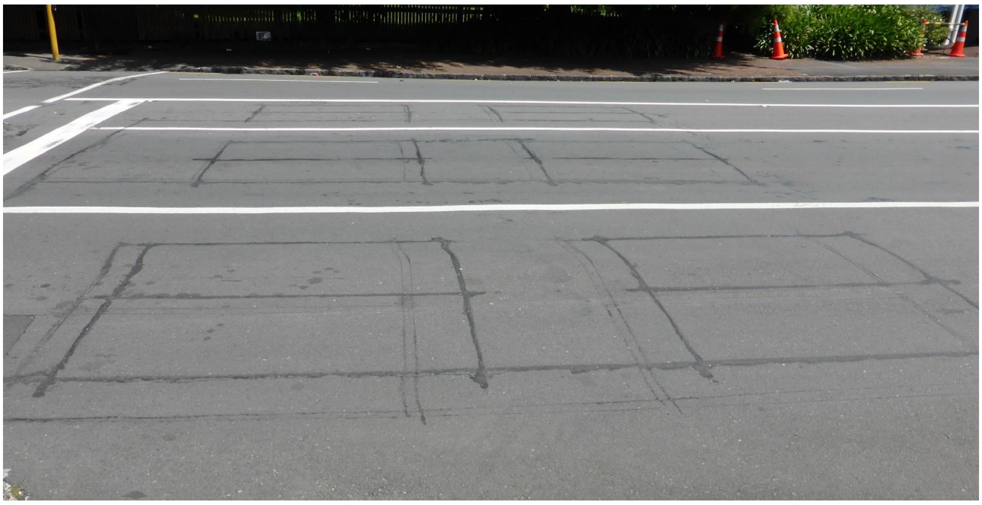

Figure 2.1: Example of an induction loop [1].

2.1.3 Bluetooth

The growth of Bluetooth enabled devices has proven to be a good method of acquiring traffic data. Nowadays a lot of smartphones and even cars are equipped with Bluetooth antennas making them sources of data for this purpose when these antennas are enabled. Bluetooth is an AVI technique since it allows to track individual vehicles throughout the network if enough scanners are deployed [FRK+14].

These sensors are proven to be cost-effective and the detection is anonymous (this is possible by hashing the unique identifier) [MNC+14]. Sensor placement is important, otherwise overlapping can occur. This overlapping can lead to an incorrect order of detection (causing one detection to be followed by a non-adjacent point) of the path although this can be easily detected and corrected. In more complex networks a sensor can detect a vehicle traveling in another corridor other than the target one [MNC+14].

Bluetooth scanners have different signal strength which might lead to some devices more likely to be detected on one zone than others. Furthermore, physical obstacles such as billboards or even weather affect the signal which will change the reliability of the data collected. [MNC+14] Signal interference also takes part in the data signal: areas that are more crowded may change the effectiveness of detection. Furthermore, since the use of Bluetooth enabled devices (especially cars) only represents part of the traffic, it is possible that this data has a particular social-economic bias. [MNC+14]

Although these types of sensors require some pre-processing of the collected data, this source is capable of producing good results. This method allows recording the locations crossed by a device as well as the time of that crossing. By knowing the distance between the points where the sensors are located, it is possible to obtain an estimation of the speed between these points.

This source has been tried for producing dynamic OD-Matrices on [BMB+12] and has proven to provide good estimations.

2.1.4 RFID

RFID - Radio-frequency identification allows automatic identification of the object that is attached to the RFID tag. This tag contains a unique identifier that allows a car to be identified each time that crosses a reader on the network. Each time this happens, the identifier can be stored as well as the time the reading occurred, allowing the calculation of the average speed between the two points where the detections happened.

[Fin10] discussed some characteristics of this technology and mentions that this technology can read the identifier with a distance between the identifier and the reading up to 5 meters, the reading time is fast (about 0.5 seconds) and that optical cover has no influence on the readings, as well as dirt. Despite the promising features, the relatively short range of detection make the implementation of this systems more difficult: the readers must be placed, for example, in gates

[1] https://www.drivingtests.co.nz/resources/traffic-lights-change/ - accessed on January 29, 2019

over the roads which might require some investment. The price of the readers is also something to take into consideration since these are not cheap.

In Portugal RFID is used for toll gates on highways with "Via Verde" (as shown on Figure 2.2) and perhaps this method could provide good information if these tags could be used for other purposes and to obtain other data like which cars pass a given point and the average speed between those.

Toll tags have already been studied to provide information about travel times between some points on [WD01] and mention that if enough tags are used it is possible to obtain good information. [Wen10] has also tried to use RFID tags to monitor travel times and mentions that for detecting cars at higher speeds, active tags should be used (active tags require its own power source) instead of passive tags (where no power source is required).

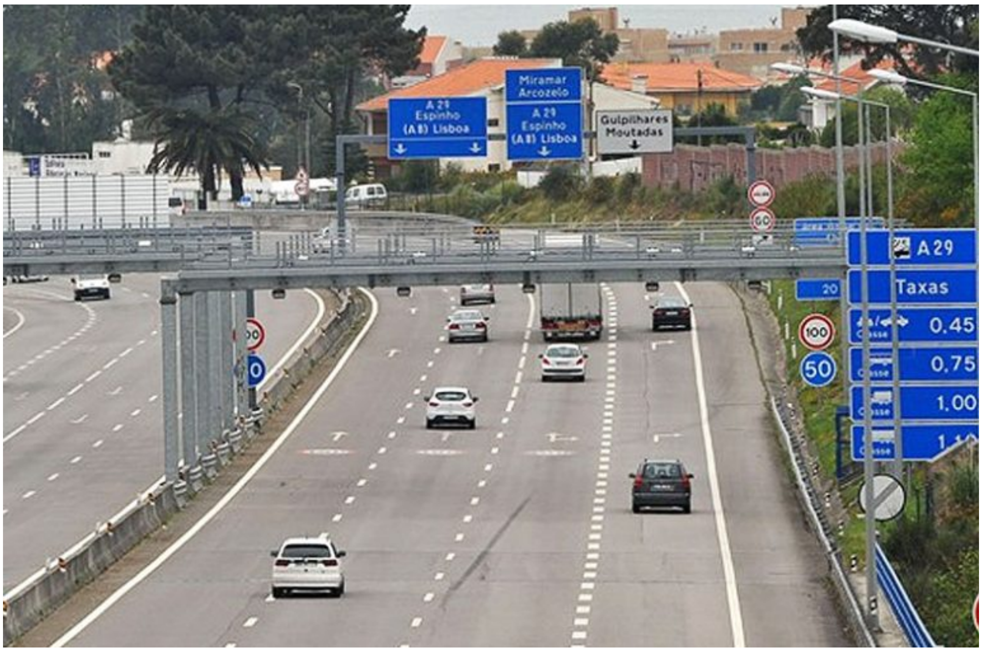

Figure 2.2: Portuguese tolls that use RFID for vehicle identification [2].

2.1.5 Mobile Phone Location

Nowadays, phones are part of almost everyone's life. This raises a big opportunity to collect information about population mobility patterns. It has been tried by [CLLR11] to collect information about individuals' location each time a connection is made to a cell tower, can this be a call (placed or received), SMS or a network connection. This allows recording a user location anonymously each time a network event is triggered - making the interval between these events important when trying to obtain an accurate location.

[2] https://jpn.up.pt/2017/01/24/aberta-via-evitar-portagem-na-a29/ - accessed on February 6, 2019

Due to the way the position is calculated, the uncertainty is considerable - averaging 320 meters - which might be something to consider when applying this method to complex networks. Furthermore, when an antenna is heavily loaded, the connection can be handed to another antenna that is less loaded, possibly introducing an even bigger error. [CLLR11]

When comparing the results obtained from [CLLR11] with data from the census, it was possible to obtain a good estimation of the users' home locations density with only 4.3% of the population being monitored. The obtained O-D Matrices from this study also revealed promising when compared to other matrices calculated with other data sources.

The wide adoption of mobile phones and in a world each time more dependent on constant connections to the internet, this method can prove to be very useful. With only 4.3% of the population being used for this test it was possible to obtain valuable information by allowing a temporal analysis of the mobility patterns and perhaps with more frequent position updates an even better data could be obtained.

An experiment to compare the results obtained when estimating O-D Matrices using surveys *versus* using mobile phone data was conducted on [TPP17] as well as some properties of those methods. The authors mention that if no other data is available other than the mobile phone data, this source can provide a good estimate of the matrix. Similar research is also conducted on [TÁD15] that also mentions that if the obtained data is pre-processed this can be used for this purpose on highways.

Further experiments using data collected from cellphones were tested as in [ICWG14] that only used Call Detail Records as the data source and has shown potential. [SK08] was able to obtain good results using location data from a mobile cellular network but still obtained worse results than when compared to AVI methods using point detectors.

2.1.6 GPS

With the wide adoption of smartphones, nowadays GPS devices are everywhere since almost every smartphone is GPS enabled. GPS - Global Positioning System provides great location accuracy and their shrinking costs is making the adoption of this technology possible to a great part of the population.

As discussed on [HWH+10], the data acquired from these devices can be sampled in two ways: the data can be collected at time intervals regardless the position of the device or these devices can be used to report their information when they are at some location, as it happens on other AVI techniques. Although the first option is preferable, some aspects need to be taken into consideration like privacy, communication load, and energy consumption.

The same research concluded that with only 2/3% penetration rate on the population, the authors were able to obtain a good estimation on the speed at the studied highways when comparing with the ones obtained from the loops, despite the fact that those were being calculated in different ways – space and time mean speed.

Collected data from GPS equipped vehicles has also been used to estimate traffic volume by [CC18] and has produced acceptable results. The increase in the number of cars equipped with GPS is expected to increase the reliability of this source of data.

However, privacy is something to take into consideration using this method since it might record more information than strictly necessary.

2.1.7 License Plate Recognition

License Plate Recognition (LPR) is another AVI method. As the name suggests, this method works by identifying a vehicle by its license plate. For this method to be implemented, cameras need to be placed on the points of the network to analyze. The captured images can then be processed by a computer vision algorithm like the one presented by [CCCC04] (an example of some of the steps on this algorithm can be seen on Figure 2.3). With this information, it is also possible to estimate the average speed between two points where the vehicle was located (if the distance between these points is known) as well as the vehicle class. This last information is obtained after validating the obtained license plate (possibly from an external source) or by identifying the vehicle type on the image.

Extracting the license plates can be a challenging task due to the increased complexity caused by, for example, different illumination conditions (day, night, rain, etc.). More recently, deep learning algorithms have provided great results when applied to computer vision problems. A deep learning approach has been proposed for this problem by [PRR16] and claims to have better results than other proposed methods. However, these algorithms usually require a big computational power which might require an investment (apart from the necessary on cameras' installation). Computer Vision has also proved to be a useful source of information, not only for surveillance purposes, but also as a traffic sensor as well [LRB09, LKR$^+$16, NSR18].

2.2 Origin-Destination Matrix Estimation

2.2.1 Overview

Origin-Destination matrices are a fundamental part of the analysis of a transportation network [Abr98] once they are planned based on information acquired from this matrix. O-D Matrices represent the number of travelers that commute between two points of a location with these points being centers that represent a part of the region (centroids).

The methods used to estimate these matrices, as discussed on [Abr98], have some different characteristics. While some methods require an existing O-D matrix (for example an outdated matrix built using information acquired from a survey) others are able to build a matrix without requiring an existing O-D Matrix (called target O-D Matrix).

Another characteristic that differentiates the methods is the way these handle the congestion problems. Some models consider that the level of congestion is negligible, meaning that the route assignment is not dependent on the level of congestion. These models consider a proportional

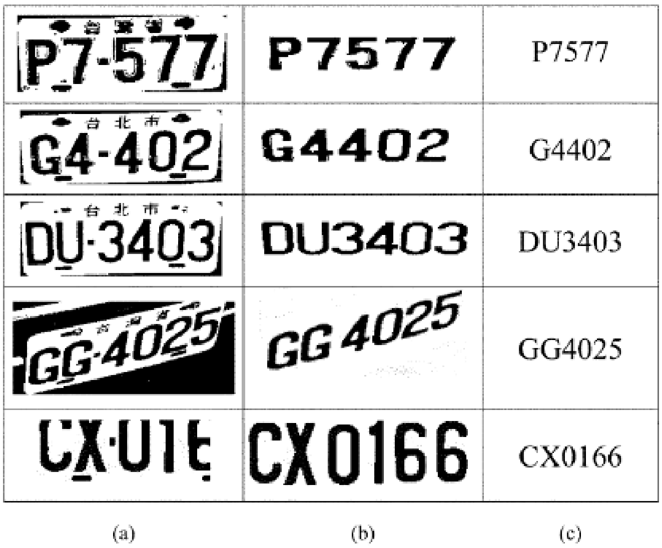

Figure 2.3: Example of License Plate Recognition from [CCCC04].

assignment. On the other hand, the models that consider the levels of congestion consider a user equilibrium assignment (according to [War52] the system is in equilibrium when no user can reduce the travel time by choosing another route) and try to satisfy this principle.

Origin-Destination Matrices can also be separated into two categories, with different purposes for different ranges of time. On one hand, static O-D Matrices considers flows as time independent, averaging the observed flows. Static matrices are used for long term planning and design of the transportation network. Because this type of matrices represents a constant flow, these are used to understand the mobility patterns of the population and the acquired information can be used to perform changes to the current infrastructures or add new ones so that the network can support the demand. [BR11] Dynamic O-D Matrices provide a valuable source of information for the short term. These matrices consider the time and represent the traffic variations through time and can be used, for example, for traffic control and management or online applications like real-time traffic predictions that can be useful for route guidance. Inside this category, two sub-categories can also be considered: one for online applications which consider a short range of recent time of observations to estimate the matrices and these are used, for example, in some route advisors. The other category uses traffic counts divided into periods of time: these can be used for example for planning a route.

2.2.2 Techniques to Estimate Static O-D Matrices

2.2.2.1 Mobility surveys and census information

Depending on the type of census, information about the population's regular trips can also be available. However, these censuses have a high time interval between them (usually 10 years, although some countries do it every 5 years) because of the high cost associated and a long time to obtain results.

Mobility surveys are surveys conducted to the population, door by door on residences or workplaces, and are done to collect information about the daily trips of each person. The main difference to the demographic census is the covered area: while the censuses are usually done in the whole country, these are usually done only in the areas that are being studied. Another differentiating characteristic is the scope of the surveys: the censuses have a wider scope, trying to capture more demographic information about the population, while the mobility surveys' main focus is to gather information about the mobility patterns of the population.

The main advantage of these surveys is that these allow knowing the reason of the trips done by the population, instead of simply knowing the number of trips done, which can be a valuable source of information. However, these are expensive to perform and therefore, their periodicity is low. Therefore, the inferred matrices using this information can become outdated after some years. Another downside of this approach is that the demand caused by other persons other than the population, such as tourists, is not considered.

2.2.2.2 Travel Demand Based Models

Some researchers have tried to build an O-D matrix using gravity models. Gravity models are used to predict interactions between two elements and require parameters that need to be tuned. [Hög76] proposes a technique to estimate those parameters on a simple network and claims to have good results, without being necessary a previous O-D Matrix. [BR11] refers that the main gravity based models do not handle external-external trips accurately (trips that have its origin and/or destination outside the zone of interest). It is also mentioned on [Wil78] that gravity models are not appropriate for inner city areas due to the short distances between the areas. The same author also mentions that this method is able to deal with inconsistencies in traffic counts.

[TW89] proposed different models apart from the mentioned Gravity model: opportunity model and Gravity-Opportunity model. Additionally, different methods were attempted to estimate the parameters based on traffic counts. The authors conclude that Opportunity and Gravity-Opportunity models require more time to compute and do not guaranty the reliability of the resulting O-D matrices produces from these models.

2.2.2.3 Entropy Maximization and Information Minimization

Entropy Maximization is a procedure that "seeks the most likely configuration of elements within a constrained situation" [JP00]. This method tries to find the flows between the points in the network with the constraints that are the observed traffic counts.

These approaches have the characteristic that both can use an existing O-D Matrix that will be updated accordingly to the observed traffic counts. An Entropy Maximization model was proposed on [Wil78] and claimed good results, although this method requires that inconsistencies on traffic counts should be removed before applying the algorithm.

[VZ78] proposed a method using Information Minimization for estimating O-D Matrices. In [VZW80] the authors of the aforementioned approaches discussed that the two methods were quite similar: the only difference between them is the unit of observation, both sharing the same advantages, such as the possibility of using an O-D matrix obtained from other sources and the good fit into small areas (as opposed to the gravity methods mentioned previously), as well as the disadvantages like the inability to deal with incorrect data. The performance of this model using real data has not been tested.

2.2.2.4 Statistical Approaches

Statistical Inference approaches have been applied for estimating Origin-Destination Matrices. Explained in [Abr98] are the following approaches: Maximum Likelihood (ML), Generalized Least Squares (GLS) and Bayesian Inference (BI). These approaches assume that the traffic volumes and the target O-D Matrix are assumed to be generated by probability distributions.

Maximum Likelihood uses a target O-D Matrix and it aims to maximize the likelihood of observing the obtained traffic counts and the initial matrix on the produced matrix. [Abr98] This method assumes that the observed traffic counts and the target O-D Matrix are independent. [Spi87] proposed one application of this approach and mentions that the model is feasible if the link counts are consistent. [Haz00] proposes a method that is able to estimate a correct O-D Matrix without the need of an existing matrix as opposed to the other mentioned method, although it can also incorporate one existing matrix. Furthermore, the author also mentions that the model can incorporate errors in the measured flows.

Generalized Least Squares is also used to produce O-D Matrices and, like the previously mentioned method, also assumes that the observed traffic counts and the target O-D Matrix are independent. One application of this is presented on [Bel91] and the authors present good results but mention that the model is sensitive to incorrect traffic counts.

At last, the Bayesian Inference approach uses the existing target matrix as a probability of the new estimated matrix and uses the observed traffic counts as a different source of information. [Abr98] The two sources are then combined to produce a new O-D Matrix. On the method proposed on [Mah83], the author explains an approach using Bayesian Inference and mentions that the main advantage of this method is the flexibility of the belief used on the previous O-D Matrix.

2.2.2.5 Gradient-Based Techniques

Gradient-Based techniques as the one presented in [Spi90] start by using an initial O-D Matrix and assign it to the network; the observed traffic counts from this matrix are then used to perform adjusts on the O-D Matrix so the resulting matrix produces traffic counts as similar as the original one. This is an iterative process in which the changes made at each iteration are made accordingly to the gradient function which tries to minimize the objective function. The authors have tested this method on real and complex networks (some with more than 1000 links) and the resulting O-D matrices were very satisfactory.

Despite the good results provided by this method, [DB05] has tried to control the changes caused by the updating process on the existing matrix thus limiting the deviation from the target O-D matrix.

2.2.2.6 Other Approaches

A Hopfield Neural Network has been used to obtain an O-D Matrix by [Gon98] on a simple network using traffic counts but this application has not been tested on real networks.

Another attempt of using neural networks to predict trip distribution was done on [MTU00] but the model struggled to obtain good results.

Recently, [TK17] has tried to estimate O-D Matrices using Markov Chains. When the Markov property was not violated, ie, the transition probability on one node does not depend on the previous nodes' transition probabilities, the obtained results were quite good. However, when this property was violated the results were very unsatisfactory, thus affecting the reliability of the obtained matrix. Therefore the authors mention that this approach is suitable for weakly connected networks as in a network between cities but when used on strongly connected networks like in a city this method is not reliable.

2.2.3 Techniques to Estimate Dynamic O-D Matrices

When it comes to estimating dynamic Origin-Destination Matrices (matrices that vary over time), [BR11] mentions that the proposed techniques for this problem use mainly two approaches: State-Space Modelling and Kalman filters.

In State-Space Modelling, the first step is to define an initial state and then specify a transition equation and measurement equation. While the first, transition equation, describes the state over time, the measurement equation relates the state to the observed indicators which in this case are usually the measured traffic counts.

Kalman filtering is an approach used to make the estimate of the next state by using a series of observations over time. This approach handles measurement and state errors. As mentioned on [BR11], the disadvantages of this approach are that intensive matrix operations are required as well as the need for enough data.

One implementation of this approach can be seen on [CT99]. This proposal is able to estimate time-varying O-D Matrices without the need of a previous one and also addresses the impact of intersections on the produced matrix.

A study of AVI technologies to gather information about traffic has been conducted on [ZM06] using synthetic data and the authors analyzed the errors obtained from different penetration rates of the used AVI methods. The authors mention that in order for AVI methods to provide useful information, enough penetration rate is required.

A practical use of AVI methods like Bluetooth to estimate time-dependent O-D Matrices was tested on [BMMC10] and [BMB$^+$12] using also Kalman filters and the results demonstrate that this method of traffic data collection can already provide good results.

[Haz08] proposes a Bayesian Inference method to estimate O-D Matrices applying it on a small part of a real network.

A different approach is proposed on [CPM+13] and the authors claim that by using the "Quasi-Dynamic" estimations (presented by the authors) the estimates are more accurate since the Kalman filtering estimates strongly rely on the initial O-D Matrix. More recently, this method was extended on [BRA+18] in order to eliminate the need for an initial O-D Matrix. This method is applicable where path choices have small importance on roads like highways.

Neural networks have also been tried for predicting the flow on points of a network by [HAS98] and the results were promising. However, the authors mention that studies with real data should be made in order to validate the obtained results.

2.3 Artificial Societies

Artificial societies are described in [Saw03] as multi-agent systems used to perform social simulations - a multi-agent system is described by the same author as a series of self-ruling agents that operate in parallel and interact with each other. There are various categories of artificial societies as described in [Dav01]: open societies, where agents can join the society without restrictions; closed societies, where it is not possible for an external agent to join the community; a community is considered as semi-open if there is an institution responsible for deciding whether a potential new agent can or not join the community; at last, a community is said to be semi-closed when new agents are not allowed to enter but they have the possibility to instantiate a new agent inside the community.

These systems are based on a number of assumptions and are used to generate data from which patterns can be extracted, simulating the evolution of the population over time [Bra10]. The same publication mentions some potential applications for such systems such as: verification - where a hypothesis is tested; prediction - where the future state of a system is forecast by using a number of parameters relevant to the evolution of the society; optimization - where different strategies are tested and the compared; and other potential applications.

An example of an application of this concept is demonstrated in [NMC+02] where the authors develop a multi-agent system to simulate interactions in an electricity marketplace where there are agents like energy generation companies, transmission companies, consumers, among other types of agents. This system is used to make a prediction in different time ranges, from day-ahead planning up to years-ahead planning. Other applications of this system include the possibility of testing potential changes on the rules before implementing them in real systems and observe the behavior of the system.

Other use of this type of system can be seen in [TC02] where, at the beginning of the century, the authors developed a multi-agent system to simulate the pay-TV subscriber market - with this system it was possible to estimate the consequences of eventual changes in the existing conditions. The same publication gives some examples of multi-agent systems used for commercial purposes such as an application to simulate the paths taken by customers and test different placements for products and the costumers' purchases for each situation, or a system used to predict the CD sales for the Japanese pop market. The same publication also identifies strengths and weaknesses inherent to agent-based modeling, some of those described below:

Strengths:

- Heterogeneity - high diversity of the agents that compose the population can be introduced;
- Lack of assumptions - the system as a whole is not constrained by any restriction which can lead to diverse behaviors;
- Communication - the agents can communicate with each other and modify their behaviors based on this

Weaknesses

- Lack of data - if there is no adequate data available, the modeling of the system can be affected;
- Establish behaviors - identifying and implementing the possible agents' behaviors can be difficult

Further research on the potential of multi-agent systems for modeling human behavior and test the response of the population to certain policies is demonstrated on [BG10]. Due to the growing importance of environmental issues, the authors suggest that such a system could be developed to test the impact of different policies and their impact on the environment. For this purpose, different types of agents could be modeled for different levels of decision; this system would also benefit from the interactions between the agents that can change their behavior. However, this same publication mentions that the verification and validation of these systems are not easy - real data is necessary to compare to the results of the simulation and guarantee the reliability of the system.

A synthetic population initialized with data from the census is presented has been initialized [3]. In this example, the population from Melbourne, Australia is synthesized, establishing the relationships between individuals (both familiar and personal relationships) as well as other personal characteristics such as age and zone of residence. It should be stressed that this population is not used to simulate its evolution, only for demographic purposes. Some efforts have already been reported, which apply the concepts of agents and multi-agent systems to model demand relying on artificial societies so as to analyze different aspects of decision-making on trip variables [RLCB02, RBB$^+$02, RL05c, RL05b, RL05a].

[3] https://github.com/agentsoz/synthetic-population

2.4 Summary

Through this chapter, several methods or collecting information regarding traffic were mentioned as well as their strengths and weaknesses. Table 2.1 summarizes some of the properties inherent to each method.

Table 2.1: Summary of the presented Data Collection Methods.

Method	**Collection Periodicity**	**Cost**	**Reliability**	**Penetration**	**Speed**	**Path Reconstruction**
Surveys	Low	High	-	-	No	Depending on surveys
Traffic Counts	High	-	Prone to failures	High	Aggregate average speed	No
Bluetooth	High	Low (readers)	Data requires preprocessing	Low	Average speed between points	Yes
RFID	High	Medium (readers and active tags)	High	Depends on the tags adoption	Average speed between points	Yes
Mobile Phone Location	High	None	Low accuracy on location	High	Average speed between points	Approximate
GPS	High	None	High	Low	Yes	Yes
License Plate Recognition	High	Medium	High	High	Average speed between points	Yes

Analysing the summarised data it is possible to conclude that AVI methods usually provide more information than methods like surveys or traffic counts because the data that is captured is for each individual vehicle instead of aggregate information and it is possible to reconstruct the path followed by each vehicle (inside the area covered by the deployed methods) . Except for the methods using GPS or mobile phone location, the collected data is limited to the range of the placed sensors, but the users' privacy is also something to consider. Depending on the existing environment (for example, the presence of RFID tags), the decision of which collection system to implement should be taken.

Regarding the methods proposed to estimate Origin-Destination matrices, it is clear that solutions for this problem have been proposed since a long time ago and using different approaches. Surveys and census can provide good results but the downside is clear, with the high associated cost. Travel Demand Models are not appropriate when short distances are considered, but entropy maximization methods are good for this purpose. However, this last approach shows some problems with incorrect data. Statistical approaches have also been proposed with good results with the possibility of combining an outdated matrix or a matrix using other sources of information. Gradient-Based methods have been tested on real and complex networks and have provided good results.

Other approaches like using neural networks or Markov chains have also been tested and while they have a good performance on simple networks, on more complex networks they do not provide the same good results.

Multi-agent systems have proven to be a good approach to represent and simulate human behavior. These societies have the possibility of being used for multiple purposes, from verification of hypothesis, prediction of the evolution of the population up to optimization of certain parameters. Examples of practical applications of this type of systems have been presented and their conclusions.

After analyzing the state of the art, it is possible to conclude that more than one of the proposed methods is able to infer Origin-Destination Matrices correctly. However, these methods do not allow to predict the evolution of those matrices throughout the years. Approaches for estimating these matrices, like mobility surveys, are expensive thus discouraging the decision makers to perform these surveys periodically.

Ultimately, the decisions made to the urban network when using these matrices, estimated using traffic counts or through mobility surveys, might not be the most correct ones, either because the information can be outdated (in case of mobility surveys as a source of information), or because the evolution of the population and the consequent change of the mobility patterns is not taken into consideration.

Chapter 3

Methodological Approach

This Chapter starts by giving a high-level view of the proposed solution in Section 3.1, details the process of the data collection in Section 3.2, and describes in Section 3.3 the initialization of the population and the simulation process. Section 3.4 details how the survey results were synthesized and Section 3.5 specifies how the matrices are generated based on the population and the survey results. At last, Section 3.6 describes how this approach is going to be evaluated and Section 3.7 summarizes the chapter.

3.1 Preliminaries

In order to provide a system that is able to predict the evolution of Origin-Destination Matrices, this project aims to do this prediction by simulating the evolution of a population by using data from multiple sources.

To initialize the population and to collect parameters regarding the evolution of the population, data from the demographic census from 2011 was used, as well as other data provided by *Instituto Nacional de Estatística* (Portuguese National Statistics Institute) on its website. Information about the mobility of the population was initially expected to be available from a mobility survey conducted this year (2019) on the parish of Pombal, Portugal, but this data was not ready on time. Nevertheless, the case study for this project was focused on the population of Pombal parish.

A multi-agent system is developed for this purpose of simulating the evolution of a population, although the agents have no interaction with each other. For this objective, the only events happening in the population are aging of the population, incoming population (either from births or migrations) or outgoing population (caused by deaths or migrations). Each step of this simulation simulates the evolution of the population in one year.

This approach can be separated into two modules: one responsible for the evolution of the population, and another responsible for the mobility dynamics of the population, which will produce the matrices (ODMs). This last module will produce more than one Origin-Destination Matrix namely for each degree of schooling (in Portugal there are 4 possible degrees of schooling), for the active worker population, and at last, one matrix which aggregates all these matrices, which corresponds to the estimated flow of the whole considered population. Because no information about the schedule of each person and their trips was available, only static matrices are inferred.

By having two separate modules, the two modules can work independently: it is possible to estimate the evolution of the population without the goal of estimate the ODMs and the opposite is also possible - produce matrices without the evolution of the population. However, in order to produce those matrices, a population must be provided.

For the development of this project, Python was the chosen programming language. This language is known for its versatility and has a wide number of libraries available which simplified some implementation steps. Additionally, the need for handling and processing data when reading the available information from the sources of data was considered, which is an area Python is known for.

This approach relies on some assumptions such as:

- there are no changes in the territory – the changes on generation or attraction points such as housing areas or workplaces, respectively, is not considered;
- there is no limit of occupation of zones – the number of trips generated to/from each area is considered to be possibly unlimited;
- external trips are not considered – trips with an origin/destination from/to outside the study area are not considered;
- only commuter trips are taken into consideration – due to the nature of the data artificially generated from mobility surveys, sporadic trips are not considered;
- mobility patterns are constant – it is assumed that, for example, the students in a given degree of schooling that live in zone A always attend the school at zone B;
- parameters are directly proportional – the data available for the municipality of Pombal (which contains other parishes other than the parish of Pombal) was resized to correspond to the parish of Pombal.

With this approach to the problem, it is expected that it is possible to estimate the evolution of the Origin-Destination Matrices on a medium term. Because of the assumptions made, the forecasts made by this system in the long term are not expected to be reliable because changes in the territory are likely to happen. Nonetheless, the current approach is expected to be able to estimate the medium term evolution, helping the decision makers to make better decisions, leading to more efficient use of the existing resources and ultimately reducing the time of the trips and the environmental impact of those.

3.2 Data Collection

The first step is to collect the necessary data to model the simulation of the population. For this purpose, INE (Instituto Nacional de Estatística - Portuguese National Statistics Institute) was the source of the extracted information and this information is publicly available on their website.

3.2.1 Population Initialization

First, information about the demographic census was collected. The census in Portugal are conducted every 10 years and by the time this work was conducted (2019), the most recent information available was from 2011. Even though this information is older than desirable, the spatial granularity of the available information is valuable since it allows to know the distribution of the population throughout the area.

In order to produce the OD Matrices, it was fundamental to delimit the smaller areas within the study area from which trips will be generated or attracted. Among the existing spatial granularity provided in the 2011 Census, one of those must be chosen depending on the desired level of granularity. The adopted zoning system consisted in 47 zones and for each one the age distribution was available for the following intervals: 0-4 years; 5-9 years; 10-13 years; 14-19 years; 15-19 years; 20-24 years; 25-64-years and >64 years. Additionally, for each of these zones, information is available about the number of students, retired persons and also about the number of people working in each activity sector (primary, secondary and tertiary). For this purpose, the distribution between male and female persons was not considered.

Because some of the age intervals were wider than desirable (25-64 years and >64 years), which could lead to a misrepresentation of the population, it was necessary to fetch more information. To do this, INE also provides information about the age distribution of the population with a thinner granularity. The available age interval for this information is of 5 years, from 0-4 years up to 80-84 years and also older than 84 years and this age distribution is available between 2011 and 2017. However, this information is only available for municipalities and not for smaller areas like the ones considered for origins and destinations of trips. In order to use this information, the absolute values cannot be considered but ratios can be used (such as the growth of the population between 2011 and 2017 or the percentage of the population in a given age group).

3.2.2 Population Evolution

To represent the evolution of the population, 3 events other than aging were considered and data about each of these were gathered:

- mortality – the number of deaths occurred between 2011 and 2017 was collected and divided by the number of persons in the population in each of those years; furthermore, INE also provides an age distribution of deaths for a population of 100 000 persons - this will allow modeling the probability distribution of the deaths that will be simulated;

- natality – the number of births per year between 2011 and 2017 was collected and divided by the size of the population in each of those years; additionally, the mothers' age group when each birth occurred was gathered for the most recent year available (2017);
- migration – for this event, only the population that changes their place of residence to a different area are considered (this means that migrations inside the same population, like between zones, are not considered); INE only provides the migratory balance (this is, the number of people that enter minus the people that leave) between 2011 and 2017; this information is not enough to simulate the evolution of the population since the age group of the population that joins or leaves is missing - the way this was handled is explained later.

Although there are more recent data available from 2018, these were monthly values. Even though the annual values could be estimated using values from one or more months, eventual variations on these numbers that occur throughout the year could lead to inaccuracies resulting from this estimation. For this reason, the annual values until 2017 were preferred.

3.2.3 Population Activity

When estimating the trips made by the population, one aspect to consider is the unemployment rate of the population. This rate is useful to know which percentage of the population is not employed and thus trips from those persons will not be considered.

The most recent unemployment rate is for the year 2018 and is available for the following age groups:

- 15-24 years
- 25-34 years
- 35-44 years
- > 44 years

The younger age group overlaps with the ages that are considered to be students' ages. Persons in the population between 15 and 17 years are considered to be students and therefore the unemployment rate should not be applicable for those. The unemployment rate for this age group will be considered for the population between 18 and 24 years old instead, which might lead to some inaccuracies resulting from this assumption.

3.2.4 Shapefiles

Also provided by the demographic census are the shapefiles. These can be filtered to display different levels of spatial granularity of the population or to display only some parts of the area and are useful to have a visual representation of the population that is being used or even help on other purposes as exemplified later.

3.3 Population Initialization and Simulation

With the collected information, after some processing, the population can be initialized, as well as the components that will be responsible for events like births, deaths, migrations.

As mentioned earlier, the baseline information for the initial population is the one provided in the 2011 census due to their spatial granularity. However, these can be considered as outdated as of the date of this project which could lead to an incorrect representation of the population.

To minimize the effects of this, the more recent data from 2017 was used: the ratio between the size of the population between 2017 and 2011 was calculated and, for every zone and for every age group, the number of persons within that age group in that zone was multiplied by that ratio. This way, an estimation of the number of persons in each zone in 2017 is done.

However, two of the age groups available in the census were too wide to get an accurate representation of the population, namely the intervals between 25 and 64 years and the age group older than 64 years. Also available from 2017 is the age distribution of the population with a 5 year age interval. With this information, it is possible to calculate the percentage of the population in each of the smaller intervals within the larger interval (for example, the number of the persons between 25-29 years divided by the number of persons between 25 and 64 years). This is repeated for each age group and, with this process, the number of persons within each 5-year interval can be estimated for each zone.

The population can then be initialized once that for every zone the number of persons within each age group is known. For every person, a random number within the age interval will be generated and used as age, and a person will be placed in the population in the corresponding zone. With this process, it is expected to have a representation of the population in 2017 - more recent information would be preferable to get a better representation of the actual population, but it was not available.

In every step of the simulation, the aging of the population is simulated by incrementing the age of each person in one year for every step.

3.3.1 Mortality

In order to estimate the number of deaths per year, the age distribution of the population is used. As an input, for every year between 2011 and 2017, the number of persons within each age group is used, the label being the number of deaths for the corresponding year divided by the size of the population in that year. Using this data, different Python's scikit-learn regressions were trained and the best one was chosen according to the R^2 Score. For the training process, the input values (number of persons in each age groups) were normalized and cross-validation was used with 7 folds (leave-one-out strategy). This estimator will be used to output the number of persons to die by passing the age distribution of the simulated population and multiplying this percentage by the size of the population.

Having the number of deaths to simulate, it is necessary to know the persons' ages with more probability to die. The used data source also provides this information by giving the number of deaths in each age for a population. A curve fitting using a Python library named SciPy is then done using this age distribution and every step of the simulation, the persons' ages will be generated according to this distribution. For each of the generated numbers, a person with that age is randomly selected and removed from the population.

3.3.2 Natality

A similar approach was taken to simulate the birth of new members of the population. However, the tested regressions failed to perform well only given the age distribution of the population. This might be because other factors that are not the age distribution, such as economic factors, have an impact on the births. Because this approach would provide unreasonable results, a simple linear regression was trained, also using Python's scikit-learn, with the historical number of births divided by the size of the population. To calculate the number of births, the result of this regression is then multiplied by the size of the population.

Knowing the number of births that will occur on the population, it is necessary to know where the newborns will reside. To do this, it is assumed that their parents will not change the place of residence. The age group of the mother at the moment of each birth is known: this distribution will be fitted into one of the possible distributions, being chosen the best one. For each birth, one number is generated according to this distribution, representing an age group, and a person within that range is randomly selected - the place of residence of this person will be the one where the newborn is placed. With this process, it is ensured that the newborns have a higher probability of having a place of residence where there are more persons in an age group where it is more likely to have children than on places where are more persons on other age groups.

3.3.3 Migrations

The information regarding migrations is limited: it is only known the migratory balance, the number of people that enter the population minus the number of people that leave. This balance is known for every year between 2011 and 2017. No information regarding which age group is more likely to enter or leave the population is known, neither the number of people that enter or leave the population.

In order to simulate this event appropriately, this approach was followed: given the age distribution of the population in 2011, a population is initiated. Since the number of births and deaths are known for every year until 2017, it is possible to simulate the evolution of the population from 2011 until 2017 using these historical values. For this purpose, the places of residence can be ignored since only the age distribution of the population is needed. This simulation was done five times and the age distributions of the resulting population were averaged and compared to the real age distribution of the population in 2017.

The difference between the simulated population and the actual one was calculated for each age group. The disadvantage of this approach is only considering migrations as possible events, ignoring other possible causes for the differences between two populations, the simulated and the real one. Having the difference of each age group, a probability distribution is fitted accordingly.

To predict the number of migrations, a regression is trained with the evolution of the number of the migratory balance divided by the size of the population throughout the years. This regression will then be used to estimate the number of migrations in the population by multiplying the output of the regression by the size of the population. For each migration, using the distribution previously mentioned, one age group is generated and the following process occurs: if the migratory balance is negative, a person within that age interval is selected and removed from the population. Otherwise, an age between that interval is generated, and a person with that age is placed in the population - the zone where each person will reside is selected using the probabilities of where the persons within that range reside.

3.4 Mobility Dynamics Data Collection through Surveys

When this project was started, the goal was to merge the simulation of the population with data about the mobility dynamics of the population collected from the conducted surveys. However, the results from the surveys were not ready on time for this project, so it was necessary to find an alternative. The result of these surveys would include information about the origin (place of residence) and the destination of that person (could this be for instance their school or workplace) as well as other information about each person such as age or, in case of students, future plans like going to university.

Since these results were not available, the surveys' results will be artificially generated based on some real information. These results are written on the template where the real results were expected - this way the real results of the surveys can easily replace the artificial information.

This process can be split into three steps:

1. **Gather a list of attraction points** - some research was conducted to fetch a list of possible points of attraction points. These points could be schools (from elementary schools to high schools) or places where people could work. A list of schools was gathered as well as a list of workplaces. Each of these points was categorized, with the schools being categorized depending on the education level (the schools with more than one schooling degree are considered one time for each degree available), while the workplaces were categorized with their activity sector: primary, secondary and tertiary. For the primary workplaces, places with activities like agriculture, flowers production, and pig-farming places were considered, while on secondary activities places with industries were considered. Finally, for tertiary activities, other activities like services, insurances, sales, health were considered. For each of these potential activity points, a class was manually given, estimating their size and the number of persons that might work there. Furthermore, for each attraction point, schools or workplaces, the geographical coordinates are given.

When reading these lists of point, using the coordinates it is possible to know in which zone of the area each point is placed by matching these coordinates with the shapefile provided in the demographic census. This shapefile contains the limits of each of the zones, as previously explained, and using the Shapely library in Python it is easy to calculate the zone in which the point is contained.

2. **Calculate the number of surveys to generate** - from the data available in the demographic census, it is known, for each zone, the number of people that are students or work in each of the three sectors. Based on this the number of answers to the survey is generating by multiplying a random number by the number of people in each of those 4 occupations. The random number is used as a percentage and is generated with a normal distribution with a mean of 30 and a standard deviation of 5 - with this distribution, it is ensured that a reasonable part of the population is sampled while keeping some unpredictability.

3. **Survey generation** - initially, the survey results were expected to be received in more than one file, each one containing the results from a type of survey that was collected. To simplify, only two types of surveys were artificially generated: one for students and one for workers. Another important point missing by moving from the real surveys to the synthesized ones is the lack of demographic information about each person - no information is known about the person which corresponds to each row of the survey, could this be information about their age, occupation (education level or activity sector) and, if it is a student, the future plans (if they plan to go to the university or start working). To work around the issue of the uncertainty of the type of occupation for each of the rows, an extra column is added to mention the type of occupation (education level in case of students and activity sector for workers), which allows a brief characterization of each person. Regarding schooling, one assumption is made: all the students within each zone attending a given schooling degree, go to the same school, meaning that in each zone the students can possibly go to four schools depending on the schooling degree they are currently attending. For each zone, one school of each degree is then assigned to that zone.

   For the number of students previously calculated for every zone, one of those four schools is randomly selected and written on the survey. A similar process is followed when generating the surveys for workers: for the number of surveys for each activity sector, a number of workplaces are generated. However, since some workplaces are assumed to employ more people than another, those are differentiated with four classes, with the probabilities of choosing a workplace depending on that class - the smaller workplaces are represented with the class number one, while the bigger ones have the class number four.

By the ending of this process, two surveys are generated: one with information about the students' mobility dynamics and other for the workers, which will serve as a source of information for generating the Origin-Destination Matrices.

3.5 Trip Distribution

This step is responsible for the assignment of the trips from an origin zone to a destination zone. While the origin is considered as being the place of residence, which is known from the distribution of the population throughout the zones, the destinations are the zones to which people commute for their activity, either to attend school or work.

Because the information regarding the population' mobility patterns was artificially generated, and thus the lack of some information like the age, degree of schooling, this process was simpler than originally thought. This is because, for each entry of the survey, no information about that person is known like their education level (for workers), age, future plans (like going to the university, in case of students) or even if the workplaces are located outside the area that is being considered. Ideally, a better characterization of the regular destinations of each group of persons would be done, and a higher heterogeneity would be present in the existing data.

Different types of matrices will be produced, depending on the occupation of the population: one for each degree of schooling (there are four possible levels in Portugal), one for the worker population and finally another matrix that aggregates all the other matrices, which corresponds to the demand of the whole considered population.

The first step is to read both surveys: the students' survey results and the workers' results. For every entry of each survey, the origin of the commute is stored, as well as the destination of that trip. If the survey that is being read is from students, the type of school should also be recorded for every entry. The results of both students and workers surveys are stored separately.

This information is then grouped by the origin and, with this process, the destination zones for each origin can be known. For students, the information is also grouped by the degree of schooling. The objective behind this grouping is to known which zones attract more trips from each origin, and for students, in which zones are the schools attended by them for each degree of schooling. The percentages of each pair of origins and destinations can then be calculated by dividing the number of surveys with a destination to a given zone by the total of surveys collected from that origin. For the students' surveys, the same is done but for each degree of schooling.

One possible case is that there might be zones where there is no information available about the residents of one of those zones. This could be for instance, when no information is known about which school of a given degree of schooling is attended by the residents in a place. To work around this issue, using the shapefile provided in the demographic census, the centroid for each zone is calculated and, for each, the distance to the other centroids are calculated and stored in a min-heap. As such, every time that no data is found for an origin, data for the next closest zone is used until information it is available. This approach makes the assumption that persons that live near each other have common zones of attraction.

Having information for all the zones available, the next step is to build the Origin-Destination Matrices. Every matrix is initiated as a bi-dimensional array with the size of the number of zones of the area, with all the values as zero. Each row or column represents a zone alphabetically ordered.

In order to fill the matrices, every person in the population is analyzed. Since only the students and workers commutes are being considered, it is assumed that people younger than 6 years are not going to school yet and people older than 65 years old are considered as retired and therefore are not considered for the produced matrices.

The distribution of the students per degree of schooling is done by age, according to the ages usually seen in each degree in the Portuguese education system, as represented in Table 3.1. For this purpose, school drop out is not considered and therefore, every student inside that age range is assumed to make a trip to the school. Based on the probabilities of the zones that attracted more trips from that origin, and depending on the schooling degree, a zone is selected according to these probabilities. The cell which corresponds to a trip between that pair of origins and destinations is incremented on the matrix regarding that degree of schooling.

Table 3.1: Distribution of students by age per degree of schooling.

Age	**Schooling Degree**
6-9 years	ESC1 (1st - 4th grade)
10-11 years	ESC2 (5th - 6th grade)
12-14 years	ESC3 (7th - 9th grade)
15-17 years	ESCSEC (10th - 12th grade)

The remaining population, older than 17 years and younger than 66 years old, is considered as the worker population. However, this may not be true (for example, they can attend university or be unemployed) and therefore not have a daily commute. To represent this absence of activity when producing the matrices, the unemployment rate is used. For each of the age groups (18-24 years, 25-34 years, 35-44 years or older than 44 years), the unemployment rate is available and depending on the age of the person, that value is selected. A random number between 0 and 1 is then generated and if it is lower than the unemployment rate, that person is considered unemployed, thus it is not considered for the matrix. For the persons that are considered as employed, a process similar to that of students occurs: based on the probabilities of the zones that attract more trips from that zone, a zone is selected according to these and the value in the workers' matrix that corresponds to that origin and destination is increased.

Having all the students' and workers' matrices filled with the trips, the next step is to create a new column and a new line. These will be used to store the information about the total number of trips originated in that zone (in the last column) or attracted by that zone (last row). The values presented in the last row correspond to the sum of all the values in the other rows in the same column while the last column is equal to the sum of all the other columns in the same row.

In order to obtain an Origin-Destination Matrix that aggregates all the trips from the considered population (students and workers), all these five matrices previously mentioned are summed. This resulting in a matrix that has information about the trips of the whole considered population (other groups of the population such as the retired population is not contemplated in this) instead of representing just a part of it.

Each of these matrices is then stored into a *csv* file. In order to simplify the viewing of these, the rows and columns are labeled with the name of the zone corresponding to each zone.

3.6 Approach Evaluation

This section describes the tests that are done to assess the correctness of the designed implementation. This is done by verifying how this approach behaves in normal conditions and how it responds to the changes done in the system by testing them in different scenarios and comparing its behavior with the expected behavior. With this process, the implementation of the approach is verified.

The validation of this approach would be possible if real data was available, then it would be possible to compare the simulated parameters with the real parameters, could these parameters be about the population (like the distribution per zones, age distribution) or the produced Origin-Destination Matrices. Since no data was available on time, the process will test the implementation and verify if it is working as expected.

The setup of these tests is the following: first, the control results are obtained - for this, the simulation is repeated 5 times, each one with 5 steps, and all the necessary parameters are collected. These results are then averaged and used as the control results. For each experiment done, the simulation is also repeated for 5 times, each one for 5 steps, with the results from the 5 simulations also being averaged. The values used for comparison are the average values for each scenario, the control, and test scenarios, for each step of the simulation.

Two different categories can be used to classify these tests: one where the behavior of the population is compared to the behavior of the real population based on the historical values; and another one where hypothetical scenarios are tested and the result of those scenarios matches the expected outcome.

3.6.1 Comparison with real values

In these tests, the behavior of the system and its evolution is compared to the values that were used to model the system, and it is expected that the values from the simulation are similar to those or follow an identical trend.

First, the evolution of the number of births per year is assessed: the historical values between 2011 and 2017 are divided by the size of the population in each of those years. Then the simulation is run for 5 times and the number of births per year is recorded, as well as the size of the population, these values averaged, and the ratio between the average number of births per year and the size of the population is calculated. This simulation is expected to simulate these numbers for 2018 and onwards, so the trend from the simulation should follow the same pattern from the historical values.

Next, a similar process is repeated using the migratory balance instead of the number of births per year. The ratios between the real migratory balance and the average of the estimated migratory balance are calculated and it is expected that the simulated values follow a similar tendency.

The evaluation of the mortality estimates cannot be done in the same way since the estimation of the number of deaths per year is done using the age distribution of the population. Because the age distribution from the population that is being simulated can be different from the population that is being used to train the estimator, the results from the comparison would not be comparable.

For the three estimations done (number of births, migratory balance, and number of deaths), the R^2 Score is calculated to evaluate the accuracy of these estimations. This score ranges between 0 and 1, the higher the value the higher the better the obtained model fits the sample and provides better results.

The following assessments are done in order to verify the result of the curve fittings that were done, this is if the numbers generated from the fitted distribution match the desired distribution. To do this, the histograms from the real distribution and the fitted distribution are compared, thus allowing to evaluate the fitted distribution. A good distribution should be similar to the real one.

In this approach, three distributions were fitted and are going to be evaluated: the distribution for the age of deaths; the distribution for the age group of the mothers' age at the time of each birth, where each value from the distribution corresponds to a value in the array containing the age groups; and the distribution for which age groups are more likely to migrate. Each distribution is chosen using the Kolmogorov-Smirnov test, which evaluates how two distributions compare to each other. A low score in this test indicates a bad fit, and therefore, the distribution with the highest score is the chosen one for each case.

3.6.2 Hypothetical scenarios

For these tests, the behavior of the population in hypothetical scenarios is going to be assessed and compared to the expected output. These tests are done to test the implementation in unusual conditions and allow to evaluate both the evolution of the population and the Origin-Destination Matrix generation process.

The list below contains the tests that will be performed and the expected results of each one. While the first four tests the evolution component of the implementation, the remaining three tests assess the mobility dynamics component of the system.

1. **Increase childbirth** - the number of births per year passed to the estimators will be multiplied by 2 - it is expected that, when comparing the number of births per year with the control results, a proportional increase should be observable.

2. **Increase mortality** - the historical number of deaths per year will be also doubled - the expected result is that the number of deaths per year during the simulation is also duplicated.

3. **Change migrations trend** - the historical migratory balance values passed to estimate the migratory balance are the opposite (positive values become negative ones, and negative ones become positive values) - the estimated migratory balances should demonstrate the contrary trend when compared to the control scenario (if the migratory balance was negative it should be positive or if it was positive it should now be negative).

4. **Change deaths' age distribution** - the probability distribution used to simulate the age of deaths of the population will be nearly flat, meaning that people of all ages will have a similar probability of passing away - the number of persons in each age group after 5 steps of the simulation will be compared and, in this scenario, the age group with a higher probability of being selected in the control scenario should have a higher number of people when compared to the control scenario, while the remaining age groups should have a slight reduction in the number of persons.

5. **Change population location** - every person of the population is going to have the same place of residence - with this test, the trip generation from all the zones should decrease, except for the zone where the population is now concentrated which should generate all the trips of the population.

6. **Change school locations** - the zones where the schools are located should be changed to a single zone - the expected result is that the zones where the schools were previously located should have a lower trip attraction and the zone where the schools are now placed should reflect this change by having a higher trip attraction.

7. **Change workplace locations** - all the considered workplaces will have their location changed to one different zone - it is expected that the trip attraction for the places where the workplaces were previously placed will decrease, and, on the other hand, the attraction for the zone where all the workplaces are concentrated is expected to increase.

3.7 Summary

This Chapter describes the proposed approach step by step, from the data collection process to the generation of the matrices.

First, the population is initialized using the census' zoning and by resizing that population from the census using more recent information. Then, the evolution of the population is simulated and the Origin-Destination Matrices are estimated for the resulting population in each step. A visual representation of this flow can be seen in Figure 3.1.

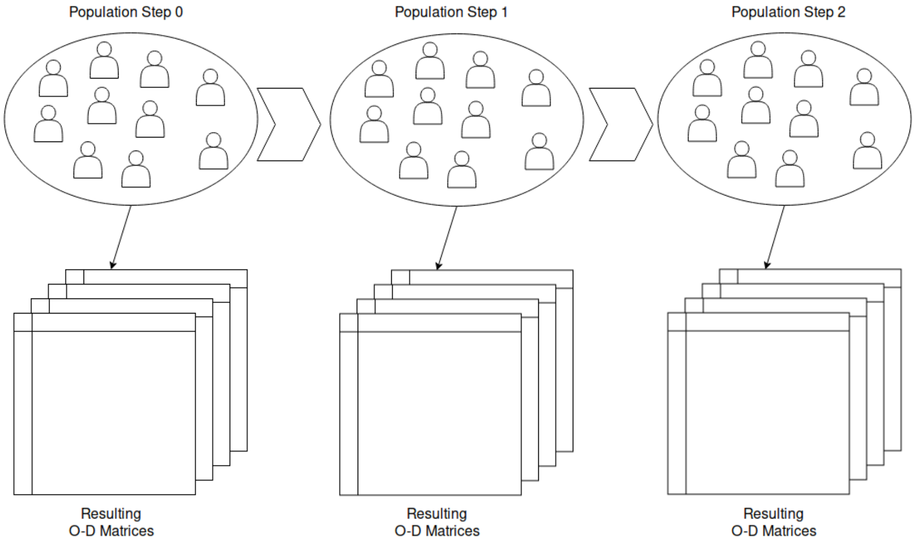

Figure 3.1: Workflow of the proposed approach.

Between each step of the population, these events are considered: aging of the population, mortality, natality, and migrations (either incoming population or outgoing).

Because the results from the mobility surveys were not available for this project, how these were artificially generated is also explained: two lists of possible attraction points were gathered (one for schools, other for workplaces) and the surveys' results are generated based on the distribution of the population and from these lists of points of attraction.

The matrices are then generated combining the population with these results, each matrix containing the flows of each subset of the population. At last, this implementation is tested both by comparing its behavior when compared to the historical values and also by analyzing its comportment in different scenarios and comparing it to the expected behavior.

Chapter 4

Result Analysis

In this Chapter, a practical application of the approach explained in the previous chapter is presented in Section 4.1 and the results of the made tests are presented and discussed in Section 4.2. Finally, 4.3 summarizes the conclusions from this chapter.

4.1 Case Study

The approach explained earlier was developed using the population of the parish of Pombal as a case study, a city located in the center of Portugal. Initially, the results for the mobility surveys conducted on this parish were expected to be available during the development of this project, hence the choice of this parish. This parish is split in 47 zones according to the demographic census used as a baseline for the population.

One possible issue present in the existing data was found in the shapefile provided on the census, and possible with the distribution of the population through the zones as well. When visualizing the shapefile, it was found that one of the zones was repeated more than once on the parish - this might be because these areas are not part of a specific zone and all these unassigned areas are merged into this zone. This might lead to some inaccuracies on the representation of the population throughout the parish since the population contained in this zone is spread across the area of the parish, although the number of persons in this part of the population is low (around 2% of the population of the parish).

Many of the data was only available for the municipality of Pombal. The problem of this is that the municipality includes other parishes other than the parish of Pombal. To get around this situation, the approach taken was to consider that the values for the parish and the municipality were proportional. For example, while resizing the population from the data available from the 2011 census to the population of 2017, it was considered that the population of the parish has decreased its size in the same proportion as the municipality. The number of births, deaths, and migrations per year were also considered as percentages of the total population of the municipality and then multiplied by the size of the parish' population. Ideally, data about the exact population should be used, but this approach should provide a reasonable approximation.

In order to use this data extracted from INE's website, the pages containing the information were downloaded. These pages were then parsed using Python's BeautifulSoup and the desired information was extracted. The downside of this approach is that it requires to download manually different files to use data about other population or more recent information. To avoid this problem, the implementation of a crawler could be considered, but the way the search is implemented on this website made this too complex to develop this solution for this scope.

For the survey generation process, a list of 12 schools was collected using as source the website of the school grouping [1] as the main source of information, each one of those categorized according to their education level. Each school was considered one time for each degree taught. For the list of workplaces, 53 different places were gathered [2]. The used criteria to search for workplaces for each type of activity are the following:

Table 4.1: Examples of categories considered for each activity type

Activity Type	**Activity category**
Primary	Fruit and vegetable production, food production, production of flowers and plants, pig-farming
Secondary	Industry (glass, rubbers, plastics, among others), footwear production, metallurgical industry
Tertiary	Health services (hospital), shopping, hospitality, computer services, car repair, restaurants, insurances

For each of the attraction points gathered, either schools or workplaces, the geographical coordinates of these points were also obtained. When reading those points, by matching the geographical coordinates with the shapefile, it is possible to know the zone in which that point is placed.

4.2 Discussion

4.2.1 Births and Migration Evolution

Figure 4.1 shows the evolution of the birth numbers, where the real values are in blue and the predicted values are displayed in red. The comparison is done relative to the size of the population because the historical numbers are for the municipality, while the simulated values are for the parish which is smaller.

[1] https://www.aepombal.edu.pt/escolas/

[2] https://portalnacional.com.pt/leiria/pombal/empresas/

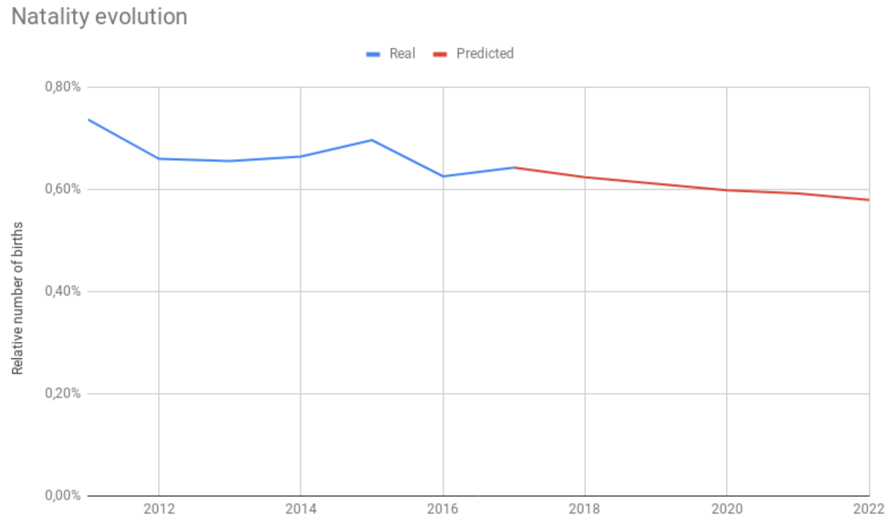

Figure 4.1: Comparison between the predicted and historical number of births per year, on a relative basis.

By looking at the Figure it is possible to conclude that the estimated values follow a similar trend observed in the historical values: the global trend of the historical values is the decrease of the natality, and the same pattern is observed in the prediction made by the system. The R^2 Score for this estimation is 0.416, which is not a satisfactory result and can be explained by the irregularity of the historical values. Because natality can be affected by factors like economic or social, these values can be irregular, making its estimation difficult. Although it seems like the predicted behavior follows the real trend, these external factors can influence this prediction.

A similar experiment is conducted to the values of the migratory balance: the real balances were divided according to the size of the population, as well as to the average of the predicted migratory balance for each of the steps. The obtained values are represented in Figure 4.2, with the historical values in blue and the estimated ones are in red.

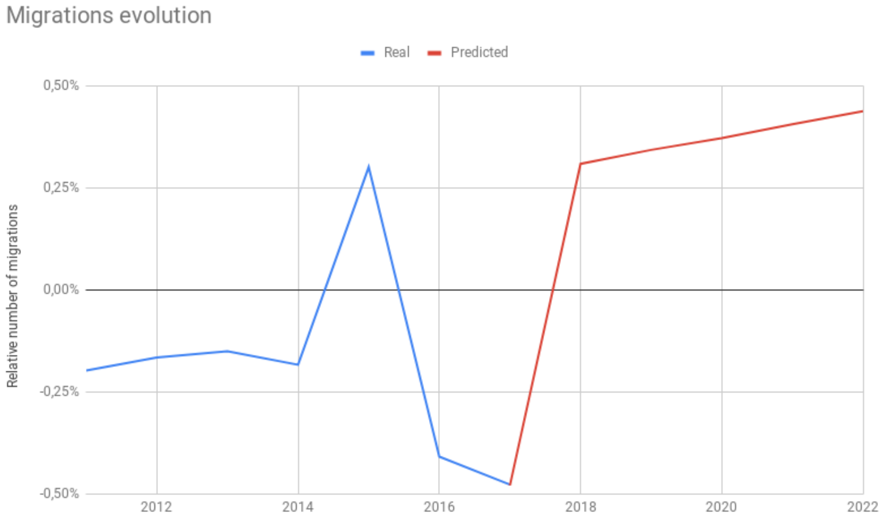

Figure 4.2: Comparison between the predicted and historical number of migrations per year, on a relative basis.

With the exception of the year of 2015, the migratory balance is always negative and the global trend is to be lower each time. The predicted values follow this trend, being the migratory balance lower each step. For this model, the obtained R^2 Score was 0.073, which is a low score. This score is possibly be explained because all the historical values but one are negative, thus lowering the score. Just like the natality, these values are influenced by external factors like the creation or elimination of workplaces and although the predictions done by the system seem to follow a correct trend, these might not be accurate.

As mentioned earlier, the evaluation of the mortality values cannot be done in the same way because this estimation is done using the age distribution of the population, and the population that is used for training is the municipality of Pombal, while the simulation is done on the parish of Pombal, which have different age distributions. For this purpose, different types of regressions from Python's scikit-learn library were tested and evaluated using the R^2 Score. For this case, it was found that Automatic Relevance Determination Regression (ADR Regression) provided the best score. The R^2 Score for this model was 0.854, thus meaning that the model is able to predict well the mortality of the population given the age distribution.

4.2.2 Curve Fitting

To perform the fitting of the distributions, the Python library SciPy was used, and the developed code was based on a tutorial[3]. This tutorial was adapted to the needs of this project: only integers are expected (representing either a value or an index of an array of values), therefore the output value is rounded to the nearest integer number, and only values within a given range are expected - if the value generated from this distribution is out of a given limit (providing a lower and upper

[3] http://www.insightsbot.com/blog/WEjdW/fitting-probability-distributions-with-python-part-1

bound), a new number is generated from this distribution to replace that invalid value. This is done to prevent invalid values like negative ages or invalid indexes.

One of the distributions that were fitted was the age of deaths: this information is for the "Center" region of Portugal and shows how many deaths occur in each age for a population of 100 000 people. In this case, it was found that Generalized Logistic was the best distribution, and the results are shown in Figure 4.3

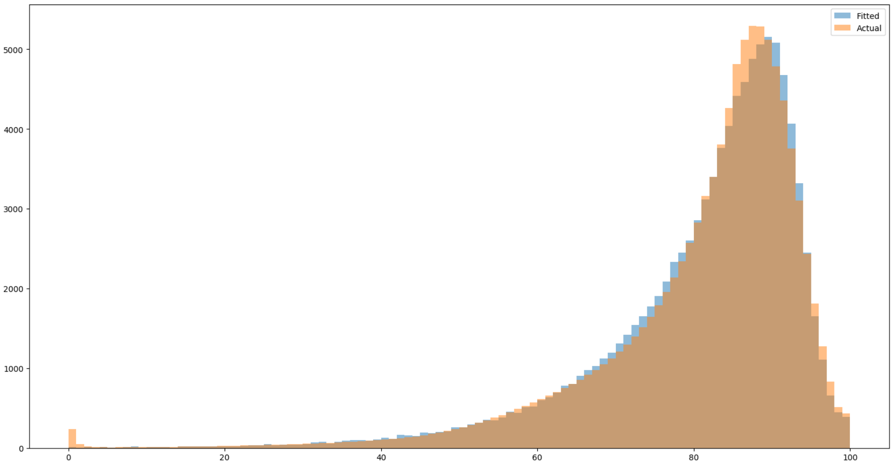

Figure 4.3: Comparison between the ages generated and the real ages.

From the chart, it is possible to conclude that the resulting distribution is quite good since the values generated (in blue) from this distribution match the real ones (displayed in orange) in almost the whole range of values. The biggest discrepancy can be seen on the lower values, where the fitted distribution gives a smaller probability than the real one, and in the values with the higher probability of being generated, where there can be seen some difference between the two distributions.

A distribution was fitted to generate the age group of the mothers' age at the time of each birth and each value represents the index of the array of a given age group. The existing age groups are the following: 10-14 years, 15-19 years, 20-24 years, 25-29 years, 30-34 years, 35-39 years, 40-44 years, 45-49 years and older than 49 years. Each value corresponds to an age group, being the smallest possible value 0 (corresponding to the age group between 10 and 14 years), and the largest possible value 8 (which corresponds to the last age group, 50 years or older) - however, in this distribution all the values are between 1 and 6.

In this distribution, the most adequate one was found to be Johnson SU. The results from this distribution can be seen in Figure 4.4

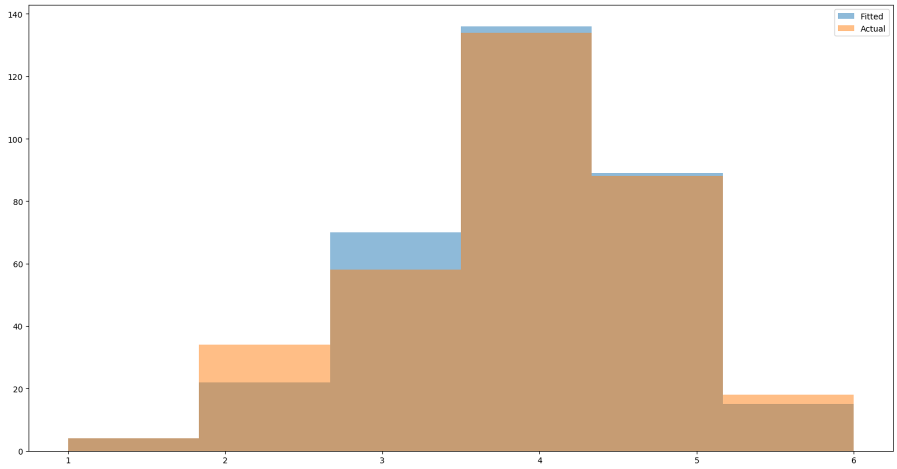

Figure 4.4: Comparison between the mothers' generated age groups and the real age groups at the time of birth.

The resulting distribution for the mothers' age group at the time of the birth is close to the real distribution. On the third age group in the chart (25-29 years), the distribution assigns a higher probability than the real one, whereas the opposite happens on the second age group (20-24 years) and in the fourth (30-34 years). Nonetheless, the distribution is able to generate these ages with a good approximation to the real values.

At last, the distribution for which age groups are more likely to migrate is fitted. As explained earlier, the distribution of these age groups was calculated by simulating the evolution of the population using the real number of deaths and births per year and the difference between the age distribution of the simulated population and the real population is used for this. The disadvantage of this approach is that it does not differentiate the age groups that are more likely to enter or leave the population. For this fitting, all the positive values obtained from this experiment were not considered because the migratory balance is expected to be negative for this population, meaning that for this distribution the migratory balance for those age groups will be considered as 0. Each value from this distribution is an index of the array of the possible age groups. For this purpose, it was found that the best distribution was the Generalized Normal that results in the distribution shown in Figure 4.5.

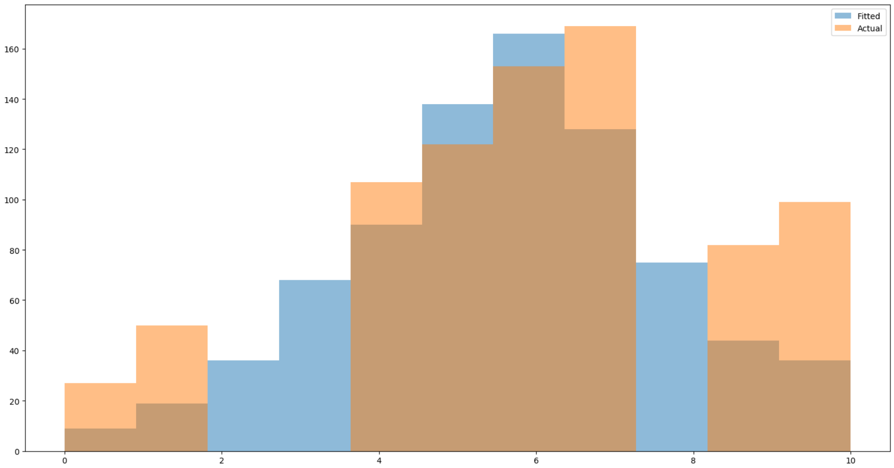

Figure 4.5: Comparison between the migrations' generated age groups and the real age groups.

This distribution seems to struggle to generate the desired probability, but this is because some age groups were considered to have a null probability, as previously explained, while the distribution continues to generate those values. The chosen distribution attributes a lower probability on both the lower end and higher end of the possible range of values, but the overall result is able to generate a reasonable approximation given the previously mentioned issues.

When initializing the simulation, instead of testing all the available distributions on the utilized library, only the three distributions mentioned earlier are tested. This allows to cut down a big amount of time necessary to initialize the simulation and this option can be easily changed to fit other possible distributions by changing a flag.

4.2.3 Hypothetical scenarios

A summary of the tests in hypothetical scenarios presented earlier in Section 3.6 and the expected result for each of those are displayed in Table 4.2.

Table 4.2: Proposed tests and expected results.

Scenario	**Expected Output**
Double historical values of births per year	During the simulation, the number of births per year should be approximately twice as many when compared to the control test.
Double historical values of deaths per year	During the simulation, the number of deaths per year should be approximately twice as many when compared to the control test.
Use the opposite migratory balance	During the simulation, the population should have a positive migratory balance, instead of the negative balance from the control test.
Flatten age of deaths probability distribution	After 5 steps, the older age groups should have more persons when compared to the control test, while the remaining age groups should have a slight reduction in their size.
Move the residence zone of the entire population to "Granja"	The number of trips from "Granja" should increase while the trips from other zones should be nonexistent
The existing schools will all be located in the zone "Outeiro das Galegas"	The matrix resulting from the workers should be similar to the control one but the matrix resulting from students should have a reduced trip attraction in the zones where the schools were placed and a bigger trip attraction on the mentioned zone.
The location of all existing workplaces will be in "Aldeia dos Redondos"	The matrices resulting from students should be similar to the control ones while the workers' matrix should have a bigger attraction on the mentioned zone and a smaller one on the remaining zones.

The results for the first four tests are presented on Subsection 4.2.3.1 while the results for the remaining three tests are given on Subsection 4.2.3.2.

4.2.3.1 Evolution of the Population

For the first test, the historical number of births per year are multiplied by 2 and the expected result is that the simulated number of births per year follows this change by being nearly the double than in the control scenario. Figure 4.6 displays average the number of births in each of the scenarios: in red are the control values while the blue is the test scenario in which the values were doubled.

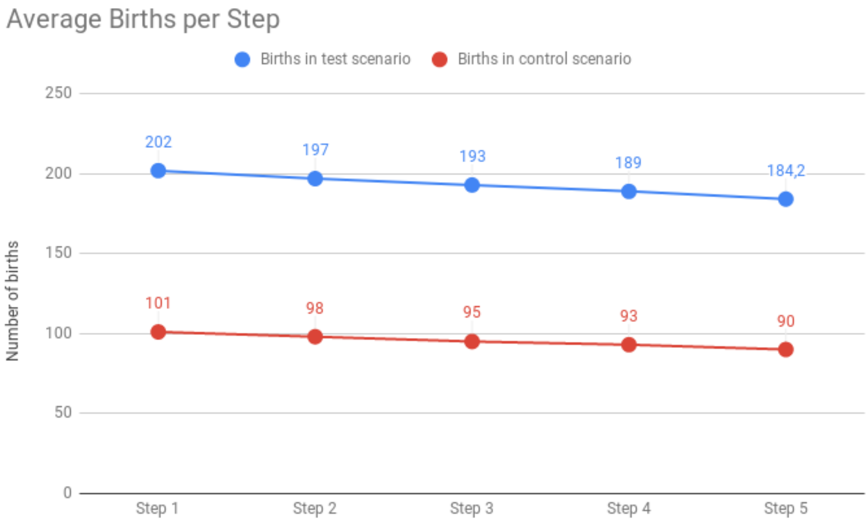

Figure 4.6: Average number of births per step in control and test scenarios.

In both cases, the trend is similar, with some differences towards the last steps which can be explained by the difference in the size of the population (the number of births is calculated using the number of persons in the population). The test scenario has nearly the double of births per year when compared to the control scenario, which is the expected result.

Next, a similar experiment is done but instead of doubling the number of births per year, the number of deaths is duplicated. Just like the previous experiment, it is expected that the number of deaths per year is close to the double of the control scenario. The results of this experiment are displayed below in Figure 4.7.

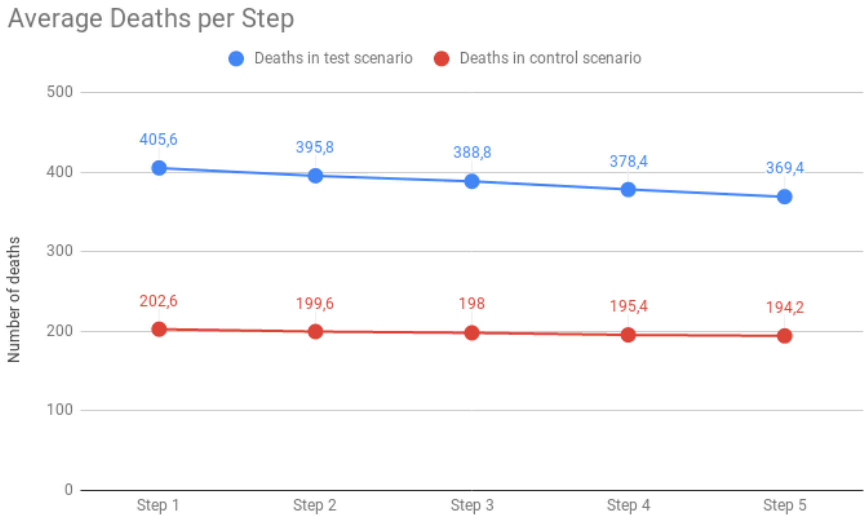

Figure 4.7: Average number of deaths per step in control and test scenarios.

The number of deaths per step in the control scenario is approximately the double of the number from the control scenario in the first two steps. However, as the simulation continues, the ratio between the number of deaths in both experiments changes - towards the end the number of deaths per year is slightly lower than the double of the control scenario. This can be explained because the age distribution of the population is changed as a consequence of the increase in the number of deaths, and the number of deaths is estimated using this age distribution.

In the next test, the opposite historical migratory balance was used: while in the control scenario, all the balances were negative except for one year (which means that there are more outgoing people than incoming), in this test, these values were multiplied by -1, meaning that for this test, historically, there are more people joining the population than leaving. Figure 4.8 shows the simulated migratory balances for each case, with the control scenario values in red and the results from this test in blue.

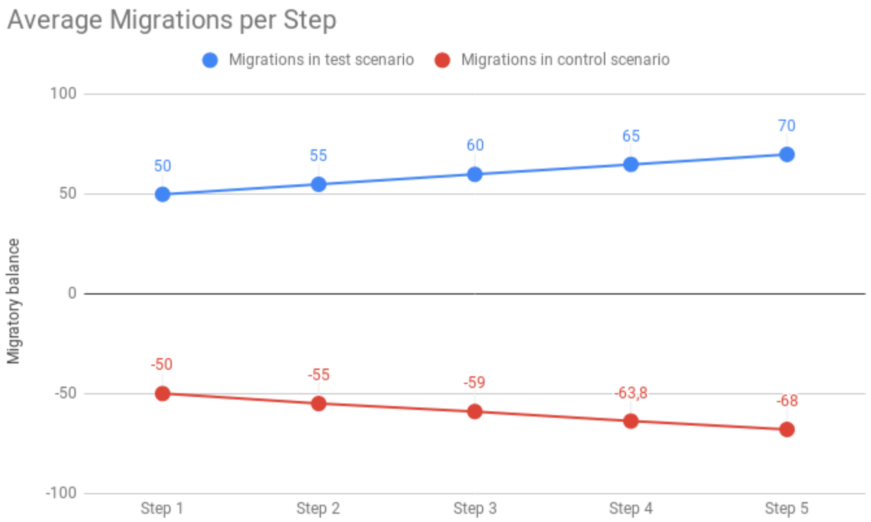

Figure 4.8: Average migratory balance per step in control and test scenarios.

The results from this experiment meet the expected ones: in this case, the population is more likely to receive new members than it is to release members to other populations. The absolute values for each step in both cases are almost the same since the historical numbers were only multiplied by -1.

The following experiment consists in changing the probability distribution of the ages of deaths so that all ages have a similar probability. Results for this experiment are displayed in Figure 4.9, with the age distribution of the population after 5 steps in the control scenario in blue and the distribution for this scenario is displayed in red.

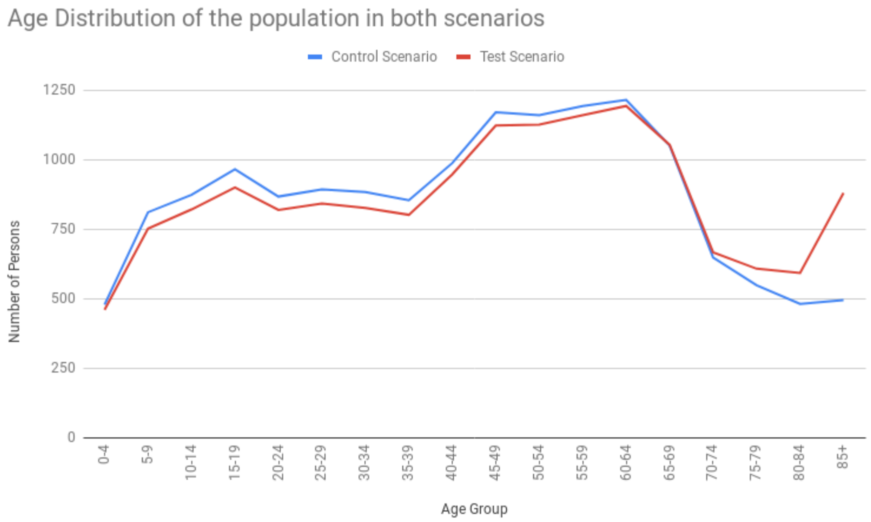

Figure 4.9: Average number of deaths per step in control and test scenarios.

The main difference between the age distributions in two scenarios can be seen in the latter age groups: as expected, the older age groups have more persons in the test scenario because the probability of those ages being generated is lower, resulting in more people in that range. On the other hand, the remaining age groups demonstrate a reduction in their size, since their probability of passing away is higher than on the control scenario. The coincidence between both scenarios in the age groups between 60 and 74 years of age can be explained by the similarity in the probability distribution for both scenarios.

4.2.3.2 Mobility Dynamics

In this experiment, the whole population of the parish of Pombal is going to have *Granja* as a place of residence. The number of trips with origin in each of the zones in the parish is going to be measured and the difference between the averages from the control scenario and test scenario is displayed in the Figure 4.10. It is expected to see an increase in the demand from *Granja* and a reduction in the trips from the remaining zones.

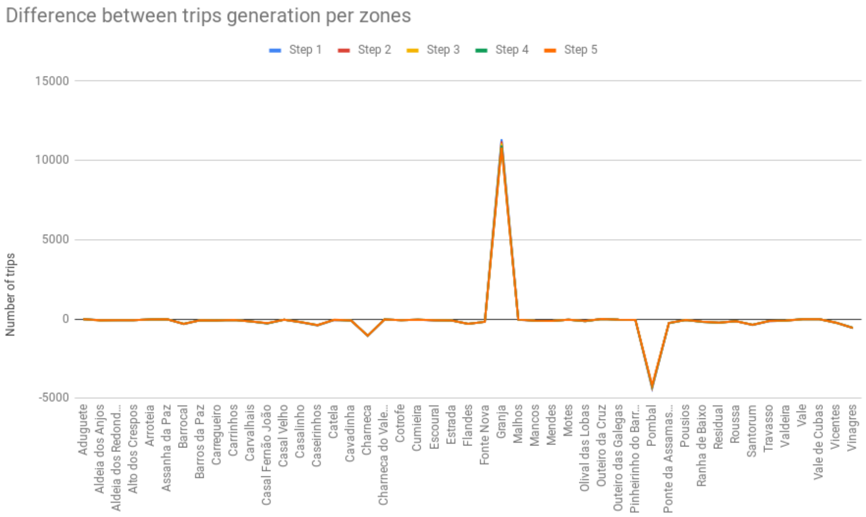

Figure 4.10: Difference in trips generated from each zone in control and test scenarios.

Only the zone of *Granja* shows a positive variation in the number of trips from that zone. On the other hand, the remaining zones have a reduction in the number of trips with that origin: the decrease is proportional to the number of persons that reside in each zone, with the biggest reduction being seen in the zone of Pombal, which is where more people reside.

The following experiment is to assess the distribution of the student population per school. All the schools will be placed in *Outeiro das Galegas* and it is expected that the trip attraction will increase for this zone, but to decrease in the zones where the schools were originally placed. The zones where the schools are located for the control test are shown in Table 4.3.

Table 4.3: Zones in which schools of each type exist.

School Type	Zones
ESC1	Casalinho, Caseirinhos, Charneca, Pombal, Santorum, Vicentes
ESC2	Charneca, Pombal (2)
ESC3	Charneca, Pombal
ESCSEC	Pombal

Below, figures 4.11, 4.12, 4.13, and 4.14 show the difference in the trip attraction between the test scenario and the control scenario, for each of the corresponding matrices to the schooling level.

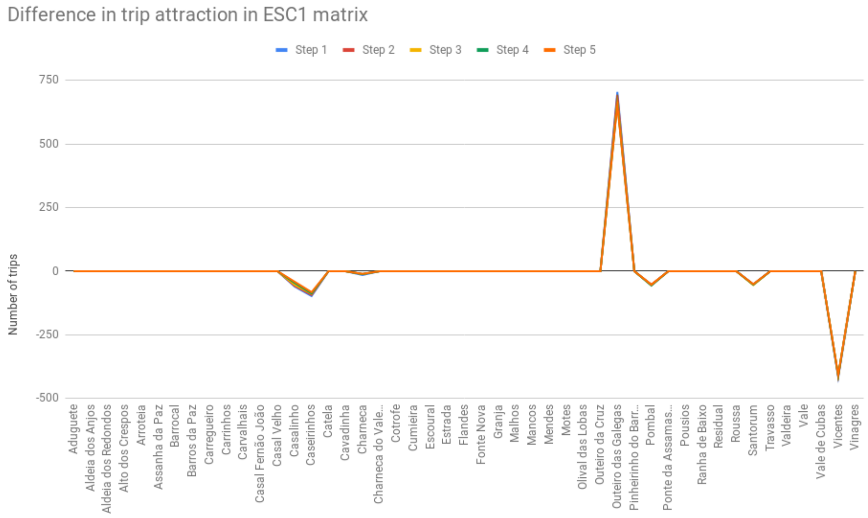

Figure 4.11: Difference in trips generated to each zone for the ESC1 matrix.

It is possible to observe a decrease in the zones where this type of schools was previously located and an increase in *Outeiro das Galegas*, where all the schools are all placed in this experiment. This means that results are the expected ones.

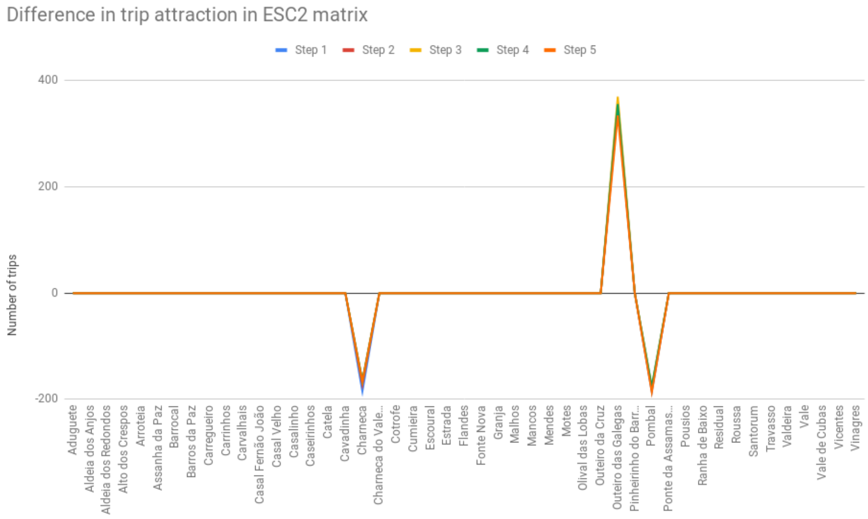

Figure 4.12: Difference in trips generated to each zone for the ESC2 matrix.

For this matrix, there is a decrease in the number of trips to *Charneca* and *Pombal* can be observed, the zones where these schools were placed. On the other hand, more trips are generated to the zone where the schools were artificially placed,

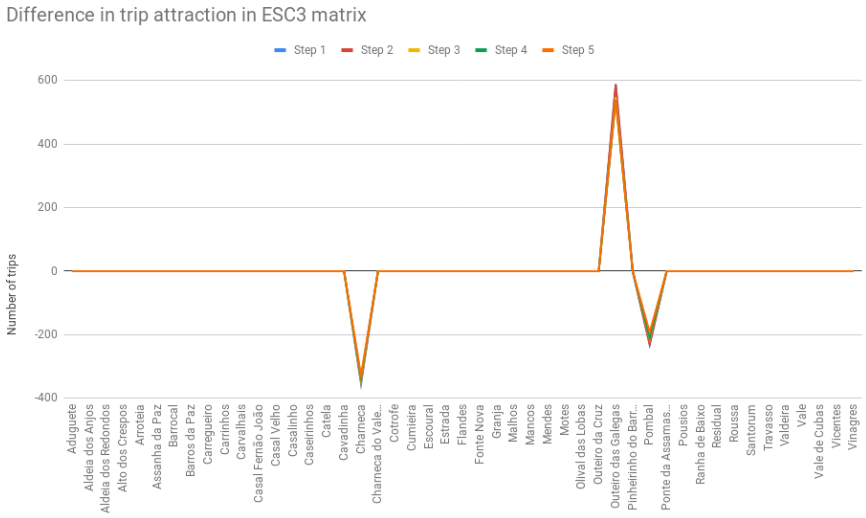

Figure 4.13: Difference in trips generated to each zone for the ESC3 matrix.

Just like the previous experiment, in which the schools of this education level were placed in the same zones, a reduction in the trip attraction for such zones can be identified. On the contrary, the number of trips to *Outeiro das Galegas* increases.

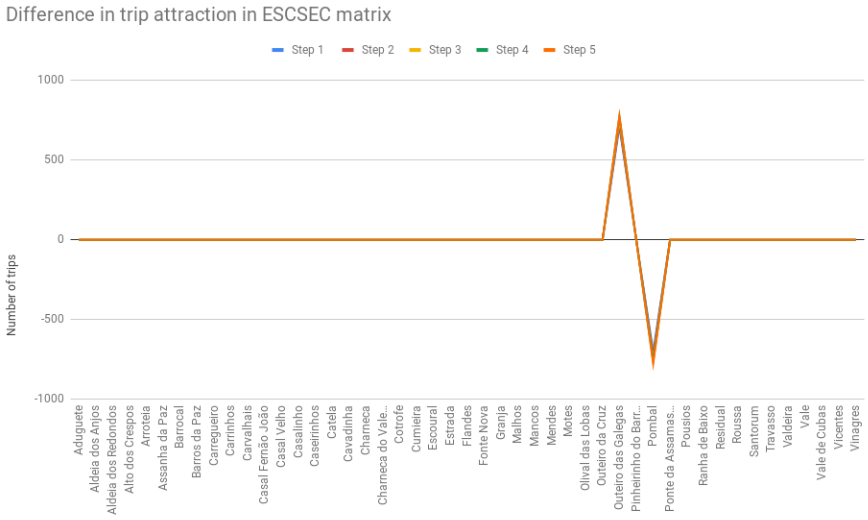

Figure 4.14: Difference in trips generated to each zone for the ESCSEC matrix.

Since there is only one school of this type, the reduction in the trip attraction is only visible in this zone where the school was placed, namely the zone of Pombal. The opposite happens with the zone in which the school was placed, where an increased trip attraction can be observed.

At last, this experiment will change the location of the existing workplaces from the existing zones to the zone *Aldeia dos Redondos*. In the control test, workplaces exist in the following zones: Aldeia dos Anjos, Carvalhais, Casal Fernão João, Caseirinhos, Charneca, Fonte Nova, Granja, Mancos, Pombal, Residual, Roussa, Santorum, and Travasso. The difference in the trip attraction between this scenario and the control one is displayed in Figure 4.15 and it is expected that these zones have a negative difference and that *Aldeia dos Redondos* has a positive variation.

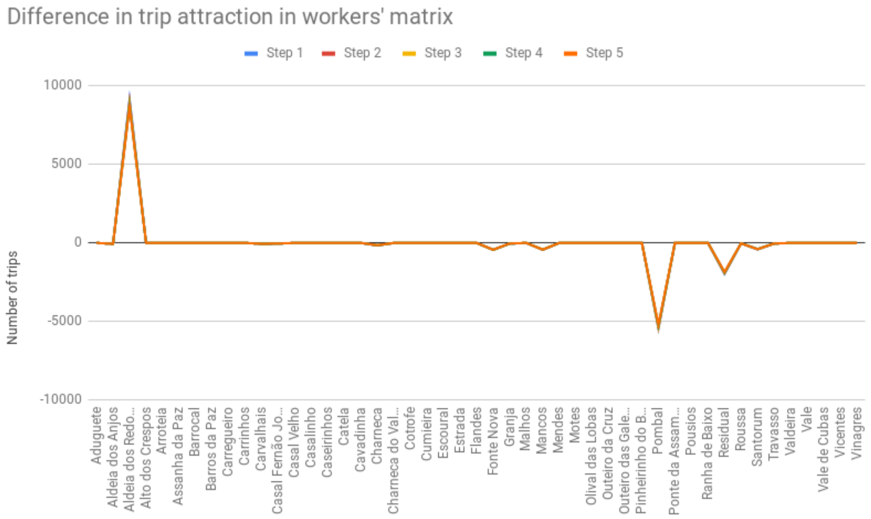

Figure 4.15: Difference in trips generated to each zone for the workers' matrix.

As expected, it is possible to see a decrease in the number of trips attracted to the zones mentioned earlier on. The zone in which the workplaces are now located has a positive variation, as it was expected.

4.3 Summary

The application of this method to the parish of Pombal was presented, as well as were some of the problems found during the implementation, such like the lack of data for the studied parish.

Next, the test results were then discussed: the irregularity of the historical values makes the prediction of those values difficult, as demonstrated by the scores from the natality and migrations estimators. However, the results from these estimations follow the trend existing in the real values. On the other hand, it seems to be easier to estimate the number of deaths per year, given the obtained scores.

The results from the tests done on hypothetical scenarios have matched with the expected results, meaning that the implemented solution seems to be correct. However, this approach could not be validated because there was no information available to do it.

Chapter 5

Conclusions

This Chapter provides a summary of the work carried out, starting with a brief overview of main remarks given in Section 5.1. The main contributions of this project are discussed in Section 5.2, whereas the limitations of devised approach are identified in Section 5.3 alongside with some suggestions of potential improvements for further development.

5.1 Overview

This report starts by describing how traffic data can be collected and the existing approaches for estimating Origin-Destination Matrices. The identified problem is that the current methods do not provide a way to estimate the evolution of those matrices throughout the future. The work carried out in this dissertation focuses on this problem, trying to predict the evolution of a population within an area by simulating its evolution and the consequent evolution of the Origin-Destination Matrices.

Because of the data that was expected from the mobility surveys was not ready on time, this approach could not be validated. This data would have allowed having more information about the demography of the current population as well as real data about its mobility patterns, which had to be artificially generated for this project. Because of this, the module responsible for producing those matrices is also simpler than desired, since there was no information to classify each person and the patterns for each group of persons. However, this implementation was tested and verified: the evolution of the simulated population was compared to the historical values and it was possible to observe that the trend followed in the simulation is similar to the one observed in the real values. Furthermore, some testing of the approach in controlled scenarios was done and the results matched the expected ones.

A downside of this approach is that only frequent trips are considered, ignoring other spontaneous trips done by the population. Depending on the information collected by the surveys, trips to leisure activities could also be considered, but since this information was not generated when the results from the survey were synthesized, those trips were not contemplated in the matrices. Another downside, inherent to the process of estimating ODMs through mobility surveys, is that trips caused by people other than the population such as tourists are not considered.

Overall, this process does not eliminate the needs of conducting mobility surveys: this approach is only able to make an estimation on a medium term due to its limitations. This proposed solution does not consider the possibility of the creation or elimination of attraction points such as workplaces. Other potential factors such as economic, political or social, which influence the evolution of the population are not treated because they are hard to predict. As such, this approach is only reasonable for estimating the evolution of those matrices in the interval between those surveys, allowing to predict the demand of the population, and thus make the necessary changes in advance.

5.2 Main Contributions

This dissertation is focused on finding solutions for predicting the evolution of the mobility needs of the population. As such, the main contributions of this project are the following:

- **Creation of an artificial population based on census data and other data sources** - it is explained how an artificial population can be initialized using these data sources;
- **Simulation of the evolution of a population** - given the initial population and the collected parameters regarding the evolution of the population, the evolution can be simulated on a medium term;
- **Estimation of ODMs based on the population** - the possibility to infer Origin-Destination matrices given a population and the results from the mobility surveys is explained.

Both the process of initializing a population and simulating its evolution can be used for other purposes rather than estimating ODMs only. Additionally, this dissertation resulted in a scientific paper reporting on the main results and contributions herein discussed, which was submitted and currently under review for possible presentation at the IEEE International Smart Cities Conference (ISC2 2019).

5.3 Future Work

When developing this project, some aspects were identified for improving this approach. As previously mentioned, it is assumed that there are no changes in the territory throughout the simulation, such as the creation or elimination of points of trips generation (like housing areas) or attraction (for instance workplaces or schools). By introducing some unpredictability on these factors, the simulation could be closer to reality.

Currently, there is no limit of occupation per zone, which means that theoretically it is possible that the whole population could be concentrated in one single zone. To work around this problem, the number of houses in each zone could be considered in order to limit the number of persons that live in one zone.

Another possible improvement would be to consider external trips, could these be trips with an origin outside the area that is being studied, or to an external area (such as trips to the university or to workplaces in other areas). In this approach, trips other than daily commutes to school or to work are not considered, but by having information about those trips, those could be considered, thus resulting in a more accurate demand.

This approach could also be used to test events that may affect the evolution of the population. These events could be, for example, an increase of the number of incoming population (instead of using the historical values to estimate the migratory balance) or the rise of the number of births per year. The effects caused by this in the matrices could then be predicted. The introduction of these events could be simplified so that the evolution used these custom values instead of following the historical trend.

Additionally, dynamic Origin-Destination Matrices could be estimated, depending on the information from the mobility surveys. If results of those surveys include information about the schedules of each person, the inferred matrices could provide information about the demand of the population throughout the day, instead of aggregating the daily demand in one single time frame.

The idea of simulating the evolution of a population to estimate future demands raises the opportunity to explore other contexts. An example of a practical application could be the prediction of the required trash collection in an area.

References

[AAA$^+$17] I. Alam, M. F. Ahmed, M. Alam, J. Ulisses, D. M. Farid, S. Shatabda, and R. J. F. Rossetti. Pattern mining from historical traffic big data. In *2017 IEEE Region 10 Symposium (TENSYMP)*, pages 1–5, July 2017.

[Abr98] Torgil Abrahamsson. Estimation of origin-destination matrices using traffic counts-a literature survey. 1998.

[AFR19] Ishteaque Alam, Dewan Md. Farid, and Rosaldo J. F. Rossetti. The prediction of traffic flow with regression analysis. In Ajith Abraham, Paramartha Dutta, Jyotsna Kumar Mandal, Abhishek Bhattacharya, and Soumi Dutta, editors, *Emerging Technologies in Data Mining and Information Security*, pages 661–671, Singapore, 2019. Springer Singapore.

[BAR15] J. Barros, M. Araujo, and R. J. F. Rossetti. Short-term real-time traffic prediction methods: A survey. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 132–139, June 2015.

[Bel91] Michael GH Bell. The estimation of origin-destination matrices by constrained generalised least squares. *Transportation Research Part B: Methodological*, 25(1):13–22, 1991.

[BG10] Stefano Balbi and Carlo Giupponi. Agent-based modelling of socio-ecosystems: a methodology for the analysis of adaptation to climate change. *International Journal of Agent Technologies and Systems*, 2(4):17–38, 2010.

[BMB$^+$12] Jaume Barcelo, Lídia Montero, Manuel Bullejos, Oriol Serch, and C Carmona. Dynamic od matrix estimation exploiting bluetooth data in urban networks. *Recent Researches in Automatic Control and Electronics*, pages 116–121, 2012.

[BMMC10] Jaume Barceló, Lidin Montero, Laura Marqués, and Carlos Carmona. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. *Transportation Research Record: Journal of the Transportation Research Board*, (2175):19–27, 2010.

[BR11] Sharminda Bera and KV Rao. Estimation of origin-destination matrix from traffic counts: the state of the art. 2011.

[Bra10] Jürgen Branke. *Artificial Societies*, pages 44–48. Springer US, Boston, MA, 2010.

[BRA$^+$18] Dietmar Bauer, Gerald Richter, Johannes Asamer, Bernhard Heilmann, Gernot Lenz, and Robert Kölbl. Quasi-dynamic estimation of od flows from traffic counts without prior od matrix. *IEEE Transactions on Intelligent Transportation Systems*, 19(6):2025–2034, 2018.

[CC18] Hyun-ho Chang and Seung-hoon Cheon. The potential use of big vehicle gps data for estimations of annual average daily traffic for unmeasured road segments. *Transportation*, pages 1–22, 2018.

[CCCC04] Shyang-Lih Chang, Li-Shien Chen, Yun-Chung Chung, and Sei-Wan Chen. Automatic license plate recognition. *IEEE transactions on intelligent transportation systems*, 5(1):42–53, 2004.

[CDBN11] Andrea Caragliu, Chiara Del Bo, and Peter Nijkamp. Smart cities in europe. *Journal of urban technology*, 18(2):65–82, 2011.

[CLLR11] F. Calabrese, G. Di Lorenzo, L. Liu, and C. Ratti. Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36–44, April 2011.

[CPM$^+$13] Ennio Cascetta, Andrea Papola, Vittorio Marzano, Fulvio Simonelli, and Iolanda Vitiello. Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, 55:171–187, 2013.

[CT99] Gang-Len Chang and Xianding Tao. An integrated model for estimating time-varying network origin–destination distributions. *Transportation Research Part A: Policy and Practice*, 33(5):381–399, 1999.

[Dav01] Paul Davidsson. Categories of artificial societies. In *International Workshop on Engineering Societies in the Agents World*, pages 1–9. Springer, 2001.

[DB05] Javier Doblas and Francisco G Benitez. An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix. *Transportation Research Part B: Methodological*, 39(7):565–591, 2005.

[Fin10] Klaus Finkenzeller. *RFID handbook: fundamentals and applications in contactless smart cards, radio frequency identification and near-field communication.* John Wiley & Sons, 2010.

[FRK$^+$14] João Filgueiras, Rosaldo J. F. Rossetti, Zafeiris Kokkinogenis, Michel Ferreira, Cristina Olaverri-Monreal, Marco Paiva, João Manuel R. S. Tavares, and Joaquim Gabriel. *Sensing Bluetooth Mobility Data: Potentials and Applications*, pages 419–431. Springer International Publishing, Cham, 2014.

[Gon98] Zhejun Gong. Estimating the urban od matrix: A neural network approach. *European Journal of operational research*, 106(1):108–115, 1998.

[HAS98] Yang Hai, Takamasa Akiyama, and Tsuna Sasaki. Estimation of time-varying origin-destination flows from traffic counts: A neural network approach. *Mathematical and computer modelling*, 27(9-11):323–334, 1998.

[Haz00] Martin L Hazelton. Estimation of origin–destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, 2000.

[Haz08] Martin L Hazelton. Statistical inference for time varying origin–destination matrices. *Transportation Research Part B: Methodological*, 42(6):542–552, 2008.

[HBB+00] Robert E Hall, B Bowerman, J Braverman, J Taylor, H Todosow, and U Von Wimmersperg. The vision of a smart city. Technical report, Brookhaven National Lab., Upton, NY (US), 2000.

[Hög76] Per Högberg. Estimation of parameters in models for traffic prediction: a non-linear regression approach. *Transportation Research*, 10(4):263–265, 1976.

[HWH+10] Juan C Herrera, Daniel B Work, Ryan Herring, Xuegang Jeff Ban, Quinn Jacobson, and Alexandre M Bayen. Evaluation of traffic data obtained via gps-enabled mobile phones: The mobile century field experiment. *Transportation Research Part C: Emerging Technologies*, 18(4):568–583, 2010.

[ICWG14] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[JP00] Ron Johnston and Charles Pattie. Ecological inference and entropy-maximizing: An alternative estimation procedure for split-ticket voting. *Political Analysis*, 8(4):333–345, 2000.

[LKR+16] G. Lira, Z. Kokkinogenis, R. J. F. Rossetti, D. C. Moura, and T. Rúbio. A computer-vision approach to traffic analysis over intersections. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 47–53, Nov 2016.

[LRB09] P. F. Q. Loureiro, R. J. F. Rossetti, and R. A. M. Braga. Video processing techniques for traffic information acquisition using uncontrolled video streams. In *2009 12th International IEEE Conference on Intelligent Transportation Systems*, pages 1–7, Oct 2009.

[Mah83] MJ Maher. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6):435–447, 1983.

[McN00] Michael G McNally. The four step model. 2000.

[MNC+14] Gabriel Michau, Alfredo Nantes, Edward Chung, Patrice Abry, and Pierre Borgnat. Retrieving dynamic origin-destination matrices from bluetooth data. 2014.

[MTU00] Mikhail Mozolin, J-C Thill, and E Lynn Usery. Trip distribution forecasting with multilayer perceptron neural networks: A critical evaluation. *Transportation Research Part B: Methodological*, 34(1):53–73, 2000.

[NMC+02] M North, C Macal, R Cirillo, G Conzelmann, V Koritarov, P Thimmapuram, and T Veselka. Multi-agent modelling of electricity markets. In *Proceedings of the Agent 2002 Conference on Social Agents: Ecology Exchange and Evolution*, pages 215–226, 2002.

[NSR18] J. Neto, D. Santos, and R. J. F. Rossetti. Computer-vision-based surveillance of intelligent transportation systems. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–5, June 2018.

[PRR16] Rohith Polishetty, Mehdi Roopaei, and Paul Rad. A next-generation secure cloud-based deep learning license plate recognition for smart cities. In *Machine Learning* *and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 286–293. IEEE, 2016.

[RBB$^+$02] Rosaldo J. F. Rossetti, Rafael H. Bordini, Ana L. C. Bazzan, Sergio Bampi, Ronghui Liu, and Dirck Van Vliet. Using bdi agents to improve driver modelling in a commuter scenario. *Transportation Research Part C: Emerging Technologies*, 10(5-6):373 – 398, 2002.

[RL05a] Rosaldo J. F. Rossetti and Ronghui Liu. *Activity-Based Analysis of Travel Demand Using Cognitive Agents*, chapter 7, pages 139–160. Emerald Publishing Limited, Kidlington, UK, 2005.

[RL05b] Rosaldo J. F. Rossetti and Ronghui Liu. An agent-based approach to assess drivers' interaction with pre-trip information systems. *Journal of Intelligent Transportation Systems*, 9(1):1–10, 2005.

[RL05c] Rosaldo J. F. Rossetti and Ronghui Liu. *A Dynamic Network Simulation Model Based on Multi-Agent Systems*, pages 181–192. Birkhäuser Basel, Basel, 2005.

[RLCB02] Rosaldo J. F. Rossetti, Ronghui Liu, Helena B. B. Cybis, and Sergio Bampi. A multi-agent demand model. In *Proceedings of the 13th Mini-Euro Conference and The 9th Meeting of the Euro Working Group Transportation*, pages 193–198, Bari, Italy, June 2002.

[Saw03] R Keith Sawyer. Artificial societies: Multiagent systems and the micro-macro link in sociological theory. *Sociological methods & research*, 31(3):325–363, 2003.

[SK08] Keemin Sohn and Daehyun Kim. Dynamic origin–destination flow estimation using cellular communication system. *IEEE Transactions on Vehicular Technology*, 57(5):2703–2713, 2008.

[Spi87] Heinz Spiess. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5):395–412, 1987.

[Spi90] Heinz Spiess. A gradient approach for the od matrix adjustment problem. 1:2, 1990.

[SRM$^+$16] M. Sandim, R. J. F. Rossetti, D. C. Moura, Z. Kokkinogenis, and T. R. P. M. Rúbio. Using gps-based avl data to calculate and predict traffic network performance metrics: A systematic review. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1692–1699, Nov 2016.

[TÁD15] Reza Tolouei, Pablo Álvarez, and Nicolae Duduta. Developing and verifying origin-destination matrices using mobile phone data: the llitm case. In *European Transport Conference*, volume 2015, 2015.

[TC02] Paul Twomey and Richard Cadman. Agent-based modelling of customer behaviour in the telecoms and media markets. *info*, 4(1):56–63, 2002.

[TK17] Alexandr Tesselkin and Valeriy Khabarov. Estimation of origin-destination matrices based on markov chains. *Procedia Engineering*, 178:107–116, 2017.

[TPP17] Reza Tolouei, Stefanos Psarras, and Rawle Prince. Origin-destination trip matrix development: Conventional methods versus mobile phone data. *Transportation Research Procedia*, 26:39–52, 2017.

[TW89] OZ Tamin and LG Willumsen. Transport demand model estimation from traffic counts. *Transportation*, 16(1):3–26, 1989.

[VZ78] JH Van Zuylen. The information minimizing method: validity and applicability to transport planning. *New developments in modelling travel demand and urban systems*, 1978.

[VZW80] Henk J Van Zuylen and Luis G Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980.

[War52] John Glen Wardrop. Some theoretical aspects of road traffic research. In *Inst Civil Engineers Proc London/UK/*, 1952.

[WD01] John Wright and Joy Dahlgren. Using vehicles equipped with toll tags as probes for providing travel times. Technical report, 2001.

[Wen10] W Wen. An intelligent traffic management expert system with rfid technology. *Expert Systems with Applications*, 37(4):3024–3035, 2010.

[Wil78] Luis G Willumsen. Estimation of an od matrix from traffic counts–a review. 1978.

[ZM06] Xuesong Zhou and Hani S Mahmassani. Dynamic origin-destination demand estimation using automatic vehicle identification data. *IEEE Transactions on intelligent transportation systems*, 7(1):105–114, 2006.