U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

U.PORTO
FACULDADE DE ARQUITECTURA
UNIVERSIDADE DO PORTO

USING MOBILE PHONE OPERATORS DATA IN TRANSPORT PLANNING

FEDOR KUZNETSOV

Dissertação submetida para satisfação parcial dos requisitos do grau de

MESTRE EM PLANEAMENTO E PROJECTO URBANO

Orientador: Professor Doutor Álvaro Fernando de Oliveira Costa

JULHO DE 2023

MESTRADO EM PLANEAMENTO E PROJECTO URBANO 2022/2023 - FEUP / FAUP

DEPARTAMENTO DE ENGENHARIA CIVIL

Tel. +351-22-508 1901

✉ mppu@fe.up.pt

Editado por

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Rua Dr. Roberto Frias

4200-465 PORTO

Portugal

Tel. +351-22-508 1400

Fax +351-22-508 1440

✉ feup@fe.up.pt

http://www.fe.up.pt

Este documento foi produzido a partir de versão eletrónica fornecida pelo respetivo Autor.

Nota: sempre que o documento é escrito em português do Brasil, deve ser indicado o seguinte:

Este documento foi escrito no idioma Português do Brasil.

RESUMO

A presente dissertação de mestrado investiga o valioso papel dos dados das operadoras móveis no planejamento de transportes, abordando seu potencial para revolucionar a gestão da mobilidade urbana e aprimorar os sistemas de transporte. A tese é composta de três objetivos principais. A primeira é pesquisar e analisar a literatura existente sobre o tema do uso de dados das operadoras de telefonia móvel no planejamento de transporte e outras áreas de estudos relacionadas. O segundo objetivo é a análise das técnicas, mecanismos e métodos existentes dedicados à obtenção e processamento de dados móveis para fins de pesquisa científica. O objetivo final é a aplicação dos conhecimentos recolhidos para desenvolver um estudo de caso para a região do Tâmega e Sousa.

Com base na literatura revisada, os dados das operadoras móveis encontraram uma aplicação em várias áreas de estudos geográficos, como planejamento espacial, estudos comportamentais, análise estatística, estudos de mobilidade e planejamento de transporte.

Análises aprofundadas da aplicação de matrizes 'Origem-Destino' na pesquisa de mobilidade lançam luz sobre a utilização de dados móveis em estudos focados na compreensão do comportamento de viagem das pessoas. A última parte da análise metodológica incluiu a avaliação de métodos de processamento de dados no contexto da pesquisa de mobilidade, como reconstrução de itinerário, detecção de modo de transporte e reconhecimento de padrões de localização.

O estudo de caso desenvolve no âmbito do projeto de análise de dados para a região do Tâmega e Sousa, em Portugal. O foco principal foi a análise dos padrões das horas de ponta da manhã e da noite na região central da área de estudo. A comparação posterior dos dados obtidos do operador móvel com os dados do censo nacional de Portugal de 2021 forneceu uma visão da interrelação entre estas duas fontes de dados e ajudou a compreender o contexto da sua utilização. A análise final das viagens originadas na periferia e mapas produzidos com correlação aos resultados obtidos antes concluiu a representatividade da análise do estudo de caso em relação aos padrões de comportamento da população.

PALAVRAS-CHAVE: dados móveis, processamento de dados, planejamento de transporte, análise espacial, padrões de viagem.

ABSTRACT

The present master's dissertation investigates the valuable role of mobile operators' data in transport planning, addressing its potential to revolutionize urban mobility management and enhance transportation systems. This research endeavors to analyze the utilization of mobile operators' data in the context of transport planning and its implications for urban mobility.

This dissertation has three main objectives. The first is to research and analyze existing literature on the topic of mobile operators' data use in transport planning and other related areas of studies. The second objective is analysis of existing techniques, mechanisms and methods dedicated to obtaining and processing of mobile data for scientific research purposes. The final objective devoted to application of collected knowledge to develop case study project for Tâmega e Sousa region.

Based on the reviewed literature mobile operators' data found an application in various areas of geographical studies such as spatial planning, behavioural studies, statistical analysis, mobility studies and transport planning.

In-depth analysis of application of Origin-Destination matrices in the mobility research shed a light on utilization of mobile data in studies focused on understanding people travel behaviour. Last part of methodology analysis included assessment of methods of data-processing in the context of mobility research, such as Itinerary Reconstruction, Transport mode detection and Location patterns recognition.

Case study part of the master's dissertation focused on data analysis project for Tâmega e Sousa region in Portugal. The main focus was on analysis of morning and evening rush hour patterns in the core study region. Later comparison of data obtained from mobile operator with data from 2021 national census of Portugal provided insight in the interrelationship between these two sources of data and helped to understand the context of their utilization. Final analysis of trips originated from periphery zone and produced maps with correlation to obtained before results concluded the representativeness of case study analysis in connection with behaviour patterns of population.

KEYWORDS: mobile data, data processing, transport planning, spatial analysis, travel patterns.

GENERAL INDEX

INDEX OF FIGURES

INDEX OF TABLES

SYMBOLS, ACRONYMS AND ABBREVIATURES

SWOT – Strengths Weaknesses Opportunities Threats

GIS – Geographic Information Systems

GDP – Gross Domestic Product

GPS – Global Positioning System

LDA – Latent Dirichlet Allocation

CDR – Call Data Records

IMSI – International Mobile Subscriber Identity

LAC – Location Area Code

LAU – Location Area Update´

LBS – Location Base System

OD – Origin Destination

AFC – Automatic Fare Collection

LBSN – Location Based Social Networks

HMMs – Hidden Markov Models

SVM – Support Vector Machines

INE – Instituto Nacional de Estatística

CSV – Comma Separated Value

SQL – Structured Query Language

1

INTRODUCTION

1.1. Justification of chosen topic

The rapid advancements in mobile technology and the widespread usage of smartphones have generated an unprecedented amount of data, presenting various opportunities for its application in diverse fields. One area that stands to benefit significantly from the utilization of this data is transport planning. As cities worldwide grapple with escalating urbanization and increased transportation demands, innovative and data-driven approaches to enhance transportation systems become imperative. This master's dissertation aims to explore the potential of mobile operators' data use in transport planning, contributing to more efficient and sustainable urban mobility solutions.

Traditional methods of transport planning have often relied on limited and periodic data sources, leading to inefficient decision-making and suboptimal transportation services (Ewing & Cervero, 2010). In recent years, a paradigm shift has occurred, emphasizing the adoption of data-driven strategies to tackle modern transportation challenges (Basiri et al., 2018). Such an approach allows for real-time insights, dynamic analysis, and informed decision-making, promising improved mobility, reduced congestion, and enhanced user experiences.

Mobile operators' data encompasses vast amounts of information regarding users' mobility patterns, including the frequency, duration, and spatial distribution of trips (Toole et al., 2015). The ubiquity of mobile devices ensures a broad coverage of the population, generating a comprehensive and representative dataset. Additionally, this data exhibits an unprecedented level of granularity, offering insights into individual travel behavior, which is invaluable for transport planning purposes.

Mobile operators' data has already demonstrated its utility in various transportation-related applications. For instance, studies have successfully employed this data to model and predict traffic flows (Calabrese et al., 2013) and analyze public transit ridership patterns (Tseng et al., 2017). Moreover, it has been harnessed to assess the impact of urban events on transportation systems (Kung et al., 2014) and optimize public transit routes (Sun et al., 2016).

While the potential benefits of using mobile operators' data in transport planning are significant, it is crucial to acknowledge and address the associated data privacy and ethical concerns. Research on anonymization techniques (De Montjoye et al., 2013) and data aggregation methodologies (Machanavajjhala et al., 2008)

can provide guidance on safeguarding individual users' privacy while still extracting valuable insights from the data.

Transport planning profoundly influences urban sustainability, as it directly impacts energy consumption, air quality, and overall environmental performance (Barth, 2018). By leveraging mobile operators' data, planners can gain a deeper understanding of travel behavior and preferences, leading to more effective strategies in promoting sustainable transport options, such as public transit, cycling, and walking.

In conclusion, the proposed master's dissertation topic on the topic of mobile phone operators' data use in transport planning holds significant relevance and potential impact. With the increasing recognition of data-driven decision-making and the widespread availability of mobile data, this research can contribute to the advancement of transport planning practices, ultimately leading to more efficient and sustainable urban mobility solutions.

1.2. Methodology and objectives

This section of Introduction presents the methodology employed in this master's dissertation on the use of mobile operators' data in transport planning as well as the main objectives of the work. The research process involved several distinct stages, each contributing to the achievement of the research objectives. The methodology encompasses literature review, SWOT analysis, data procession, data analysis and spatial analysis.

The research commenced with a literature review aimed at gaining a comprehensive understanding of the topic's various dimensions. Literature from diverse areas of study, such as spatial planning, behavioural studies, statistical analysis, mobility studies and transport planning, was overviewed. By examining existing research, this study identified gaps, limitations, and potential avenues for the application of mobile operators' data not only in transport planning but in other areas of studies as well. In order to assess the strengths, weaknesses, opportunities, and threats (SWOT) associated with the use of mobile operators' data in transport planning, a SWOT analysis was conducted. Overview of different methods and mechanisms of data mining and procession shed light on an algorithm of utilization of mobile data in transport research.

Next part of the study focused on data procession. Raw datasets from mobile operators often contain errors, missing values, or inconsistencies, which can impact the validity and reliability of subsequent analyses. Data procession techniques were applied to enhance the quality of the data. Data analysis was conducted to derive valuable insights and relevant metrics that could be used for transport planning purposes. The processed data was subjected to various analytical techniques as well as compared with data from existing official statistics to identify patterns and trends related to transportation behavior and mobility.

Spatial analysis techniques were employed to study the geographic distribution of mobility patterns, traffic congestion, and travel behavior within the study region. Geographic Information Systems (GIS) and spatial visualization methods were used to create informative maps, aiding in the interpretation of complex mobility data.

Thus, this master's dissertation is developed around three following objectives:

a) Research and analysis of existing literature devoted to use of mobile operators' data in transport planning and other related areas of studies;

b) Analysis of existing techniques, mechanisms and methods dedicated to obtaining and processing of mobile data for scientific research purposes;
c) Application of collected knowledge to develop case study project for Tâmega e Sousa region.

1.3. Structure of dissertation

The structure of this master's dissertation is divided in 5 chapters. *Chapter 1* is Introduction, which contains general framework of chosen topic and its relevancy, as well as methodology and main objectives of work.

Chapter 2 presents literature overview and divided in 7 sections. In section 2.1. the overview of main countries, institutions and areas of studies that use mobile operators' data in scientific research is presented. Section 2.2. contains overview of studies conducted in the area of statistics. Section 2.3. presents review of studies in the area of spatial planning and settlement analysis. Section 2.4. includes outline of bibliography in the sphere of social and behavioural studies. Section 2.5. reviews use of mobile operators' data in projects in developing nations. Section 2.6. contains overview of literature dedicated to mobility studies. In the Section 2.7. SWOT analysis of mobile operators' data conducted by author is presented.

Chapter 3 is divided in 3 sections and each of the sections are divided in 3 subsections. Section 3.1. includes overview of different data mining mechanisms. Section 3.2. contains review of data pre-processing methods. In the section 3.3. overview of data procession algorithms is presented.

Chapter 4 is separated in 3 sections. Section 4.1. presents overview of Tâmega e Sousa region study project. Section 4.2. outlines data processing steps. Section 4.3. is divided in 4 subsections and presents results of the case study project.

Chapter 5 present conclusion of the master's dissertation summarizing all findings that were obtained during the literature review, analysis of existing methods and case study and linking them to the main objectives of the work.

2

OVERVIEW OF MOBILE OPERATORS' DATA USE IN DIFFERENT AREAS OF STUDIES

2.1. World leaders and primary usage directions of mobile operators' data

To date, the data of mobile operators have firmly occupied an information and statistical niche as a source for numerous businesses, scientific, international, and government organizations. The United States and European nations such as France, Belgium, the United Kingdom, and Estonia conduct the broadest range of studies involving the use of information from mobile operators. Due to its high level of informatization of society, relatively small size, and established long history and tradition of working with "Big Data" the last functions as a unique "experimental laboratory" for a wide variety of scientific research based on the geolocation of mobile phones. We can highlight Italy, Germany, Austria, Italy, the Czech Republic, Switzerland, Portugal, Spain, Sweden, Israel, Japan, and China among other significant nations. Thus, research focused primarily on Europe, the United States, and several Asian countries. We can separate the main areas of studies that can benefit from using telecommunication data in 4 different categories:

- Accumulation, organization and analysis of statistics;
- Spatial planning and settlement systems analysis;
- Behavioural, social, demographic and population distribution studies;
- Transport planning and mobility studies.

The research groups, laboratories, and other scientific departments of the MIT in Boston and the University of California at Berkeley (USA), the University of Tartu in Estonia, the Ghent and Louvain universities in Belgium, the Fraunhofer Institute in Munich, Orange Lab in Paris, and IBM Research in New York are among the scientific organizations using the data of mobile operators for fundamental and applied research. Austrian Institute of Technology and Vienna University of Technology, Karolinska Institute in Stockholm, the Zurich Institute of Cartography and Geoinformation, Israel Institute of Technology in Haifa, National University of Ireland in Maynooth, Polytechnic University of Milan, and the Universities of Liege, Prague, Pisa, Oxford, and Cambridge conduct research to a lesser extent. Some individual studies involve specialists from numerous other scientific and educational institutions in the United States, Spain, France, the United Kingdom, the Netherlands, Hungary, Japan, China, and India.

2.2. Accumulation, organization and analysis of statistics

Mobile data becomes more popular in population statistics each year. There are several traditional sources of demographic statistics, including household surveys, annual censuses, and administrative population registers. Even if it is considered to be quite detailed and generally accurate, the data extracted from the current demographics is quite static. Another disadvantage of this type of statistical source is that these registers are typically only published once a year, which is insufficient for developing time series with high detail. Numerous state statistical agencies are therefore keenly interested in the potential use of alternative data sources. Despite the fact that mobile operator data is not as precise as national census data, a comparison of these sources reveals a high correlation.

A group of researchers from France and Spain concluded in the study published in 2014 that the use of three data on population from three different sources (census data, social media, and mobile phones) can provide information with very high correlation (it is worth noting that mobile phone data correlates with census data more than social media data does) (Lenormand et al., 2014). The use of clustering methods can lead to a precise determination of the location of residences, workplaces, and places of recreation and leisure, as well as aid in comparing the results with official statistical data and developing recommendations for their improvement (Csaji et al., 2013; Tiru, 2014).

In comparison to other types of data, the data from mobile operators offers an enticing trade-off: it is highly sensitive to the movements of users while also being precisely localized in space. Based on the mechanism behind how this type of data locate users in real-time, mobile positioning information makes it possible to implement the concept of emphasizing points of indication within the framework of so-called "social time" (Ahas et al., 2015) in addition to in terms of methodology. The total population and population density are the main demographic valuables in the most social-demographic researches and studies. International and state institutions, as well as administrative and research organizations, have devised methods for studying population distribution and density based on data from mobile phone operators since the beginning of the twenty-first century. Due to the characteristics of the data (cost and level of penetration), initiatives at the municipal level were the starting point for examining the data's potential applications.

One of the first projects carried out implementation of mobile phone data was conducted in 2005 (Ratti, 2005). Researchers carried out systematical analysis of information regarding population distribution changes and labor commuting based on "A1" mobile phone operator data from Austria. The visualization phone activity for this project is shown in the Figure. 2.1. Tourism was the one of the first industries interested in using data collected my mobile operators for its needs. Eurostat have ordered global research to investigate how information regarding tourist flows can be obtained from the data collected my mobile operators (Eurostat, 2014). Special emphasis was placed on identifying the benefits and drawbacks of accessibility, relevance, and data cost, as well as the technological and methodological issues that arise from the use of mobile data. According to the study, this type of data can serve as an effective tool in implementing research for touristic industry.

The geopositioning of mobile phone data is also severely hindered by regulatory restrictions that are primarily attributable to regulatory and legal differences between states. A conclusion was reached that, in order to collect statistics on the cross-border movements of European Union citizens, it is necessary to establish a central structure capable of collecting national-level statistics and compiling them according to a common European methodology. As a result, mobile phone data could be useful for measuring a variety

of touristic indicators as a supplementary source without necessarily supplanting the primary traditional statistical sources.

Using this sort of data to supplement the existing methods could be a valuable option. The sample size of sociological surveys of visitors, which can be quite costly, could be reduced, for instance, by collecting information in a mixed manner. As additional benefits of using telecommunication data, scientists cited timeliness, improved spatial and temporal correspondence with reality, and the ability to calibrate existing data. Mobile positioning data enables the retrospective linking of specific events and locations with digital footprints left by visitors.

Additionally, experience from the so-called the Estonian "tourist barometer" demonstrates that monitoring could be used as a tool for planning and administration of the entire tourism sector (Ahas et al., 2007; Jarv, 2013). It was possible to obtain information about the number of travelers and their country/region of origin by conducting a deep analysis of mobile data on Czech Republic numerous popular touristic domains. This data was subsequently used to construct a tourism marketing plan and strategies (Vogelová et al., 2012).

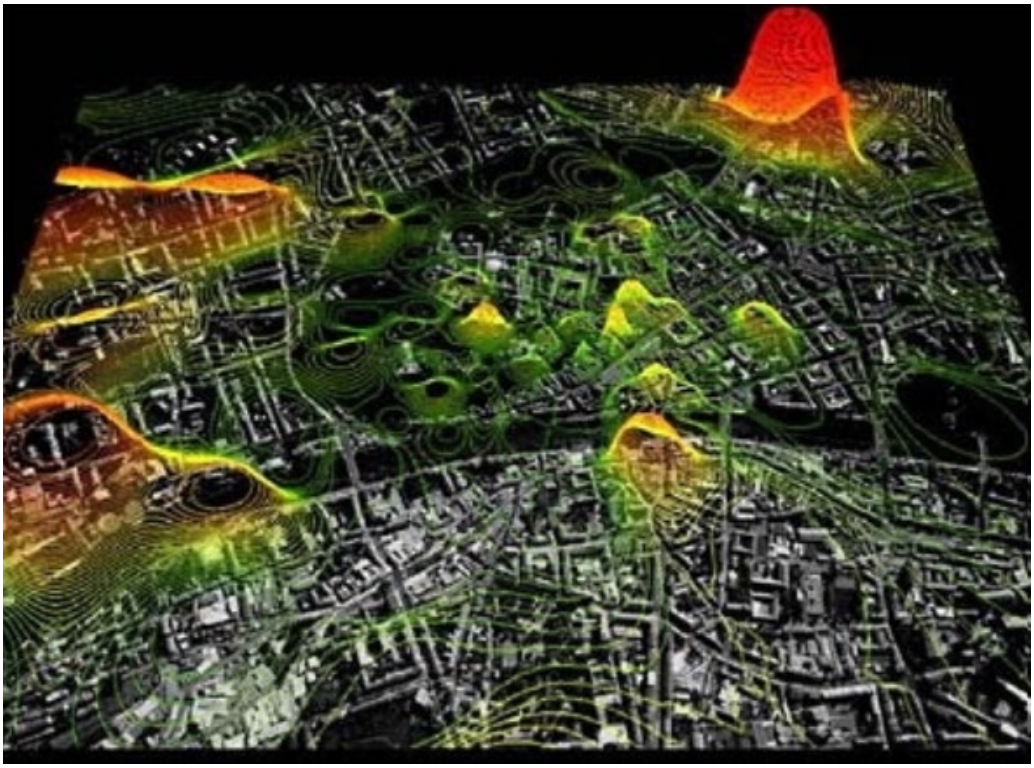


Fig. 2.1 – Street map of Graz, Austria, overlaid with an electronic visualization of cellphone activity

Source: (Graz in Real Time, 2005)

In the 2010s, numerous studies in the area focused on international tourists and the evaluation of the attractiveness of tourist destinations in countries such as Germany, Montenegro, France, Japan, Ireland, China, Indonesia, etc. were published. The most tourist-attractive countries are anticipated to demonstrate the greatest interest in incorporating mobile positioning data into national tourism statistics. In 2015, an experiment was conducted in France to examine the viability of using telecommunication data to compile

tourism statistics and the likelihood of these data replacing conventional methods of collecting statistics. The experiment provided answers to a number of questions pertaining to the most pressing issues associated with this type of geoinformation (access to data, regulatory constraints, methodology, and the quality of assessments), laying the groundwork for the pan-European trend of incorporating data from mobile operators into statistics. In 2016, comparable research was conducted in certain regions of Indonesia. In general, French and Indonesian studies have identified the advantages and disadvantages of this information source. As one of the vulnerabilities of mobile phone data, the absence of descriptive data about the mobile subscriber and his reason for visiting the city or territory was cited. At the same time, experts underscore that the primary problems for the statistical authorities were not methodological issues, but rather barriers to data access, its confidentiality, and relatively high "initial costs" of using information based on mobile networks.

In addition to the discovered benefits of using data from mobile phone operators, we can also designate the high accuracy of measuring short-term fluctuations (typical of tourist visits) and the economy, as compared to the quarterly data on tourist traffic currently used in western statistics. It is worth noting that there is competition with another type of Big Data that can give us insight into travel movements: bank card information, which is already being used to track short-term trends (it is worth noting that bank card data is obtained monthly, whereas mobile operator data is obtained daily).

Gradually, supranational initiatives to integrate data from mobile operators into statistics began to emerge, following the emergence of national initiatives involving the use of a new source of statistical information. The EU launched the largest Big Data undertaking in recent years in 2016 with the ESSNET initiative (ESSnet Big Data, 2021). Its primary objective is to complement regular development of formally published statistics by integrating data from mobile operators using pilot studies. The primary objective of the aforementioned studies is to develop and implement specific applications in a variety of fields (registration of visitor flows, assessment of population mobility, etc.).

2.3. Spatial planning and settlement systems analysis

Mobile data about population return movements are valuable for settlement system studies, such as those examining agglomeration structures. Using data from telecommunication operators regarding population movements on daily and weekly basses, it is possible to comprehend the boundaries of zones of interaction between various territories and to delimit suburban and core agglomeration areas. All of this is consistent with contemporary approaches to delimitation of agglomerations.

As a starting point to study and delimitation of agglomeration structures with the help of mobile phone data we can consider the work the study of "home–work" systems (on the example of Estonian cities) performed by researchers of Tartu University and construction on the base of this work of correspondence matrices (Ahas et al., 2009). The same researchers later developed a methodology for the allocation of functional areas and included the data of mobile operators in existing approaches of highlighting fundamental components of the settlement system of Estonia (Novak et al., 2013).

Allocation of agglomerations and the research of the organization of urban spaces on functional and structural levels with use of data from mobile operators has found wide application in Czech Republic, where such studies in the last few years have been carried out in territorial planning purposes. Simultaneously, the first experience of successful application of mobile phones data in the process of Prague

delimitation have led to the increase in interest of other agglomerations of the country and leaded in allocation of different urban areas in the whole Czech Republic (Ouředníček et al., 2019). Related work was done by collective of Spanish scientists in 2014 and included 31 Spanish cities, but with more emphasis put on territorial structures on intra-urban scale (Louail et al., 2014).

Study of the functional and structural profile of space is tightly connected with the analysis of pulsations of population on the daily and weekly basis that happens in the agglomerations and in the cities. For example, the work of a group of Czech scientists is devoted to the Prague metropolis population fluctuations which based on the data from mobile operators was engaged in establishing different sorts of patterns throughout the day for different areas of the region of Prague on the base of correlation with commercial and residential real property (Nemeškal et al., 2020). Researchers found that cell phone data helped to differentiate parts of the city as transport, service, working and residential kinds of districts. It could be possible to understand in which parts of the city user reside or work based on information on how long and at what time they stayed. Study of these daily rhythms gave knowledge about real use of various parts of the city that were later used for spatial planning.

The work of an international team of scientists published in 2014 provides an example of a study of extended seasonal cycles of people's activities (Deville et al., 2014). Using datasheets from mobile operators, a group of researchers analyzed over one billion call records in France and Portugal to determine the population density. Scientists have demonstrated conclusively that the summer population of settlements comprising the largest urban agglomerations is lower than in other seasons. On the other hand, many rural, coastal, and mountain resort areas observed a substantial summer population increase. The data from mobile phones allowed for the identification of summer and winter areas-attractions, the scale on which these attractions occur, and also the objects that radically alter seasonal rhythms. In the Ile-de-France region almost all municipalities have bigger population density in the winter than in the summer except the ones surrendering the area of biggest international airport of Charles de Gaulle, the Versailles palace complex, and the amusement parks of Disneyland and Asterix Park, where this pattern is broken. In a similar manner, the Costa de Caparica resort region near Lisbon has a greater summer population than winter population. In the Figure. 2.2 the seasonal changes in population distribution in Portugal and France are shown.

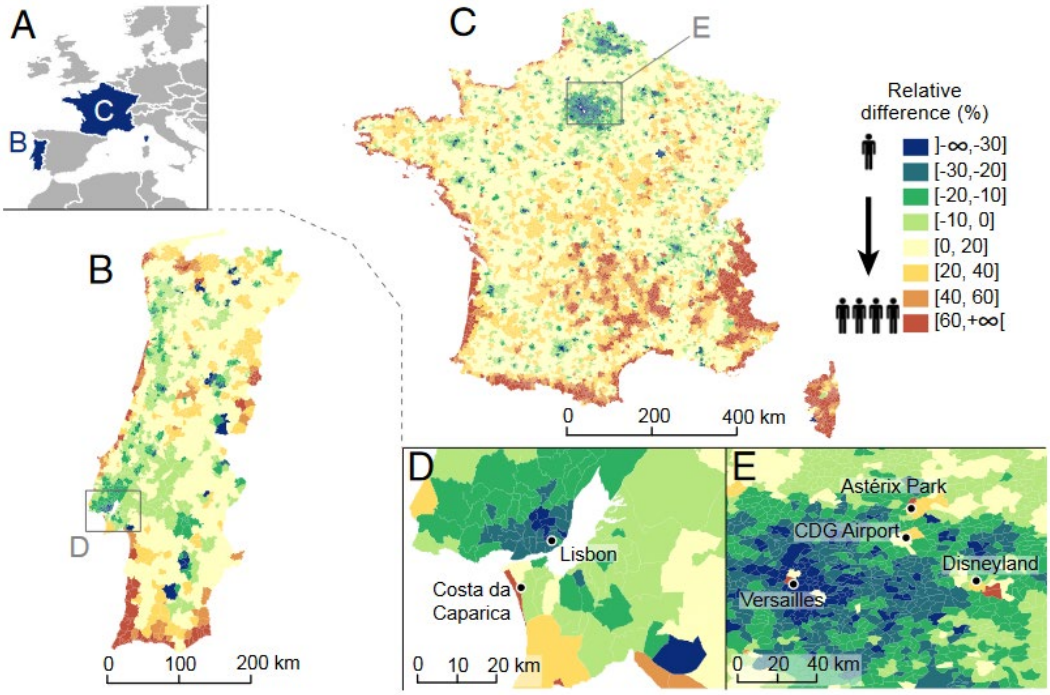


Fig. 2.2 – Seasonal changes in population distribution in Portugal and France

Source: (Deville et al., 2014)

Using Big Data to analyze how people interact in space and the patterns of this interaction can help to objectively identify the boundaries of settlements. Therefore, we should mention the work of a group of Belgian and British researchers on "natural" segmentation based on information from mobile operators in their respective countries (Blondel et al., 2010; Ratti et al., 2010). On the basis of these studies, we can conclude that the "mobile" dividing corresponds well with the municipal boundaries, but in a number of cases, previously only theorized unanticipated spatial structures were discovered. Mention should be made of the potential for a more in-depth analysis by supplementing mobile phone operator data with additional user information. By contributing to the research on the language of communication, researchers from Belgium contributed to understanding of certain sociocultural aspects of communication. In addition to Belgium, American research (Blumenstock and Fratamico, 2013) reveals a significant interest in the use of data provided my telecommunication operators for ethnic and ethnocultural zoning, as well as the study of segregation based on ethnicities.

2.4. Behavioural, social, demographic and population distribution studies

By applying and analyzing information based on mobile data metrics (average volume of calls per region, call destinations, etc.) with a variety of socioeconomic variables (such as income level), statistical models are developed that can provide us with information about trends and patterns in the territory's development. In terms of social and property differentiation, the mobility focused studies make the clearest use of

telecommunication data. For example, paper published in 2010 (Eagle et al., 2010) that focuses on social contacts, demonstrates that the variety of relationships between individuals have a close relation with local communities' economic development. The relationship the humans' social networks and their mobility is discussed in the work of researches from Portugal which based on data from mobile operators shows the connection between the intensity of social connections and human movements (Pithakkitnukoon et al., 2012). Technological capacities of mobile communications combined with the behavioral theoretical base led to the emergence of behavioural research about urban areas (Ahas et al., 2010). At the same time due to the limited possibilities of obtaining highly detailed knowledge about mobile users, works in this direction were most of the time carried out in within certain cases, considering individual characteristics in limited sample conditions.

Urban Behavioral models based on data from mobile operators were developed for variety of major cities such as Boston, Los Angeles, New York, London, Singapore and Beijing (Calabrese et al., 2013; Jiang et al., 2017). In them with the use of digital footprints of mobile subscribers, a picture of points of attractions of individual components of the urban fabric was drawn, which served as tool to a better understanding of the spatiotemporal specifics of the functioning of urban space. A lot of works are focused on the how individual social strata behave. Here it is worth highlighting the study on mobility in terms of age (on the example of Estonia and Prague), ethnicity (on the example of Russian and Estonian-speaking inhabitants of Tallinn), sex (on the example of inhabitants of the suburbs of Tallinn), nationality (on the example of foreigners in Milan) groups (Novak and Temelova, 2012; Silm et al., 2013; Silm and Ahas, 2014; Bajardi et al., 2015; Masso et al., 2018). Behavioral models also include long-distance travel models built based on data from mobile operators in Israel, USA and UK (Bekhor et al., 2013; Calabrese et al., 2013; Birkin et al., 2017). In them, by comparing data on trips received from mobile operators with other socio-economic information, scientists have expanded our understanding of people's behavior during long trips. In particular noticeable differences were found in the actual structure of behavioral preferences from the declared ones (obtained as a result of sociological surveys before trips). Such results showed an underestimation of the impact of the "situational" factor on changing the original plans of travelers. In the Figure. 2.3 the Spatial distribution of daily mobility at block group level in Boston is presented.

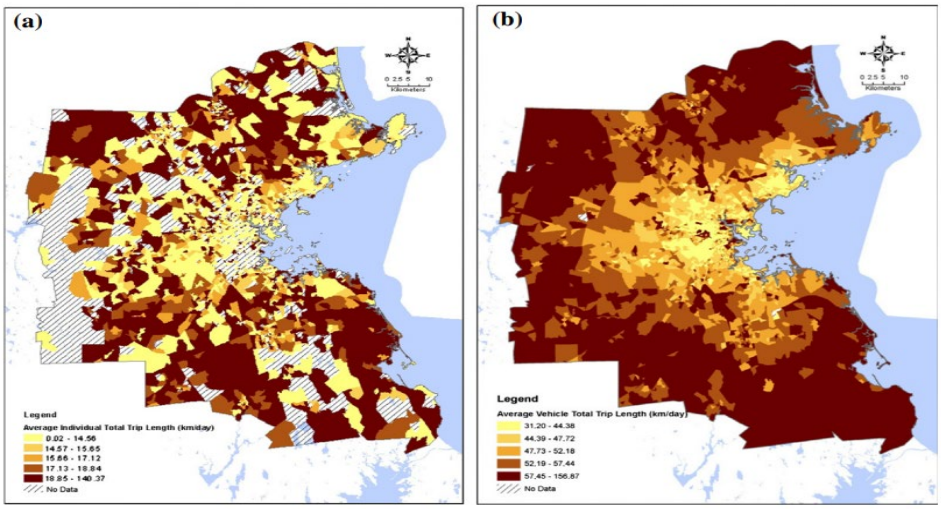


Fig.2.3 – Spatial distribution of daily mobility at block group level in Boston

Source: (Calabrese et al., 2013)

In the framework of the "Real Time Rome" project, on the example of people gathering for the concert of Madonna and to celebrate the victory in the final of World Cup of Italian national football team (Real Time Rome, 2006), analyzing the effects of different events on the existing settlement systems is a promising direction for the use data provided my mobile operators. In the Figure 2.4, there are clearly visible activity peaks in the center of Rome after the Italian national football team won the 2006 FIFA World Cup final (the national team returned to Rome on the evening of 10 July). During the 2012 Summer Olympics in London, a comparable project on an even grander scale was intended to be executed, but was ultimately canceled. The Festival of Light in Ghent, Belgium, was another cultural event that was the subject of research (Versichele et al., 2012). All designated projects were concentrated on attraction points and their spatial organization. One of the most significant conclusions made by researchers in this study was the confirmation of the "attenuation" of distance principle, which was formulated by G. Allson and was based on the first law of V. Tobler (Tobler, 1970; Olsson, 1970), which states that increasing distance from any event decreases the number of visitors for this particular event. In addition, researchers have identified seasonal patterns, such as the rise of long-distance travelers during the off-season and winter.

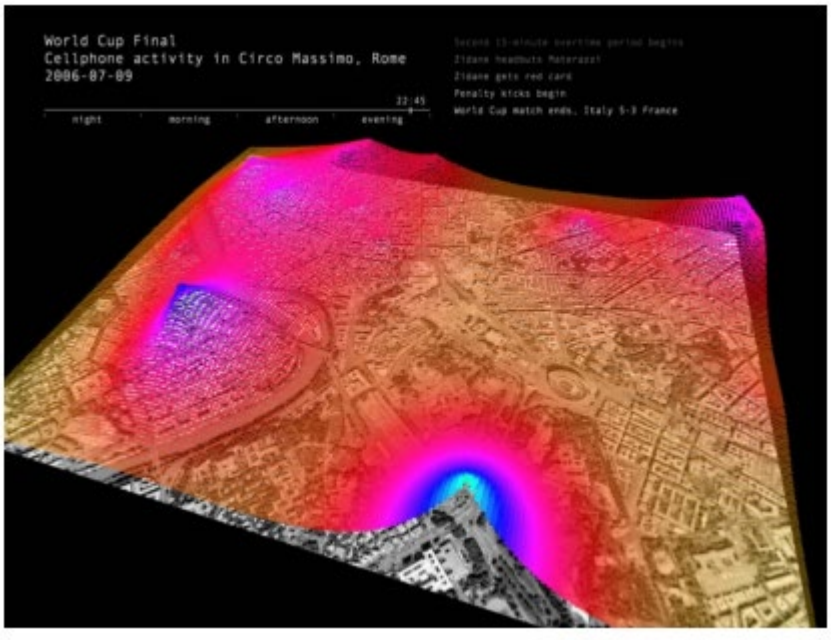


Fig. 2.4 – What does Rome look like during special events

Source: (Real Time Rome, 2006)

As an example, for the paper focused on political event and utilizing the use of telecommunication data, we can cite the study of the so-called "March of the Million" protest in Israel. With the help of mobile phone data and new sociological tools, it was possible to determine social distribution of the protest (it was discovered that a large proportion of protestors came from the lower income classes and that half of the protestors resided in Tel Aviv) (Stelman, 2012). Thus, information from mobile phones may not only provide valuable information for research but also aid in avoiding various political speculations (such as frequently occurring disputes regarding the overall number of protesters).

2.5. International initiatives in developing nations

Particularly in the disciplines of demographic and socioeconomic development, research on the development of developing countries demonstrates a keen interest in the possibilities of utilizing data from mobile operators. In a number of these nations (particularly those impacted by epidemics, natural disasters, conflicts, etc.), the national survey to determine population indicators has not been conducted in years. Consequently, the population's size, structure, and distribution appear to be highly approximative. In addition, "inaccurate" demographics can contribute to significant errors in calculating key economic indicators such as GDP. In such states, according to the Director of Financial Advisory and Banking Services of the World Bank, M. Jugale, “Big Data” is of increasing importance for the use of national statistical services, allowing you to "jump over" the stage of traditional methods of information collection and proceed directly to collecting statistics using satellite images and data from mobile phone operators. In a milder form, this type of "statistical breakthrough" is applicable to developing nations, whose statistical bases lag far behind their American and European counterparts. Too rapid urbanization is a challenge for developing countries. In these nations, the development of infrastructure outpaces urbanization, putting excessive strain on the extant road network. As a consequence, there is an extremely high pressure on the transport systems which leads to congestion, which at the same time could have a negative effect on overall human development and economic growth as people spend more time on the commute for their daily activities. Research done by the scientists from IBM shows possibility of monitoring the journeys of individuals and using mobile phone data to better plan and manage transportation services (Berlingerio et al., 2013).

In Morocco, Sri Lanka, Senegal, India and Bangladesh, mobility studies and the construction of transport models were conducted with the assistance of European and American experts. Results of study published in 2014 conclude that in developing countries, city dwellers travel longer distances than suburbanites in their daily commutes, which is not typical for residents of European or American metropolitan areas. (Smith -Clarke et al., 2014; Scepanovic et al., 2015) An international group of researchers developed the methodology to study poverty levels using mobile phones data as part of a project to examine the social and economic condition of Côte d'Ivoire. Various patterns of citizen mobility were developed for this country based on the frequency of calls that have been done throughout the different times of the day. The researchers then contrasted the resulting models with information from other sources (news about specific events during the period under review, census data, economic activity, the poverty index, information on power plants and power systems, etc.). The results demonstrated a high correlation between numerous indicators, disclosing indicators derived solely from phone call data. Concurrently, the issue of the high cost of obtaining statistical data in poor countries was resolved.

Real-time mobile phone data can be a vital source of information when emergency services require current information about peoples in peril location and movements, such as during natural disasters, armed conflicts, and epidemics (Deville et al., 2014). During the Haiti 2011 earthquake, data from the mobile operator Digicel was used to produce highly accurate estimates of population displacement (Lu et al., 2012). As demonstrated by the example of Haiti, information on population displacement during or after natural disaster can be of great assistance in organizing the distribution of medication, water, and food. In situations where there is a risk of disease transmission, telecommunication data could help to assess the geographical potential of disease transmission, import routes, and define quarantine zone parameters. In the context of studies on diseases and the nature of their spread, it is important to mention the work of Swedish scientists who used mobile phone data to determine how malaria spreads in Kenya (Lu et al., 2012; Tizzoni et al.,

2014). By analyzing migration of people during the epidemic, Swedish researchers have developed a model of the disease's import route.

2.6. Transport planning and mobility studies

One of the most rapidly expanding applications of mobile phone data is seen in mobility studies and transport planning. Sweden, Finland, Spain, and Germany, as well as a number of other nations, produces studies on transport traffic analysis using telecommunication data. Automobile navigation equipment and software have utilized mobile phone data in tandem with GPS navigation for quite some time.

Operator-collected data are not the only ones used in urban and transportation research. For example, information regarding the use of travel documents (cards) by passengers can provide information regarding direct routes (Namiot D. et al., 2018). In certain instances, mobile phones with specific programs installed can collect location data.

As more and more of the public transport vehicles are being equipped with the GPS systems, the case for implementing the final point regarding data integration becomes even stronger. The exchange of data will allow for the estimation of passenger transportation burden. When it is possible to analyze a vehicle's GPS traces, the synergistic effect will be particularly significant. Counting motorists is typically a complex but highly relevant urban analytics task.

IMT is one of the first institutions that successfully utilized use of data collected by mobile operators in transport research, and the 2006 study “Real Time Rome” (Real Time Rome, 2006) was one of the earliest practical works on this path, followed by a series of studies on urban monitoring. The proposed platform initially facilitated the monitoring of mobile device motion (or the determination that no motion was present). Results of this work could show as a different information about population and its mobility:

- Where people spend their day;
- Concentration and migration of people during special events;
- Which of the city's historical sites attract the greatest number of tourists;
- Transportation pedestrian and mobile phone (users) traffic;
- Movement and concentration of foreigners (they were determined based on the IMSI (International Mobile Subscriber Identity) phone numbers of foreign mobile network operators).

Consider the aforementioned paper in greater detail (Deville, Pierre, et al., 2014). Based on operator-recorded mobile device activity, the primary objective is to estimate population density. The mathematical apparatus is a Voronoi diagram (Aurenhammer et al., 1984).

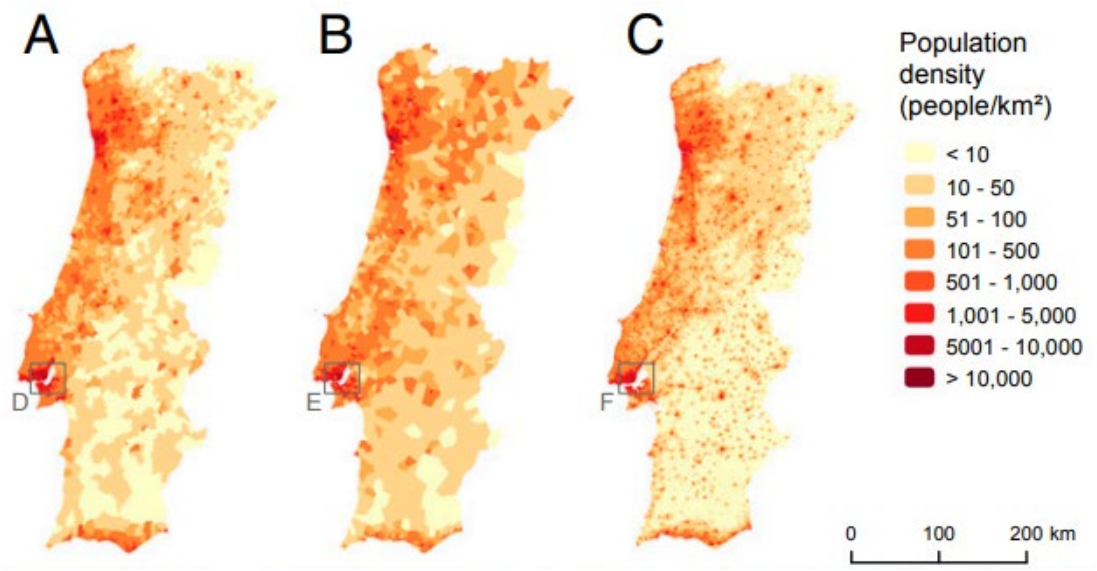


Fig. 2.5 – Comparison of predicted population density datasets for Portugal

Source: (Deville, Pierre, et al. 2014)

In the Figure. 2.5 a comparison of predicted population density datasets for Portugal is presented. The variation in subscriber density suggests that a number of users have relocated. In this instance, the territory is a geo-square, which is a geographically delimited region. The minimum limits of such areas are determined by the network's technical capabilities (the characteristics of base station locations). Obviously, small-location data can be aggregated. In particular, up to specific administrative entities. Therefore, we have pairs of regions (origin-destination) along with the number of relocating subscribers for each pair. Clearly, these mobile subscriber movements should be reflected in transport activity as a subscribers relocate to a new region. Then it is possible to count values of this “relocations” for a different scenarios (time of year, time of the day, movement patterns, etc.).

In the study done by Farrahi and Gatica-Perez (2011), user movements are classified using model based on LDA (Latent Dirichlet Allocation). To illustrate the user's movements, each location is initially classified as home-based. In addition, subscriber movements are recorded every 30 minutes according to the newly introduced classification, three steps of these classification are shown In the Figure. 2.6.

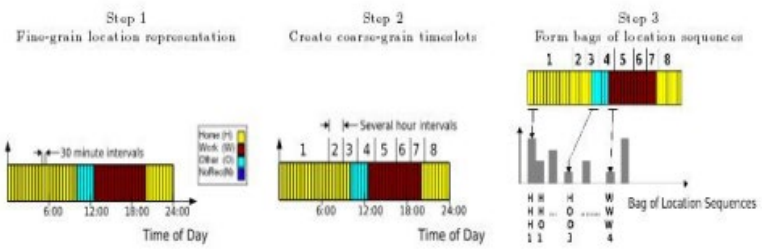


Fig. 2.6 – Three steps of movements classification

Source: (Farrahi and Gatica-Perez, 2011)

Article published in 2016 (Elias, Daniel, et al., 2016) describes one of the outcomes of the SOMOBIL initiative. This is the planning framework for transportation, based on data from telecommunications companies. A result of this work is the restoration of modes of transportation utilized by mobile subscribers when traveling. Another work published in 2013 (Schneider, Christian M., et al., 2013) examines movement patterns. The authors contend that 17 distinct patterns adequately characterize movements of 90 percent of the world's population. Moreover, an individual's movement patterns are typically stable for several months. The authors conclude that Modeling Markov Chains for periods of high frequency travel followed by periods of lower activity can be used to simulate the daily mobility of humans. The authors note that models of human mobility at varying scales influence contemporary society and the environment.

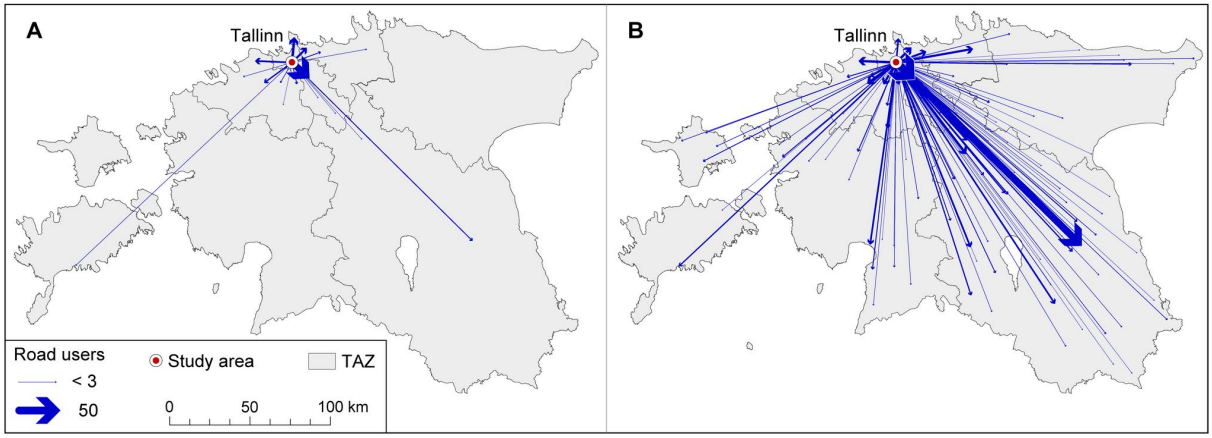


Fig. 2.7 – Spatial distribution of the last call activity of the day in Estonia. A: Monday-Thursday, B: Friday

Source: (Järv et al., 2013)

Is it also worth to mention a study carried out for Tallin by Estonian researchers to identify who are the main "perpetrators" of the road network congestion in evening rush hours (Järv et al., 2013). The results of this study disproved the dissertation of leading role in Friday traffic jams of residents of the suburbs and, quite the opposite, the main role of city residents traveling outside of it with purposes of recreation. Spatial distribution of the last call activity of the day in Estonia is presented in the Figure 2.7. One of the most important result of the study was that it has shed light on the nature of traffic congestion in the Friday evening. Over 60% of trips of Friday evening rush hour (different from other days of the week) were not associated with work, and motives of motorists that day were outside their daily work activities. Thus, the largest traffic jam of the week was due to an increase in overall individual mobility in society, which coincides with some of the earlier research which leisure was designated in as the main motivation for the movement of people in post-industrial society (Schlich et al., 2004).

2.7. SWOT analysis of mobile operators' data

To conclude overview of use of telecommunication operator's data in variety of studies fields the SWOT analysis would be concluded, the results of which are shown in the Table 2.1.

Table 2.1 – SWOT analysis of mobile operators' data.

Source: (complied by author)

Strengths	Weaknesses
Large sample size: mobile operators collect vast amounts of data from millions of users, providing researchers access to extensive and diverse datasets, enabling comprehensive studies	**Data quality and bias**: mobile data may suffer from inconsistencies and biases, as it primarily reflects the behavior and preferences of mobile phone users, excluding those without access to mobile networks
Real-time insights: mobile data is generated in real-time, allowing researchers to analyze trends and patterns as they occur, leading to timely and relevant findings	**Limited variables**: researchers might not have access to all the relevant variables needed for a comprehensive study, as data provided by mobile operators could be limited to specific metrics
Cost-effective: utilizing existing mobile data eliminates the need for costly data collection efforts, making it a more economical option for research projects	**Regulatory and ethical concerns**: the use of mobile data for research may raise ethical and regulatory issues, such as ensuring data privacy and obtaining proper consent from users
Broad geographical coverage: mobile networks cover vast geographical areas, allowing researchers to study phenomena across different regions and urban-rural divides	**Data ownership**: mobile operators may have proprietary rights over the data, leading to potential restrictions on usage and publication
Anonymized data: mobile operators can provide anonymized data, protecting user privacy while still offering valuable information for research purposes	
Opportunities	**Threats**
Developing countries: mobile operators' data could be a variable and sometimes the only one reliable source of actual information about population for areas with restricted access to more traditional statistical sources	**Data security risks**: storing and sharing large-scale mobile data can expose sensitive information, leading to potential security breaches and privacy infringements
Flexibility: mobile data can be used in a broad spectrum of projects and studies	**Data access limitations**: mobile operators might not grant access to data or may impose restrictive terms, hindering researchers' ability to conduct certain studies
Data processing: advancement in methods and instruments of processing and cleansing of data can lead to a more accurate representation	**Technological barriers**: analyzing massive and complex datasets from mobile operators may require specialized tools and expertise, posing technical challenges for some researchers
Comparability with AI: mobile operators' data can be utilized in tandem with rapidly developing AI systems which could lead to a more fast and deep analysis of real-time processes	**Competing interests**: mobile operators may prioritize their commercial interests over research needs, leading to conflicts in data sharing and usage
	Public perception and trust: the use of mobile data in research can raise concerns among users about how their data is being used, affecting public trust in both researchers and mobile operators

3

DATA MINING AND PROCESSING

3.1. Data mining

3.1.1. Call Data Records

In the process of servicing telecommunications devices (mobile phones, etc.), telecommunication operators collect a large amount of service information. First of all, this data is necessary to support operators' own business processes. This is the main reason they collect this data. In particular, files (logs, magazines) with records of the details of conversations (interactions) – the so-called Call Details Recordings (also called Call Data Records, CDR) serves as the basis for billing. In other words, it is the basis for assessing (calculating) the economics of the operator. In fact, this is the main set of data required by the operator (Horak R., 2007). These records, of course, can be used for other purposes (for example, to search and verify information about the location of a particular subscriber, etc.), but it is the economic component that remains the main one. And this, in turn, means that such information becomes a relatively "cheap" source of data for third parties. There is no need to create your own data collection services, install and maintain sensors, etc., if you learn how to get useful information from operators' data, which the operators collect in any case and which are available for any place (districts, etc.) where there are subscribers of these operators. And the “penetration” of mobile devices ensures that data collection will automatically cover all regions. It is these considerations that determine the interest shown by researchers in relation to the mobile data.

CDR can be obtained from any type of subscribers’ activities, receiving or outgoing calls, SMSs, data transfer trough internet connection (i.e., 3G or 4G technologies). CDR store information about time and geolocation of the records. International mobile subscriber identity (IMSI) is anonymized according to the countries legislation on data privacy and encrypted in the anonymized-imsi (AMSI). When mobile device changes its’ location area code (LAC) a record is created from a location area update (LAU).

Typically, a CDR includes the following key information:

- **Calling Party Number (CgPN)**: The phone number or identifier of the caller initiating the communication;
- **Called Party Number (CdPN)**: The phone number or identifier of the recipient or the party being called;
- **Call Start Time**: The date and time when the call was initiated;
- **Call End Time**: The date and time when the call was terminated;

- **Call Duration**: The length of the call-in seconds or minutes;
- **Call Type**: Indicates whether the call is voice, video, SMS, data, etc.;
- **Call Result**: The outcome of the call (e.g., success, failure, missed);
- **Location Information**: The cell tower or geographical location where the call was made (LAC and Cell IDs).

In the Table 3.1 you can see an example of how rows with information from typical CDR may look like.

Table 3.1 – Example of rows from mobile operators' CDR

Source: (complied by author)

AIMSI	Time	LAC	Cell ID	Technology	Type
#.....	2022-08-07 13:05:46	854263	C5	3G	SMS
#.....	2022-08-07 13:24:13	854263	C5	3G	Start Voice
#.....	2022-08-07 13:35:05	352478	C6	4G	End Voice
#.....	2022-08-07 13:42:17	352478	C6	4G	Data
#.....	2022-08-07 14:00:25	145693	C7	2G	SMS
#.....	2022-08-07 14:05:36	145693	C7	2G	Start Voice
#.....	2022-08-07 14:07:12	248789	C9	4G	End Voice
#.....	2022-08-07 14:09:32	248789	C9	4G	Data

Obviously, data collected by operators is not the only type of metadata used in urban and transport research. For example, information about the use of travel documents (cards) by passengers can directly provide route information **(**Namiot D. et al., 2018). In some cases, information about movements can be collected by specially installed programs on mobile phones. The problem is that carrier data is hard to compete in coverage. Naturally, the movements of passengers (mobile phones) who move, for example, in a city train, will be reflected in the data of operators. From the data collected by the operators, it will be difficult to isolate exactly rail passengers, but these data will contain information on other movements for which there is no information on the validation of travel documents. What exactly do operators collect? It is important to note here that in this case we are talking about information related to the activity of their subscribers. In fact, the very fact of some kind of activity (call, message, etc.), as well as the place where it happened. CDR are typical metadata. The record includes information about the number that made the call, the number that received the call, call duration, device identification, etc. The location is determined using information about the base stations serving the mobile device at any time. Technically, the operator has the ability to link the data from the CDR to the user's profile and, accordingly, also use the information from the profile (gender, age, etc.), but this cannot be done for privacy reasons. In addition, for privacy reasons, such links are unlikely to be shared with third party data analytics organizations. Therefore, such data are not considered further, and age and sex can be estimated even without profile data (by similarity of behavior). Also, for privacy reasons, the identification of mobile devices is usually replaced with a one-way hash file. It should also be noted that in the CDR data provided for analysis (again, for confidentiality reasons), time stamps of events can be changed, coordinates are specially "blurred", etc. In the Figure 3.1 the simple mechanism of

depersonalization of mobile operators' data is provided, where personal information about subscriber (name, mobile phone number, etc.) ID is changed to an alternative ID (codified numerical value) using the chosen depersonalization algorithm *f*.

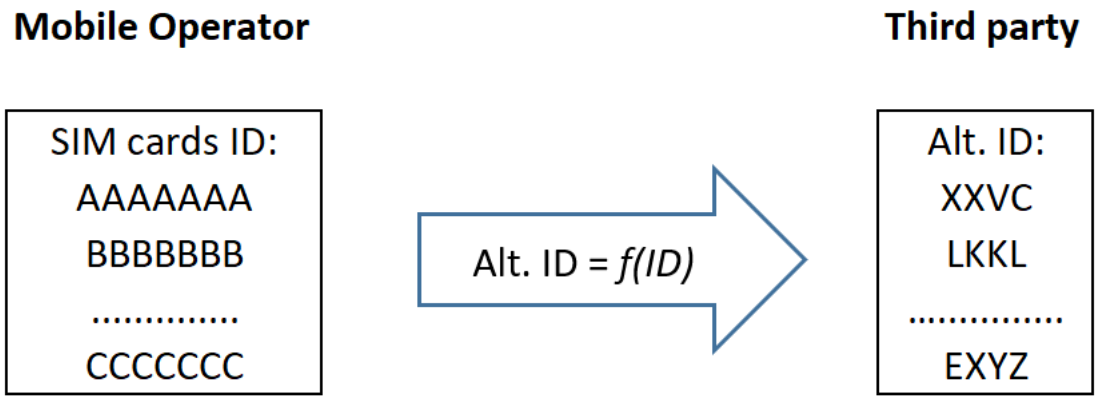


Fig. 3.1 – Depersonalization of mobile operator's data

Source: (complied by author)

In the more recent studies focused on mobile data there have emerged a new type of records, so-called Passive Call Data Records. They have the same structure as usual CDR but they can be collected from any data exchange between device and mobile network, while to generate usual CDR there is need in active usage of telecommunication (call, SMS) passive records can be obtain information about several locations even while device is not active. Due to that Passive CDR have a higher frequency of data collection and by that bigger volume of overall obtained information which can lead to a more precise geopositioning. Overall passive CDR are more suitable for a broad area of studies that would benefit from bigger data samples and more precise location records.

3.1.2. Geopositioning

Mobile operator data is primarily concerned with location. Mobile positioning involves monitoring the location coordinates of mobile devices. The basis of cell phone communication is the exchange of data between subscribers and the network operator. Data exchange is essential to determining the location of subscribers.

There are numerous positioning systems, including device-based, network-based, and GPS-based. The majority, however, do not provide feedback to the operator, and mobile operators lack access to the GPS coordinates of mobile devices. Instead, each subscribers' activity is linked to the mobile network using the coordinates of the nearest connected telecommunications cells.

Radio waves are used to locate phones, and positioning is accomplished using a variety of techniques, including cell ID, triangulation with heading angle, and distance from the antenna. Due to distinct network standards (GSM, CDMA, 3G) and different location purposes, different geopositioning methodologies are

utilized. Moreover, the use of mobile positioning data in geographical research involves various approaches and algorithms.

Two types of mobile positioning can be distinguished, active and passive. Active mobile positioning is utilized for mobile tracking, in which the location of a mobile phone is determined (updated) in response to a specific request using radio waves (Ahas et al., 2007). Passive mobile positioning is data that is routinely retained in the CDR of mobile operators (billing memory, inter-cell transition, home location register, etc.) (Ahas et al., 2009). The simplest method of passive mobile positioning is a "billing log" that captures actions that have been triggered. Call activity encompasses all mobile phone usage (incoming and outgoing conversations, SMS messages, GPRS, etc.). Typically, passive mobile positioning data is collected with grid cell precision. When discussing passive mobile positioning data, it is essential to define some of the key terms associated with this topic. The cellular network is supported by a collection of base stations, which typically consist of a single tower and multiple directional antennas. One antenna's radio coverage constitutes a network cell; multiple antennas comprise a cellular network. Each network cell in a mobile network has a unique identifier and geographic coordinates, allowing the location of each phone within the cell to be readily determined.

Cell ID is one of the most prevalent passive geolocation methods. The method derives its name from the fact that each cell has a specific geographic coverage area and a unique identification code. The location of mobile phone is based on geolocation of telecommunication network signal coverage areas, which are represented as cells. The cell capacity of the network and all cellular networks is not fixed; typically, the phone will transition to the antenna with the strongest radio coverage or the greatest visibility. If the network is congested or visibility is poor, phones can be routed to any station in the vicinity rather than the closest one. In a GSM network, the greatest distance between the device and the antenna cannot surpass 35 kilometers. In areas with fewer inhabitants and greater distances, GSM networks employ amplified antennas. Cell phones are located anywhere within the cells. The latter are symbolized by spherical shapes. Their radius extends from 100 meters in densely populated areas to several kilometers in sparsely populated areas. Each base station is outfitted with multiple antennas that emit multiple cells in various directions. Mobile network signals encompass multiple cells that overlap. The angle of the signal emitted by distinct antennas connected to each base station divides the mobile network. These units are known as mobile network sectors. The default configuration is characterized as three-sector base stations with sectors encompassing approximately 120° (Ratti et al., 2006). Consequently, each entry corresponds to a sector position. Mobile network areas are traditionally depicted as Voronoi polygons centered on base stations.

LBS, a location-based service, is another common method. Several methods, including cell ID, timing advance, and signal intensity, are needed to achieve optimal positioning performance (Resch, B. et al., 2005). The concept underlying the Cell ID method has been discussed previously. Originally deemed a technical parameter of the mobile station, timing advance can be used to estimate the distance between the serving station and the subscriber owing to its specification. Regardless, it is only able to estimate the segment of the ring from the service area, and the positioning is not accurate. The signal intensity is estimated using intricate models that take into account geographical area, number of neighboring cells, etc. Depending on the density of cell sites, the LBS method's accuracy ranges from a few hundred meters to a thousand. Therefore, we have the highest accuracy in cities and it decreases significantly in less populous areas (Andreas Schmidt-Dannert, 2010). In the Figure 3.2 the basic principle of LBS method is presented.

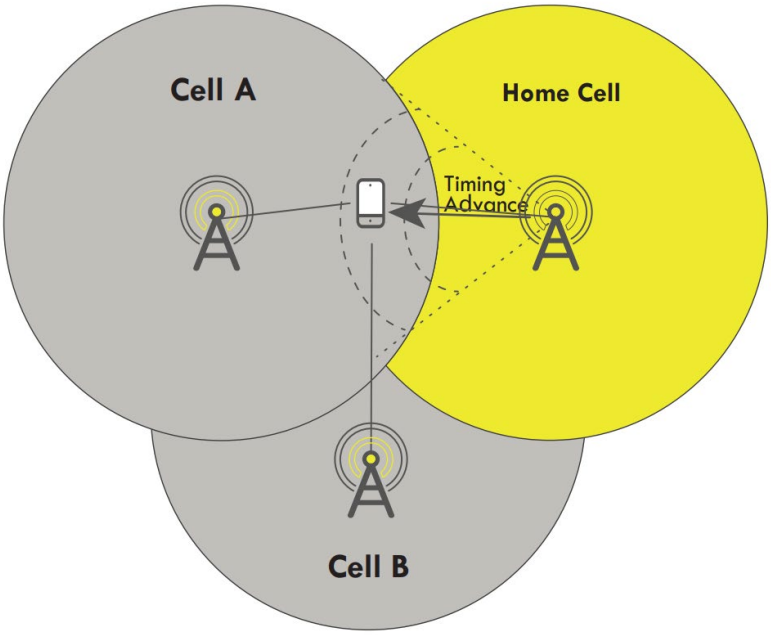


Fig. 3.2 – LBS geopoistioning method

Source: (complied by author)

Mobile operators may capture anonymized geographic data from CDR, such as location points or motion vectors, for use in scientific surveys. Privacy and surveillance concerns are essential components of mobile positioning data. For passive positioning, sources other than call activity, such as antenna erlang, are also utilized (Reades et al., 2007).

3.1.3. OD matrices

Origin destination matrices are key aspect of transport planning when it comes to representing of travel demand. The OD matrices contain condensed information about travel demand at the study location. Each element of the matrix quantifies the number of movements between origin and destination regions. Movements can refer to the aggregate travel demand of all travelers in a given time period, or to more disaggregated levels, where movements characterize travel demand for a particular mode of transport, purpose of travel, etc. For travel surveys that guarantee a statistically representative number of trips at the level of OD pair. Consequently, and because of the difficulty of observing all movements in the study area, OD matrices are usually generated using a travel demand model. The OD matrices are a consequence of the travel distribution and mode selection steps in traditional travel demand modeling using a four-step approach. The mode-specific OD matrices are then used at the traffic assignment stage. To calibrate the parameters of the travel time distribution model, it is common practice to compare the simulated network distribution of travel time with the observed distribution of travel time derived from household surveys. Once assigned, the simulated traffic volumes can be compared to the observed traffic count volumes.

Sometimes, once assigned, correction or matrix update techniques are applied to the base matrix using the traffic count values to create an updated travel demand matrix.

There are several classes of correspondence matrices. The first class includes normative, most often linear models, showing increased sensitivity to individual indicators from others. Models of the second class are statistical – from unusual univariate to multivariate components. A special place among them is occupied by mobile models. To solve problems of the thirds class, statistical models and basic genetic models are also used, but modified and more complex compared to the models of the second class. The complication of game models occurs in the form of redundant conditions corresponding to the balancing of the correspondence matrix. Models of the fourth class are entropy. They are used in the form of a nonlinear optimization of a mathematical programming problem, and their objective function is of a thermodynamic nature and includes probabilistic characteristics of behavior. The decisive role is played by non-deterministic factors of individual behavior, the propensity for collective behavior.

Traditionally in mobility planning OD matrices are used in form of simple two-dimensional arrays that represent the summarized value of trips between OD pairs. As noted earlier, current travel demand estimation practice typically derives OD matrices from large-scale travel survey data, while more recent developments in this area use various new data sources, such as data from mobile operators, to reduce data collection costs and provide more timely estimates.

Traditional sources of data used in developing origin-destination (OD) matrices in mobility studies include household travel surveys, passenger boarding information systems, and travel documents (Stopher et al., 2007). Travel surveys are valuable for formalizing and estimating behavioral choice models, such as destination or mode of transportation choices. However, there are some issues with using this data source for developing OD matrices for demand modeling, including high costs, rising non-response rates, omission of significant trips, and data accuracy concerns (Stopher et al., 2007). Another commonly used data source is automatic fare collection systems (AFC). In certain systems, AFC only records passenger boarding information, not alighting information. For instance, in London, Oyster card data was utilized, and rules were proposed to identify transfer transactions accurately. Passenger trip flow was scaled using additional information, such as count data (Gordon et al., 2018). Recent studies have attempted to combine both travel survey and AFC data to achieve more accurate trip counting and passenger flow estimation. Tamblay et al. (2016) proposed a methodology to infer a zonal OD matrix from stop-to-stop data obtained from AFC, requiring land use information and a zoning system. The model was calibrated using a passenger access survey, and the results were externally validated with a large-scale OD survey. In the similar study authors conducted obtaining OD matrices from AFC data for the city of Porto using trip-chaining method (TCM) (Hora et al., 2017). In the Figure 3.3. Visualization of the OD matrix obtained from the application of the algorithm to the case study of Porto is shown. the However, with advancements in technology and data collection methods, attention has shifted towards new sources of information, such as mobile operators' data, due to its non-intrusive collection and large sample sizes.

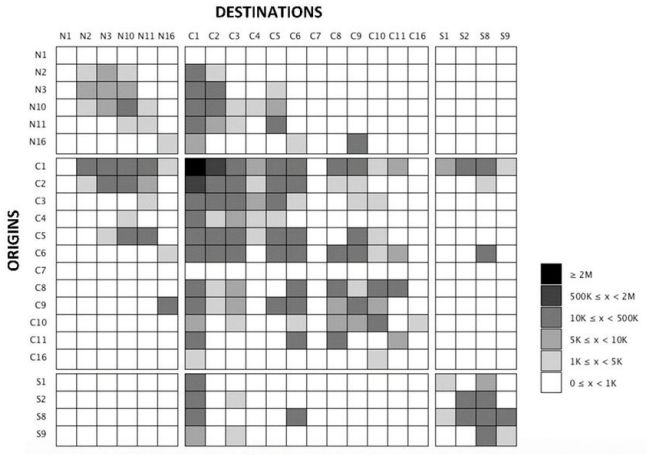


Fig. 3.3 – Visualization of the OD matrix obtained from the application of the algorithm to the case study of Porto

Source: (Hora et al., 2017)

The first study that implemented use of mobile operators' data for conducting OD matrices was done in Italy in 2000 (Bolla and Davoli, 2000). Caceres et al. (2007) calculated an OD matrix for a road between Huelva and Seville in Spain, comparing the results with road traffic counts based on motorway interchanges. Wang et al. (2012) conducted a major study in San Francisco and Boston, constructing hour-by-hour OD matrices to observe network saturation during morning peak periods, considering journeys taking less than one hour. They segmented the population into three groups based on data collection to ensure unbiased results. In another study published in 2016 authors compared OD matrices obtained from mobile operator Telefonica Chile in Santiago with data from the Santiago travel survey, using an iterative end-point fit algorithm to identify activity stops and found high correlations between the two data sources in terms of trip distribution and OD matrix (Graells-Garrido and Saez-Trumper, 2016). These examples highlight various methodologies and data sources utilized for addressing issues related to building OD matrices for transport planning purposes and the growing interest in leveraging mobile phone data for these studies.

3.2. Data pre-processing

Data pre-processing plays a crucial role in mobile data analysis as it involves cleaning, transforming, and preparing raw mobile data to make it suitable for further analysis and modeling (Han et al., 2011). By ensuring the data is accurate, consistent, and relevant, pre-processing helps in extracting meaningful insights and improving the performance of machine learning algorithms and statistical models. The first step of pre-processing is noise reduction. The second step of data pre-processing is trajectory segmentation.

3.2.1. Data noise

Mobile data can sometimes be noisy, containing outliers or irrelevant information. Removing or handling these noisy data points is essential to prevent them from influencing the analysis negatively (Lyons, 2011). Noise reduction focuses on identifying and rectifying errors, inconsistencies, and missing values in the

mobile data. Missing values may occur when certain mobile devices fail to report specific data points. Noise reduction is a critical step in mobile data analysis to improve data quality by removing irrelevant or erroneous data points that can adversely affect the accuracy of analytical models and insights. In the context of mobile data, noise can arise due to various factors such as sensor errors, network fluctuations, device-specific inconsistencies, or interference during data transmission. Addressing noise is essential to ensure robust and reliable analysis. The two fundamental steps of noise reduction are oscillation filtering and trajectory segmentation.

The most common type of noise in terms of mobile data are oscillations. Oscillations refer to repetitive and fluctuating patterns observed in trajectory data or other time-series data collected from mobile devices. These oscillations can be caused by various factors, such as measurement errors, GPS inaccuracies, or the nature of the user's movements. There are several types of occurring oscillations:

- **Trajectory Oscillations:** trajectory data from GPS traces or location-based services may exhibit oscillatory patterns, where a user appears to move back and forth between two locations or follows repetitive paths. These oscillations may arise due to GPS multipath effects, signal interference, or the user's movement within constrained spaces (e.g., indoor environments) (Chen et al., 2018);
- **Signal Noise:** mobile data collected from various sensors (e.g., accelerometers, gyroscopes) can contain noisy oscillations due to sensor inaccuracies or environmental disturbances. These oscillations can affect the reliability of movement detection and activity recognition algorithms (Li et al., 2021);
- **Cyclic Behavior:** oscillatory patterns may arise in due to the cyclic nature of certain activities or routines, such as daily commuting or weekly travel patterns. Identifying and understanding such cycles is essential for modeling user behavior and predicting future movements (Huang et al., 2013);
- **Time-Based Oscillations:** time-series data, such as mobile app usage patterns or call records, may exhibit oscillations in usage or activity levels at specific times of the day or week (Bhattacharya et al., 2015).

Trajectory oscillations are by far the most common type of oscillations that is presented in the literature as they appear on the level of data mining and geopositioning due to the mechanisms of these processes that were already described in the first part of this chapter and overall nature of telecommunication network. In the next section we would like to present two steps of noise reduction, oscillation filtering with the focus on trajectory oscillations and positioning smoothing with focus on increasing geolocation precision of raw subscriber locations collected in CDR.

3.2.2. Noise Reduction

Oscillations filtering in mobile data analysis involves the process of removing periodic fluctuations or oscillations from the dataset to reveal underlying trends and patterns more accurately. This filtering technique is crucial in understanding mobile user behavior, traffic patterns, and mobility dynamics, as it helps eliminate noise and biases introduced by recurring temporal patterns. Several methods are employed for oscillations filtering in mobile data analysis, each suited for different types of data and research

objectives. Some commonly used techniques include Fourier Transform, Seasonal Decomposition and Wavelet Transform (Zheng et al., 2015; Calabrese et al., 2013; Gao et al., 2014).

To describe simple methods of detecting and removing of trajectory oscillations we need to understand a mechanism behind their occurrence. Let's imagine we have an active mobile phone connecting with the mobile network. If it's covered by the reach of two antennas "A" and "B" it may create an oscillation called "ping-pong" effect. Due to network balancing mechanism connection may switch back and forth between two antennas and static phone may appear as changing its location. In that case signal switching between two antennas create so-called oscillation pair. It takes at least three 3 transitions of signal between antennas to identify it as an oscillation (Bayir et al., 2009). One simple way to remove an oscillation would be delete the minority location. In case if we have an oscillation pattern "ABA" that means that we would remove location "B" and retain location "A". That method won't work with more complex oscillation patterns. If we have an oscillation pattern "ABC" the one way to address that could be to use measure distance between locations. If for example two locations "A" and "C" are positioned relatively close but on other hand locations "A" and "B" are positioned far away from each other then connection "AB" is considered as an impossible jump and only location "A" and "C" are considered while location "B" is removed. For a complex oscillation pattern as "ABAB" the method of distance threshold could be used (Çolak et al., 2015). There are other methods for even more complex patterns (e.g., "AAABCAA" or "ABCA") such as using stay periods, time-window-based filtering method etc. (F. Wang et Chen, 2018). In one of the study's authors used combination of mentioned methods to detect oscillations and identified that almost 6% of CDR from their 1 Tb sample were oscillations (Wu et al., 2014).

Even after applying oscillation filtering to raw CDR arrays a lot of times geopositioning of records is still very inaccurate from real-world subscribers' locations. This is especially present in less populated areas where the network coverage is not so dense. As these devices move, their position readings can be subject to various sources of errors, such as GPS signal fluctuations, multipath interference, and other environmental factors. To address this problem researchers have developed so-called positioning smoothing method. Position smoothing algorithms aim to enhance the reliability of the collected data by filtering out irregularities and providing more stable and consistent location estimates.

One of the widely used position smoothing methods is the Kalman filter. The Kalman filter is an optimal recursive algorithm that estimates the state of a dynamic system from a series of noisy measurements. In the context of mobile data analysis, it helps improve the accuracy of location data by combining the information from various sensors and reducing the impact of outliers or noisy measurements. A study published in 2017 explored the application of Kalman filtering to reduce GPS positioning noise in smartphone-based travel surveys (Kim et al., 2017). The researchers demonstrated that using the Kalman filter significantly improved the accuracy of location data, making it a valuable tool for analyzing and interpreting travel behavior patterns based on mobile data.

Another approach to position smoothing is the use of Gaussian processes. Gaussian processes are a powerful non-parametric Bayesian method for estimating continuous functions from noisy observations. In the context of mobile data analysis, they can be used to model the relationship between time and position and predict more accurate and smooth trajectories for mobile devices. Research done in 2018 applied Gaussian processes to smooth noisy mobile phone positioning data (Srinivasan et al., 2018.). The research demonstrated that this technique outperformed conventional filtering methods and resulted in more accurate and smooth trajectories

Apart from the Kalman filter and Gaussian processes, there are other position smoothing algorithms and techniques used, such as the moving average filter, Savitzky-Golay filter, and particle filters, each with its own advantages and limitations depending on the specific application and data characteristics.

3.2.3. Trajectory segmentation

Second step of data-preprocessing is to imply a trajectory segmentation. This process refers to dividing continuous movement data, such as GPS traces or CDR, into meaningful segments or sub-trajectories. These segments represent distinct mobility patterns, activities, or events, which are essential for various applications. The trajectory segmentation process is crucial because raw data is often noisy and continuous, making it challenging to extract meaningful insights directly. By breaking trajectories into segments, researchers can identify stops, turns, mode of transportation, and other significant mobility patterns, allowing for a deeper understanding of users' behavior and preferences. In case of CDR the triangulation method is usually the most used.

In the study published in 2012, authors proposed a trajectory segmentation method for mobile phone users to analyze urban mobility patterns (Gao et al., 2012). They used anonymized data collected from mobile phones to track users' movements. The proposed method employed a combination of spatial and temporal features to detect mobility patterns, such as stops, turns, and significant changes in transportation modes. The authors applied density-based clustering to identify stay points, representing areas where users spent a considerable amount of time. They then used a time-based segmentation approach to divide the trajectories into meaningful segments, allowing them to analyze urban mobility patterns more effectively.

Other study published in 2014 utilized smart card data from the London Underground to understand travel behavior variability among users (Sun et al., 2014). The data consisted of entry and exit records of passengers using public transportation. To segment the trajectories, the authors applied clustering techniques to group similar travel patterns together. They used a density-based clustering algorithm to detect stay points, representing passengers' activity locations within the transit system. The study further employed the K-means algorithm to cluster the trajectories based on spatial and temporal characteristics, helping to identify different travel behavior patterns. By segmenting the trajectories, the researchers gained insights into passenger preferences, travel modes, and transit usage patterns within the urban environment.

3.3. Data processing

3.3.1. Itinerary Reconstruction

One key application of mobile data analysis is itinerary reconstruction, a process that aims to accurately reconstruct individuals' travel patterns and sequences based on their mobile device's location data and other contextual information. Various techniques have been developed to address the challenges posed by noisy and sparse data, irregular sampling intervals, and the need to infer users' travel sequences including trajectory segmentation, which was already observed in the second part of this chapter:

- **Spatial clustering** is employed to group similar GPS locations together, forming clusters that represent specific locations or points of interest. By clustering GPS points that are in proximity, this technique helps in identifying common destinations visited by an individual;

- **Hidden Markov Models** (HMMs) are probabilistic models used to infer hidden states based on observable data. In the context of itinerary reconstruction, HMMs are employed to model users' mobility patterns as a sequence of hidden states (e.g., different transportation modes or locations) given their observed GPS locations;
- **Probabilistic inference** techniques, such as Bayesian networks and Conditional Random Fields (CRFs), are used to model the dependencies between different location points and infer the most likely itinerary based on observed data. These techniques can take into account not only spatial proximity but also temporal patterns and user behavior, improving the accuracy of itinerary reconstruction;
- **Machine Learning Algorithms,** which can be supervised and unsupervised, play a significant role in itinerary reconstruction. For instance, Support SevVector Machines (SVMs), Random Forests, and Gradient Boosting are used for classification tasks, where the goal is to categorize locations into different trip types (e.g., home, work, leisure). On the other hand, unsupervised techniques like clustering and dimensionality reduction algorithms can be applied to discover patterns and similarities in users' travel behavior;
- **Data Fusion and Contextual Analysis** involves integrating mobile data with other contextual information, such as timestamps, device sensor data (e.g., accelerometer, gyroscope), weather conditions, and user profiles, to enhance the accuracy of itinerary reconstruction. By fusing multiple data sources, the reconstructed itineraries become more comprehensive and reflective of real-world travel patterns.

Several studies have used different techniques of Itinerary Reconstruction to achieve deeper knowledge on travel patterns of studied groups. For example, in the work published in 2014 researchers proposed a method to infer users' transportation modes and travel purposes from mobile phone data, such as GPS trajectories and call records. By combining trajectory segmentation techniques with supervised machine learning algorithms, they were able to reconstruct users' travel itineraries accurately. The study demonstrated the feasibility of inferring transportation modes (e.g., walking, driving, public transport) and travel purposes (e.g., work, leisure) from mobile data (Zhu et. al., 2014). In the other paper authors used CDR to analyze urban mobility and reconstruct travel itineraries (Silva et. al., 2017). The study involved applying data mining and trajectory analysis techniques to CDR, which contain information about call times, cell tower locations, and call durations. By processing and clustering this data, they reconstructed individuals' travel patterns and identified key locations in the city. At last, the most recent study focused on itinerary reconstruction using location-based social networks (LBSN) data, specifically from platforms where users share their locations and activities (Alharbi et. al., 2019). The researchers employed a data-driven approach based on trajectory clustering and Hidden Markov Models (HMMs) to infer users' travel sequences and identify their frequent destinations.

3.3.2. Transport mode detection

Transport mode detection using CDR appears to be quite a new area of research. Few studies were conducted in that field. Previous studies on transport mode detection mainly focused on positioning obtained from GPS technology. One common approach is to use machine learning algorithms to process the sensor data and classify the transport mode. Zhang et al. (2017) conducted a study where they used smartphone sensor data, including GPS and accelerometer data, to identify four transport modes: walking, running, cycling, and

motorized transport. They employed support vector machines (SVM) and achieved high accuracy in distinguishing between different modes. Another research work by Bulling, A. et al. (2014) focused on detecting mobility activities using a smartphone's built-in sensors, including accelerometers and gyroscopes. Their study aimed to identify not only different transport modes but also other activities like walking upstairs, walking downstairs, and standing still. The researchers used a combination of decision trees and random forests algorithms.

But compared to GPS data CDR positioning is not so precise. Yet scientists still have a lot of interest in that area for the reason of how big samples of data can be obtained and its relatively low cost. But because of the unreliable positioning CDR have compared to the GPS researches are required to develop new methods to combat that issue. One of the early studies in this domain was conducted by Toole, J. L. et al. (2015), where they proposed a method to classify transport modes using CDR. The study focused on distinguishing between four main modes: walking, cycling, driving, and public transport. By analyzing the temporal and spatial patterns of mobile phone activity, they achieved reasonable accuracy in classifying the modes of transportation. In a similar study, De Nadai, M. et al. (2016) investigated the use of CDR to understand urban mobility patterns in Italy. The researchers applied machine learning techniques to CDR, including call and text message records, to infer the mobility behavior of individuals, such as their home and work locations, and the modes of transport they commonly used. Apart from individual-level analysis, CDR has been used to study collective mobility patterns and transport mode distributions in cities. For example, Blat, J. et al. (2016) examined mobility patterns in Barcelona using anonymized CDR. They analyzed the spatial distribution of mobile phone events to estimate the flow of people and the modes of transport they utilized, helping to understand the overall mobility landscape of the city.

3.3.3. Location patterns recognition

Location patterns recognition refers to the process of extracting meaningful insights and patterns from location data generated by a variety of data sources used in transport planning and mobility research. These data sources usually include GPS traces, CDR, and other location-related information that can be used to study the movement and behavior of individuals or groups of users. Most common location patterns include “home”, “work” and “other” (Alexander et al., 2015). Information about location patterns can help in understanding of behaviour of certain social groups.

In a study conducted in 2013 in Portugal, authors identified the “home” and “work” locations of one hundred thousand users and applied clustering methods to their most frequent locations (Chaji et al., 2013). Clustering was based on characteristics such as aggregated hourly call volume by weekday. Then the results of the analysis were compared with data from national census and it was concluded that at least in three of the clusters average call pattern corresponded with data from Portuguese national institution called Instituto Nacional de Estatística (INE) regarding the average hours that people spent at home and at work.

Another paper published in 2015 focused on identifying “home” locations in Boston as the most frequent weekend and weekday destination (Alexander et al. 2015). On the basis of the number of weekday visits, the location of “work” was determined to be the location where users were the farthest from their homes. Wihalm et al. (2015) conducted a study in both Boston and Vienna by categorizing activities by location. The characteristics derived from call and land use data led to inferences about activities such as "home",

"work", "education", "leisure", and "shopping". Using a Markov relational network, activity patterns were generated, and this method was tested using survey data.

Using mobile data analysis, these studies have contributed significantly to understanding the mobility patterns and behaviors of various populations. The cited research demonstrates the significance of studying and analyzing mobility patterns in order to gain a deeper comprehension of human movement and behavior.

4

DATA ANALYSIS OF TÂMEGA E SOUSA REGION

4.1. Overview of Tâmega e Sousa region

Case study project in the framework of master's dissertation was done within the TRENMO Engenharia S.A. company from December 2022 to June 2023 and its main focus was on analyzing mobility patterns in the region of Tâmega e Sousa in Portugal using the Call Data Records provided by the telecommunication company Vodafone.

The intermunicipal community of Tâmega e Sousa (shown in the Figure 4.1) was created in 2009 from the former districts of Aveiro, Braga, Porto and Viseu. It is located in the north of Portugal and has an area of 1,831.52 km². The seat of the intermunicipal community is Penafiel. According to the national census conducted in 2021 by Portuguese national statistical bureau Instituto Nacional de Estatística (INE) its' population was 408,878 (INE, 2021), the population distribution is presented in the Table 4.1. Since 2011 the region has lost around 5,6% of its population (24,037 inhabitants). The municipality of Lousada is the only one in the area that haven't lost its' population in the 10 years since the previous census.

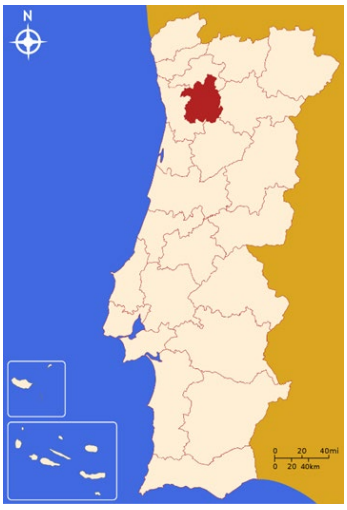


Fig. 4.1 – Intermunicipal community of Tâmega e Sousa

Source: (Wikipredia.org)

The intermunicipal community of Tâmega e Sousa consists of 11 municipalities (shown in the Figure 4.2).

Table 4.1 – Municipalities of Tâmega e Sousa region

Source: (INE, 2021)

Municipality	Population	% Of total population	Area (km²)
Resende	10 053	2,46%	123 35
Castelo de Paiva	15 597	3,81%	115 01
Celorico de Basto	17 666	4,32%	181 07
Cinfães	17 747	4,34%	239 29
Baião	17 527	4,29%	174 53
Lousada	47 401	11,59%	96 08
Marco de Canaveses	49 563	12,12%	201 89
Amarante	52 131	12,75%	301 33
Paços de Ferreira	55 623	13,60%	70 99
Felgueiras	55 883	13,67%	115 74
Penafiel	69 687	17,04%	212 24
Total	408 878	100%	1 831 52

Fig. 4.2 – Municipalities of Tâmega e Sousa region

Source: (Wikipedia.org)

Current analysis in the framework of case study mainly devoted to the movements that happens within the region of Tâmega e Sousa. Municipalities (Concelho) and parishes (Freguesia) that make up the region are considered as Core zone areas and the ones that located outside of the Tâmega e Sousa regions are considered as Periphery zone. The main focus of this analysis is on the movements within Core zone. Zoning was defined by TRENMO in order to simplify the process of analysis. Remote regions of periphery zone were merged into bigger areas while some municipalities of a closer periphery and core zones were restructured on the level of parishes like for example in the municipality of Paredes in periphery zone the parishes of Lordelo, Rebordosa and Gandra were detached in separate regions. Specific Zoning of the Core and Periphery zones is presented on the Figure 4.3.

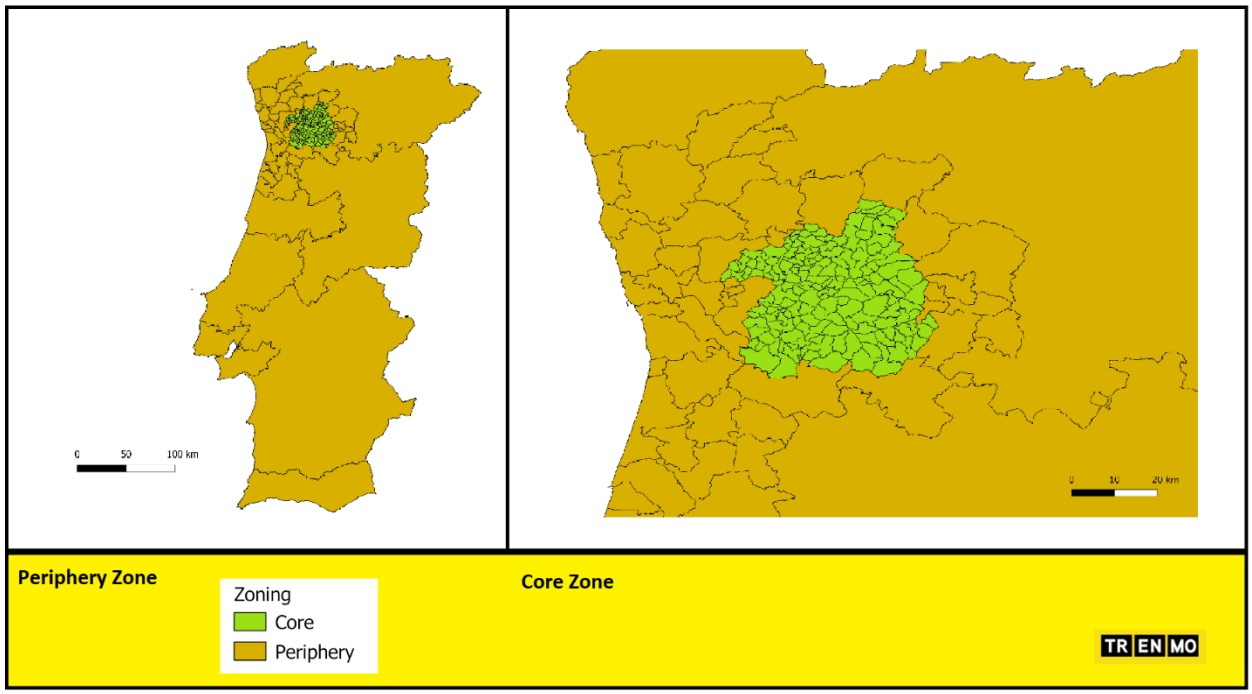


Fig. 4.3 – TRENMO zoning

Source: (complied by author)

4.2. Data extraction and pre-procession

The information about the movements of subscribers was provided by the mobile operator company Vodafone in form of CSV files (Comma Separated Values, shown in the Figure 4.5). When collecting the Data Vodafone use Cell ID geopositioning method and Voronoi polygons. Voronoi polygons are created from the location of Vodafone Portugal's antennas, with which the geographical polygons under analysis (targets) intersect.

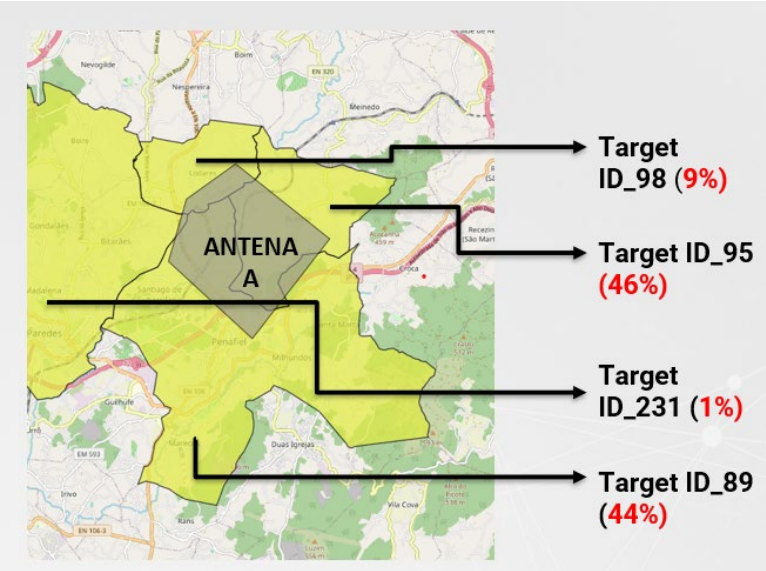


Fig. 4.4 – Vodafone data mining

Source: (Vodafone CDR presentation, 2022)

In Figure 4.4, the Voronoi of antenna A, intercept 4 target polygons with intersection ratios 1%, 9%, 44% and 46%. Users registered in antenna A in a given time period will be distributed among the respective target polygons according to these ratios.

The data is provided in the form of simple matrixes with 8 columns and big number of rows, each row represents one movement of the subscriber that happened in the set period of time. The Data is provided for three days: 16th (Wednesday), 18th (Friday) and 26th (Saturday) of November year 2022.

	A	B	C	D	E	F	G	H
1	start_interval	end_interval	poi_origem	poi_destino	poi_home	poi_work	country_name	count_ine_imsi
2	2022-11-26.20:00:00	2022-11-26.20:00:00	0	0	33	0	PORTUGAL	109.40301350219761
3	2022-11-26.20:00:00	2022-11-26.20:00:00	0	0	58	0	PORTUGAL	54.668358922966036
4	2022-11-26.20:00:00	2022-11-26.20:00:00	0	0	71	0	PORTUGAL	75.65447428765458
5	2022-11-26.20:00:00	2022-11-26.20:00:00	0	0	189	0	PORTUGAL	9207.146822870622
6	2022-11-26.20:00:00	2022-11-26.20:00:00	0	0	207	0	PORTUGAL	3297.7267670260926
7	2022-11-26.20:00:00	2022-11-26.20:00:00	89	1	1	37	PORTUGAL	8.115265301808097E-7
8	2022-11-26.20:00:00	2022-11-26.20:00:00	178	1	1	201	PORTUGAL	0.20537478512594382
9	2022-11-26.20:00:00	2022-11-26.20:00:00	38	1	1	66	PORTUGAL	0.0024909509971552312
10	2022-11-26.20:00:00	2022-11-26.20:00:00	71	1	1	16	PORTUGAL	5.427970959993974E-6
11	2022-11-26.20:00:00	2022-11-26.20:00:00	178	1	1	214	PORTUGAL	0.366136918550334

Fig. 4.5 – Raw CSV file

Source: (complied by author)

Each column accounts for a separate information about the movement of the subscriber:

- **Start_interval** – time and date of the beginning of movement;
- **End_interval** – time and date of the end of the movement;

- **Poi_origem** – original location of the movement;
- **Poi_destino** – ending location of the movement;
- **Poi_home** – location which is considered by the operator as the home location of the subscriber;
- **Poi_work** – location which is considered by the operator as working location of subscriber;
- **Country_name** – country of origin of subscriber;
- **Count_ine_imsi** – this variable represents the number of trips made in that period, but extrapolated to the total number of people who have an active SIM card.

It is important to make a few notes about the data that is provided by Vodafone and its' form:

i. **Start_interval** and **End_interval** have the same value on the provided example because the matrixes are split in the different days and time intervals, so each separate matrix represents all the movements that were made in or out of the studying zone in the fixed time interval. In our work we used the data from three different days and each day is split in six time intervals, intervals were ordered by TRENMO in order to simplify analysis of particular parts of the day, for example intervals for morning (7 a.m. to 10 a.m.) and evening (5 p.m. to 10 p.m.) rush hours are set at three hours while other intervals like night interval (midnight to 6 a.m.) may have different amount of recorded hours in it.

ii. **Poi_home** is the location which is operator considers as the "home" location of the subscriber according to its' algorithms, usually it is the location where the subscriber resides most of its´ time during the night without any "movements";

iii. **Poi_work** is the location which is operator considers as the "work" location of the subscriber according to its' algorithms, usually it is the location where the subscriber resides most of the time during the usual working hours which can be place of work, place of studying, etc.;

iv. **Country_name** is the country where the SIM card of the subscriber is registered. In the matrixes provided by the Vodafone this parameter has two values: Portugal and Outros.

v. **Count_ine_imsi –** this variable considers an active mobile phone subscribers of all of Portugal It is extrapolated to the total number of active subscribers in the time of receiving a record as well as population numbers from INE.

In order to process data from the files provided by Vodafone we need at first need to extract them from the matrixes. The matrixes present as a simple CSV files and are separated according to the 6 intervals trough out the day. Because of their big size (around 1,5 gigabytes and at least 23 million rows each) it is not possible to extract the values using spreadsheet software programs such as Microsoft Excel so the first stage of work would be to do a preprocessing and extraction of Data using SQL software such as PostgreSQL.

First step would be extraction of the Data from CSV file in the Database (shown in the Figure 4.6).

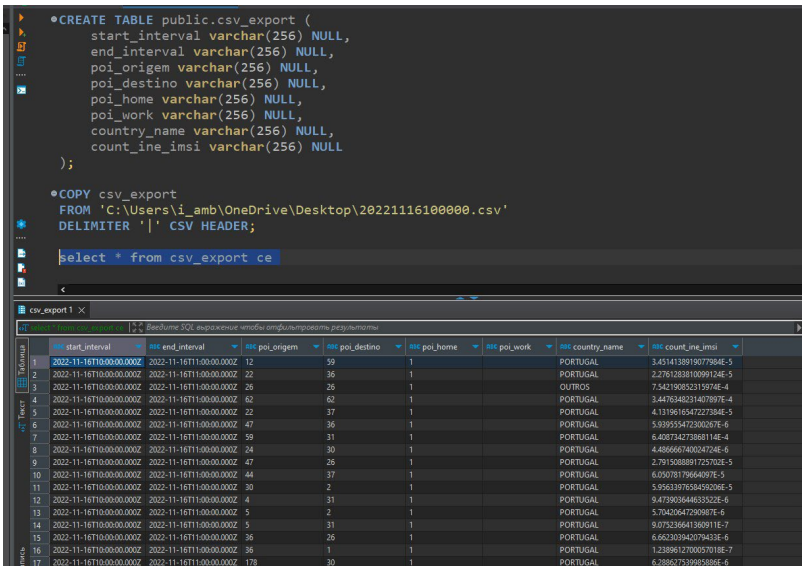


Fig. 4.6 – Data extraction

Source: (complied by author)

At the second step we could organize rows by pair of columns poi_origem and poi_destino and summarize the values of count_ine_imsi column for each existing OD pair (shown in the Figure 4.7).

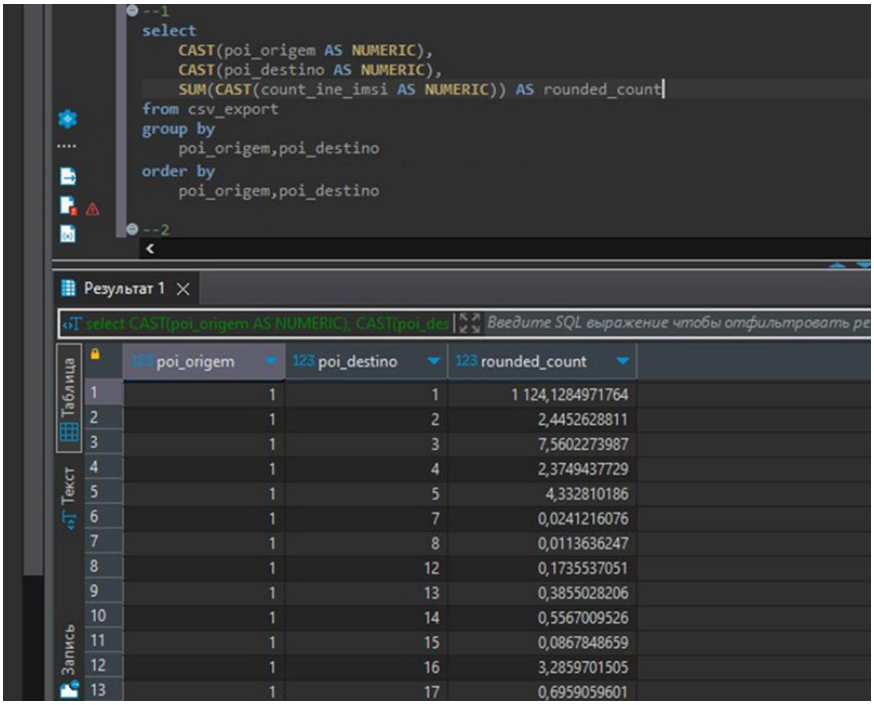


Fig. 4.7 – Organizing of OD pairs

Source: (complied by author)

The final step for Data preprocessing would be to summarize count_ine_imsi values for similar OD values and to eliminate OD pairs that have their poi_origem or poi_destino columns value as 0 as this means that operator was not able to recognize location of these records according to the zoning provided by TRENMO and it wouldn't be possible to use these records in analysis due to inability to correctly obtain their location of origin or destination (shown in the Figure 4.8). Then we can finally export processed Data to Microsoft Excel for further work.

```
COPY (
  select
    CAST(poi_origem AS NUMERIC),
    CAST(poi_destino AS NUMERIC),
    SUM(ROUND(CAST(count_ine_imsi AS NUMERIC), 1)) AS rounded_count
from csv_export
group by
    poi_origem,poi_destino
order by
    poi_origem,poi_destino
) TO 'C:\Users\i_amb\OneDrive\Desktop\testoutput5.csv' DELIMITER '|' CSV HEADER;
```

Fig. 4.8 – Data export

Source: (complied by author)

In the end we will receive CSV file with Data organized as OD pairs with final count_ine_imsi value for each existing pair (which are shown in the Figure 4.9) that already can be open in Microsoft Excel.

	A	B	C
1	poi_origer	poi_destin	count_ine_imsi
2	1	1	1290,282248
3	1	2	3,155278586
4	1	3	3,240113593
5	1	4	3,580281161
6	1	5	9,373517434
7	1	6	0,017783623
8	1	7	0,00234567
9	1	9	0,00962513
10	1	12	0,703174696
11	1	13	0,303094626
12	1	14	0,159214187
13	1	15	0,311250521
14	1	16	1,453492612
15	1	17	0,134398686
16	1	18	0,002445581
17	1	19	0,353047979
18	1	20	2,436195845
19	1	22	11,34022289
20	1	23	16,15608649

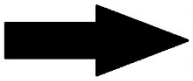
Fig. 4.9 – OD pairs obtained from data provided by mobile operator

Source: (complied by author)

Finally, we should organize received pairs according to the Vodafone location code so we will have pairs with actual locations with their core-periphery zoning and not just numbers (shown in the Figure 4.10). It is worth to mention that this analysis focuses on the core zone movements on the levels of municipalities so

all movements that happens within the border of the same municipality would be considered as inner-municipalities movements.

	A	B	C
1	poi_origer	poi_destin	count_ine_imsi
2	1	1	1290,282248
3	1	2	3,155278586
4	1	3	3,240113593
5	1	4	3,580281161
6	1	5	9,373517434
7	1	6	0,017783623
8	1	7	0,00234567
9	1	9	0,00962513
10	1	12	0,703174696
11	1	13	0,303094626
12	1	14	0,159214187
13	1	15	0,311250521
14	1	16	1,453492612
15	1	17	0,134398686
16	1	18	0,002445581
17	1	19	0,353047979
18	1	20	2,436195845
19	1	22	11,34022289
20	1	23	16,15608649



	A	B	C	D	E	F	G
1	poi_origer	poi_destin	count_ine_imsi	origem	destino	O_Type	D_Type
2	1	1	1290,282248	Cinfães	Cinfães	core	core
3	1	2	3,155278586	Cinfães	Cinfães	core	core
4	1	3	3,240113593	Cinfães	Cinfães	core	core
5	1	4	3,580281161	Cinfães	Cinfães	core	core
6	1	5	9,373517434	Cinfães	Castelo de Paiva	core	core
7	1	6	0,017783623	Cinfães	Felgueiras	core	core
8	1	7	0,00234567	Cinfães	Lousada	core	core
9	1	9	0,00962513	Cinfães	Celorico de Basto	core	core
10	1	12	0,703174696	Cinfães	Penafiel	core	core
11	1	13	0,303094626	Cinfães	Penafiel	core	core
12	1	14	0,159214187	Cinfães	Baião	core	core
13	1	15	0,311250521	Cinfães	Penafiel	core	core
14	1	16	1,453492612	Cinfães	Marco de Canaveses	core	core
15	1	17	0,134398686	Cinfães	Baião	core	core
16	1	18	0,002445581	Cinfães	Penafiel	core	core
17	1	19	0,353047979	Cinfães	Marco de Canaveses	core	core
18	1	20	2,436195845	Cinfães	Marco de Canaveses	core	core
19	1	22	11,34022289	Cinfães	Castelo de Paiva	core	core
20	1	23	16,15608649	Cinfães	Castelo de Paiva	core	core

Fig. 4.10 – Organizing OD pairs by location

Source: (complied by author)

At that moment we have finished process of pre-processing and clearance of row data that was provided by mobile operator. The same process should be done for all the matrixes that are available for chosen days. Then extracted values for each time interval should be averaged between the data from the same time intervals between two days. This would ensure that the final values remain representative and minimize the effect that could be done by any data anomaly that could happened on the one of the days of records.

4.3. Results

4.3.1. Core zone analysis

By carrying out the analysis of the matrixes provided by the Vodafone we now have information about total amount of trips that have been done by the subscribers within the area of Tâmega e Sousa region (core zone). We can look at the morning rush hour average (7 a.m. to 10 a.m.) and evening rush hour (5 p.m. to 8 p.m.) OD trips values that have been extracted from available matrixes and averaged, and their proportional distribution throughout 11 municipalities that compromise the Tâmega e Sousa region. It is important to note that for both of these intervals only the data for two out of three days were considered as the data about the records that were made at 26th of November was excluded by the reason that that day was a Saturday and the values of trips that were made on the weekend day could distort overall assessment of the morning and evening rush hours trips values. The intensity of a colour palate in all following Figures with the total OD trips and percentage values represents the proportion of each value have among the sum of values presented of a Figure. The results for the Morning rush hour Core zone analysis are shown in the Figures 4.11 and 4.12.

		Destination												
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin	Total OD
Origin	Core	46870	14248	11349	11795	13077	40203	35911	40298	39585	50465	6462	310265	
	Amarante	38996	689	27	881	76	2220	1136	2345	99	1136	63	47670	
	Baião	352	10886	21	30	865	42	39	535	9	68	1037	13885	
	Castelo de Paiva	14	8	8270	1	696	6	25	635	7	808	0	10470	
	Celorico de Basto	591	76	2	10337	5	257	39	58	13	42	2	11423	
	Cinfães	35	579	549	1	9166	7	13	1144	1	269	167	11932	
	Felgueiras	2803	218	5	370	5	35200	1693	586	109	868	2	41859	
	Lousada	1085	47	54	41	31	1561	28266	492	1628	2969	6	36180	
	Marco de Canaveses	1845	794	878	74	1496	303	473	31408	71	2378	91	39812	
	Paços de Ferreira	71	7	36	10	11	113	1738	61	37153	418	1	39619	
	Penafiel	1033	111	1498	46	528	470	2400	2937	404	41311	16	50755	
	Resende	23	833	1	1	191	5	8	74	3	12	5076	6228	
	Periphery	3697	1189	1921	3316	1316	5028	6007	2901	6187	10031	936	42530	
	Total Destination	50547	15437	13264	15109	14387	45215	41837	43176	45685	60310	7398		352364

Fig. 4.11 – Morning rush hour (7 a.m. to 10 a.m.) average OD trips in the core zone

Source: (complied by author)

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Core	15,11%	4,59%	3,66%	3,80%	4,21%	12,96%	11,57%	12,99%	12,76%	16,27%	2,08%	88,05%
	Amarante	81,81%	1,45%	0,06%	1,85%	0,16%	4,66%	2,38%	4,92%	0,21%	2,38%	0,13%	13,53%
	Baião	2,53%	78,40%	0,15%	0,22%	6,23%	0,30%	0,28%	3,86%	0,07%	0,49%	7,47%	3,94%
	Castelo de Paiva	0,14%	0,08%	78,99%	0,01%	6,65%	0,06%	0,23%	6,07%	0,07%	7,72%	0,00%	2,97%
	Celorico de Basto	5,18%	0,67%	0,02%	90,49%	0,05%	2,25%	0,35%	0,50%	0,11%	0,36%	0,02%	3,24%
	Cinfães	0,30%	4,86%	4,60%	0,00%	76,82%	0,06%	0,11%	9,59%	0,01%	2,26%	1,40%	3,39%
	Felgueiras	6,70%	0,52%	0,01%	0,88%	0,01%	84,09%	4,04%	1,40%	0,26%	2,07%	0,00%	11,88%
	Lousada	3,00%	0,13%	0,15%	0,11%	0,09%	4,31%	78,13%	1,36%	4,50%	8,21%	0,02%	10,27%
	Marco de Canaveses	4,64%	2,00%	2,21%	0,19%	3,76%	0,76%	1,19%	78,89%	0,18%	5,97%	0,23%	11,30%
	Paços de Ferreira	0,18%	0,02%	0,09%	0,03%	0,03%	0,29%	4,39%	0,15%	93,78%	1,05%	0,00%	11,24%
	Penafiel	2,04%	0,22%	2,95%	0,09%	1,04%	0,93%	4,73%	5,79%	0,80%	81,39%	0,03%	14,40%
	Resende	0,37%	13,38%	0,02%	0,02%	3,07%	0,09%	0,13%	1,18%	0,04%	0,20%	81,51%	1,77%
	Periphery	8,69%	2,80%	4,52%	7,80%	3,10%	11,82%	14,12%	6,82%	14,55%	23,59%	2,20%	12,07%
	Total Destination	14,35%	4,38%	3,76%	4,29%	4,08%	12,83%	11,87%	12,25%	12,97%	17,12%	2,10%	100,00%

Fig. 4.12 – Proportional distribution of morning rush hour average OD trips in the core zone by municipalities

Source: (complied by author)

We can see that by big margin most of the trips that happen in the region are inner-municipalities. 82,53% of all trips that happened inside the core zone were done without crossing the borders of municipalities. Municipalities with the biggest number of inner-municipal trips are Amarante, Felgueiras, Lousada, Marco de Canaveses, Paços de Ferreira, Penafiel and Resende which overall matches with population distribution in Tâmega e Sousa region. We can also see that 12,07% of all trips originated from periphery zone (outside of core zone). Later we can proceed with analysis of evening rush hour pattern which results are presented in the Figures 4.13 and 4.14.

		Destination												
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin	Total OD
Origin	Core	48056	14475	11107	11765	12673	41505	37203	40472	40567	50195	6505	314523	
	Amarante	39356	444	19	542	41	2774	1344	1968	126	1072	29	47715	
	Baião	541	11241	12	23	836	61	70	733	11	93	894	14515	
	Castelo de Paiva	37	16	8444	3	494	8	42	610	9	978	3	10644	
	Celorico de Basto	1011	79	2	10724	6	344	34	59	3	26	1	12289	
	Cinfães	57	645	631	5	9432	16	33	1424	8	483	162	12896	
	Felgueiras	2591	161	7	327	6	35580	1653	403	136	645	1	41509	
	Lousada	1088	31	40	48	21	1772	28959	444	1753	2691	1	36848	
	Marco de Canaveses	2175	708	770	50	1280	375	665	32354	97	2760	72	41306	
	Paços de Ferreira	70	10	11	6	9	135	1606	68	37886	403	3	40206	
	Penafiel	1054	88	1166	35	322	423	2757	2304	434	40955	11	49548	
	Resende	66	1052	0	1	224	5	10	93	3	11	5328	6793	
	Periphery	2812	980	1159	3230	794	4923	4847	1928	6339	6784	770	34566	
	Total Destination	50859	15454	12260	14994	13465	46416	42021	42387	46805	56900	7276		348836

Fig. 4.13 – Evening rush hour (5 p.m. to 8 p.m.) average OD trips in the core zone

Source: (complied by author)

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Core	15,28%	4,60%	3,53%	3,74%	4,03%	13,20%	11,83%	12,87%	12,90%	15,96%	2,07%	90,16%
	Amarante	82,48%	0,93%	0,04%	1,14%	0,09%	5,81%	2,82%	4,12%	0,26%	2,25%	0,06%	13,68%
	Baião	3,73%	77,44%	0,08%	0,16%	5,76%	0,42%	0,49%	5,05%	0,08%	0,64%	6,16%	4,16%
	Castelo de Paiva	0,35%	0,15%	79,33%	0,03%	4,64%	0,08%	0,39%	5,73%	0,08%	9,19%	0,03%	3,05%
	Celorico de Basto	8,23%	0,64%	0,01%	87,27%	0,05%	2,80%	0,28%	0,48%	0,03%	0,21%	0,01%	3,52%
	Cinfães	0,44%	5,00%	4,89%	0,04%	73,14%	0,13%	0,26%	11,04%	0,06%	3,74%	1,26%	3,70%
	Felgueiras	6,24%	0,39%	0,02%	0,79%	0,01%	85,72%	3,98%	0,97%	0,33%	1,55%	0,00%	11,90%
	Lousada	2,95%	0,08%	0,11%	0,13%	0,06%	4,81%	78,59%	1,20%	4,76%	7,30%	0,00%	10,56%
	Marco de Canaveses	5,26%	1,71%	1,87%	0,12%	3,10%	0,91%	1,61%	78,33%	0,23%	6,68%	0,17%	11,84%
	Paços de Ferreira	0,17%	0,02%	0,03%	0,01%	0,02%	0,34%	3,99%	0,17%	94,23%	1,00%	0,01%	11,53%
	Penafiel	2,13%	0,18%	2,35%	0,07%	0,65%	0,85%	5,56%	4,65%	0,88%	82,66%	0,02%	14,20%
	Resende	0,97%	15,48%	0,00%	0,02%	3,30%	0,07%	0,15%	1,37%	0,04%	0,16%	78,44%	1,95%
	Periphery	8,14%	2,84%	3,35%	9,34%	2,30%	14,24%	14,02%	5,58%	18,34%	19,63%	2,23%	9,91%
	Total Destination	14,58%	4,43%	3,51%	4,30%	3,86%	13,31%	12,05%	12,15%	13,42%	16,31%	2,09%	100,00%

Fig. 4.14 – Proportional distribution of evening rush hour average OD trips in core zone by municipalities

Source: (complied by author)

Here we can see again that total number of trips have changed insignificantly compared to morning rush hour values. The proportion of inner-municipal trips also remained almost the same as the value that was in morning rush hour pattern (82,75%). The biggest change that can be noticed happened in the proportion of periphery zone originated trips which have decreased compared to morning rush hour by more than 2 percent (9,91%)

4.3.2. Comparison with INE data

In second parts of analysis, we can compare the absolute and proportional average OD values that we have obtained from data extracted from mobile operator with the data from INE. The last national census was done in 2021 so the information about population residence and daily commute locations remains representative. To get required data from INE we need to extract available data on population locations of residence and their daily commutes, organize it by municipalities of study region and summarize values, data from INE doesn't require any additional pre-procession such as noise clearance because of nature of its' obtainment by more "traditional" tools such as phone surveys. After this process is done, it is possible to compare absolute "home-daily place of residence" commute (home to work, school, university, etc.) values from INE for our study region with obtained earlier average morning rush hour OD trips values. In this analysis we will not focus on comparison with evening rush hour values as they do not corelate well with INE commute values which are more relevant for morning rush hour pattern. At first, we should look at the values for INE data which are shown in the Figures 4.15 and 4.16.

		Destination												
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin	Total OD
Origin	Core	21519	5490	5935	6106	5482	28684	22579	19717	26684	27950	2116	172262	
	Amarante	16811	146	32	317	37	1006	498	784	154	570	35	20390	
	Baião	113	4501	20	25	55	47	49	236	54	86	58	5244	
	Castelo de Paiva	29	30	5189	34	391	50	39	146	35	168	5	6116	
	Celorico de Basto	151	17	26	4826	18	98	53	31	42	85	5	5352	
	Cinfães	49	76	111	16	4464	68	43	129	52	101	44	5153	
	Felgueiras	2124	88	65	599	103	25645	1737	261	421	372	40	31455	
	Lousada	536	47	26	68	54	891	16611	214	1270	1004	17	20738	
	Marco de Canaveses	682	260	156	35	133	171	194	16636	143	1002	19	19431	
	Paços de Ferreira	177	60	48	75	57	315	1745	164	24043	486	13	27183	
	Penafiel	791	113	259	102	130	349	1587	1073	460	24032	29	28925	
	Resende	56	152	3	9	40	44	23	43	10	44	1851	2275	
	Periphery	2879	1543	1626	1622	1127	3064	3651	3222	4301	6420	557	30012	
	Total Destination	24398	7033	7561	7728	6609	31748	26230	22939	30985	34370	2673		202274

Fig. 4.15 – INE "home-daily place of residence" values in the core zone

Source: (complied by author)

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Core	12,49%	3,19%	3,45%	3,54%	3,18%	16,65%	13,11%	11,45%	15,49%	16,23%	1,23%	85,16%
	Amarante	82,45%	0,72%	0,16%	1,55%	0,18%	4,93%	2,44%	3,85%	0,76%	2,80%	0,17%	10,08%
	Baião	2,15%	85,83%	0,38%	0,48%	1,05%	0,90%	0,93%	4,50%	1,03%	1,64%	1,11%	2,59%
	Castelo de Paiva	0,47%	0,49%	84,84%	0,56%	6,39%	0,82%	0,64%	2,39%	0,57%	2,75%	0,08%	3,02%
	Celorico de Basto	2,82%	0,32%	0,49%	90,17%	0,34%	1,83%	0,99%	0,58%	0,78%	1,59%	0,09%	2,65%
	Cinfães	0,95%	1,47%	2,15%	0,31%	86,63%	1,32%	0,83%	2,50%	1,01%	1,96%	0,85%	2,55%
	Felgueiras	6,75%	0,28%	0,21%	1,90%	0,33%	81,53%	5,52%	0,83%	1,34%	1,18%	0,13%	15,55%
	Lousada	2,58%	0,23%	0,13%	0,33%	0,26%	4,30%	80,10%	1,03%	6,12%	4,84%	0,08%	10,25%
	Marco de Canaveses	3,51%	1,34%	0,80%	0,18%	0,68%	0,88%	1,00%	85,62%	0,74%	5,16%	0,10%	9,61%
	Paços de Ferreira	0,65%	0,22%	0,18%	0,28%	0,21%	1,16%	6,42%	0,60%	88,45%	1,79%	0,05%	13,44%
	Penafiel	2,73%	0,39%	0,90%	0,35%	0,45%	1,21%	5,49%	3,71%	1,59%	83,08%	0,10%	14,30%
	Resende	2,46%	6,68%	0,13%	0,40%	1,76%	1,93%	1,01%	1,89%	0,44%	1,93%	81,36%	1,12%
	Periphery	9,59%	5,14%	5,42%	5,40%	3,76%	10,21%	12,17%	10,74%	14,33%	21,39%	1,86%	14,84%
	Total Destination	12,06%	3,48%	3,74%	3,82%	3,27%	15,70%	12,97%	11,34%	15,32%	16,99%	1,32%	100,00%

Fig. 4.16 – Proportional distribution of INE "home-daily place of residence" values in the core zone by municipalities

Source: (complied by author)

At the first look the proportional distribution of values from INE between core zone municipalities of origin and destination looks pretty similar with the proportional distribution of OD trips for both observed before patterns. Proportion of inner-municipalities trips is 83,95% which again is pretty similar to both reviewed before patterns. The biggest difference lies in the value of periphery zone originated trips which is 2,76% higher than in morning rush hour pattern and almost 5% higher than in evening rush hour pattern as well as at total size of recorded "trips" which is significantly lower than in both reviewed patterns. Next, we can compare absolute values from INE data with OD trips from morning rush hour (shown in the Figure 4.17).

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Core	45,91%	38,53%	52,29%	51,77%	41,92%	71,35%	62,87%	48,93%	67,41%	55,39%	32,75%	55,52%
	Amarante	43,11%	21,19%	117,61%	35,97%	48,43%	45,31%	43,84%	33,43%	155,77%	50,19%	55,55%	42,77%
	Baião	32,11%	41,35%	93,58%	81,97%	6,36%	111,50%	125,91%	44,08%	584,23%	125,86%	5,60%	37,77%
	Castelo de Paiva	201,78%	379,70%	62,74%	3591,02%	56,19%	803,47%	158,85%	22,99%	498,11%	20,79%	3755,58%	58,41%
	Celorico de Basto	25,54%	22,31%	1097,61%	46,69%	332,03%	38,08%	134,21%	53,75%	328,31%	204,32%	230,50%	46,85%
	Cinfães	138,82%	13,12%	20,21%	2833,55%	48,70%	942,95%	341,71%	11,28%	3502,16%	37,52%	26,29%	43,19%
	Felgueiras	75,77%	40,46%	1386,81%	161,75%	1927,93%	72,85%	102,60%	44,54%	385,65%	42,86%	1991,89%	75,14%
	Lousada	49,40%	99,72%	48,30%	165,43%	173,08%	57,08%	58,77%	43,53%	78,01%	33,82%	279,23%	57,32%
	Marco de Canaveses	36,96%	32,73%	17,77%	47,05%	8,89%	56,39%	41,03%	52,97%	201,14%	42,14%	20,78%	48,81%
	Paços de Ferreira	249,33%	865,71%	133,58%	735,56%	519,63%	277,81%	100,40%	268,50%	64,71%	116,40%	1308,64%	68,61%
	Penafiel	76,54%	101,96%	17,28%	223,06%	24,64%	74,22%	66,12%	36,53%	113,88%	58,17%	181,04%	56,99%
	Resende	245,32%	18,24%	318,90%	928,68%	20,93%	806,16%	276,76%	58,41%	369,72%	352,97%	36,46%	36,53%
	Periphery	77,88%	129,72%	84,63%	48,92%	85,62%	60,94%	60,78%	111,05%	69,51%	64,00%	59,50%	70,57%
	Total Destination	48,27%	45,56%	57,01%	51,15%	45,94%	70,22%	62,70%	53,13%	67,82%	56,99%	36,13%	57,40%

Fig. 4.17 – Percentage comparison of INE "home-daily place of residence" divided by the morning rush hour average OD trips in the core zone

Source: (complied by author)

Beforehand it is worth to mentions that some differences compared to the data from INE are to be expected as obtained from operator values take into account all trips that happened during the chosen time interval in the core zone and not only the "home-daily place of residence" commute ones.

We can see that total summarized value from INE data relates to total amount of OD trips from morning rush hour at the value of 57,4% which is significantly lower and can be explained by the reason that INE data represents only the daily commute type of movements, while data from operator represents all movements that happened in the set amount of time in the core region. If we look at the proportional values of total trips originated and destinated to municipalities of the core zone as well as at the proportional values of inner-municipal trips we can see that they generally lay within the interval from 35% to 75%. The biggest

percentage gaps are presented in some of inter-municipal OD pairs such as Castelo de Paiva-Celorico de Basto with the value of 3591%, however this could be explained that this OD pairs are generally characterized by a lower number of recorded trips both in mobile data and in INE which could lead to such distortions (in this particular example the value for this OD pair in morning rush hour pattern equals 1). It is worth to mention that proportional value of periphery zone originated trips from INE compared to morning rush hour value is relatively close at 70,57% taking into account the overall bigger sample of trips in morning rush hour and that "home-daily place of residence" commute pattern from INE is characterized by a bigger percentage of periphery originated trips compared to both rush hour patterns. Next, we can look at Proportional distribution of trips by origin and destination between the core zone municipalities for two movements patterns and INE data (shown in the Figure 4.18).

Origin				
	Morning rush hour	Evening Rush Hour	INE	% of population
Amarante	15,11%	15,28%	12,49%	12,75%
Baião	4,59%	4,60%	3,19%	4,29%
Castelo de Paiva	3,66%	3,53%	3,45%	3,81%
Celorico de Basto	3,80%	3,74%	3,54%	4,32%
Cinfães	4,21%	4,03%	3,18%	4,34%
Felgueiras	12,96%	13,20%	16,65%	13,67%
Lousada	11,57%	11,83%	13,11%	11,59%
Marco de Canaveses	12,99%	12,87%	11,45%	12,12%
Paços de Ferreira	12,76%	12,90%	15,49%	13,60%
Penafiel	16,27%	15,96%	16,23%	17,04%
Resende	2,08%	2,07%	1,23%	2,46%
Destination				
	Morning rush hour	Evening Rush Hour	INE	% of population
Amarante	14,33%	14,58%	12,06%	12,75%
Baião	4,38%	4,43%	3,48%	4,29%
Castelo de Paiva	3,76%	3,51%	3,74%	3,81%
Celorico de Basto	4,28%	4,30%	3,82%	4,32%
Cinfães	4,08%	3,86%	3,27%	4,34%
Felgueiras	12,82%	13,31%	15,70%	13,67%
Lousada	11,88%	12,05%	12,97%	11,59%
Marco de Canaveses	12,24%	12,15%	11,34%	12,12%
Paços de Ferreira	12,97%	13,42%	15,32%	13,60%
Penafiel	17,15%	16,31%	16,99%	17,04%
Resende	2,10%	2,09%	1,32%	2,46%

Fig. 4.18 – Proportional distribution of trips by origin and destination between the core zone municipalities for two movements patterns and INE data

Source: (complied by author)

We can see that generally the proportional distribution of origin and destination trips throughout the municipalities of the core zone corelates between two rush hour patterns, INE values and the proportional distribution of population of core zone.

Finally, we can proceed with comparison of inner-municipal and periphery originated trips proportional values for reviewed patterns from operators' data and INE data (shown in the Table 4.2).

Table 4.2 – Percentage proportion of inner-municipal and periphery zone originated trips for the core zone for two movements patterns and INE data

Source: (complied by author)

	% Of Inner-municipal trips	% Of periphery zone originated trips
Morning rush hour	82,53%	12,06%
Evening rush hour	82,75%	9,91%
INE	83,95%	14,84%

To overview all the notes regarding comparison of INE values to mobile data for morning rush hour. Firstly, there are appear some big gaps in percentages while comparing data from these two resources in the inter-municipalities OD pairs and they could be explained by a relatively small sample of records for this pairs in both sources of data and overall, these OD pairs usually represent relatively small proportion of all trips. Secondly, proportional values for inner-municipality trips and proportional total origin and destination values distributed between municipalities of core zone are consistent and correlate to a high degree between INE and operators' data. Finally, if we take into account bigger size of the sample of OD trips for morning rush hour in mobile data the value of periphery zone originated trips is pretty close in a both sources of data.

4.3.3. Periphery zone analysis

In this section the analysis of periphery zone originated trips is going to be presented. In this analysis we would focus only on the trips that originated from the periphery regions whose destination location was within the borders of a core study zone. Set of 16 municipalities of periphery zone that surrender the core study region were chosen for this analysis. At first, we can look at total morning rush hour average OD trips and their proportional distribution throughout the municipalities of origin which are presented in the Figures 4.19 and 4.20.

		Destination												
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin	Total OD
Origin	Arouca	8	16	428	2	359	7	15	105	3	133	3	1078	
	Cabeceiras de Basto	47	1	0	322	0	9	13	5	1	5	0	402	
	Castro Daire	4	20	1	0	132	2	4	24	0	5	31	223	
	Fafe	198	11	2	1429	0	926	145	65	19	116	4	2916	
	Gondomar	96	16	286	4	58	46	109	104	62	440	9	1231	
	Guimarães	260	16	6	368	6	1699	718	111	242	172	8	3606	
	Lamego	38	208	1	6	24	4	17	29	8	15	298	646	
	Mesão Frio	18	152	1	0	11	3	5	6	2	3	157	358	
	Mondim de Bastos	273	18	2	656	0	45	18	26	1	11	0	1051	
	Paredes	264	38	156	18	69	176	1664	370	2631	4886	10	10282	
	Peso da Régua	62	135	1	4	19	15	28	27	2	15	137	445	
	Santa Marta de Penaguião	28	27	0	2	8	10	9	15	1	11	27	138	
	Santo Tirso	36	4	19	23	3	152	287	33	1100	145		1801	
	Valongo	165	18	30	8	19	99	254	168	340	484	11	1598	
	Vila Real	282	97	5	32	18	38	54	77	10	77	48	736	
	Vizela	65	6	1	12	0	587	530	21	61	44	1	1329	
	Total Destination	1844	782	939	2887	727	3817	3872	1188	4481	6561	742		27840

Fig. 4.19 – Morning rush hour average periphery zone originated trips

Source: (complied by author)

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Arouca	0,74%	1,49%	39,69%	0,18%	33,30%	0,64%	1,40%	9,71%	0,25%	12,36%	0,25%	3,87%
	Cabeceiras de Basto	11,57%	0,15%	0,01%	80,03%	0,00%	2,21%	3,23%	1,33%	0,14%	1,33%	0,00%	1,44%
	Castro Daire	1,66%	9,03%	0,33%	0,21%	59,17%	1,02%	1,80%	10,91%	0,00%	2,11%	13,76%	0,80%
	Fafe	6,79%	0,38%	0,07%	49,01%	0,01%	31,74%	4,99%	2,24%	0,65%	3,99%	0,14%	10,47%
	Gondomar	7,83%	1,28%	23,28%	0,35%	4,70%	3,70%	8,88%	8,48%	5,03%	35,72%	0,75%	4,42%
	Guimarães	7,21%	0,44%	0,16%	10,21%	0,18%	47,12%	19,92%	3,07%	6,70%	4,78%	0,22%	12,95%
	Lamego	5,83%	32,24%	0,10%	0,87%	3,73%	0,61%	2,57%	4,43%	1,20%	2,31%	46,12%	2,32%
	Mesão Frio	5,10%	42,52%	0,16%	0,05%	2,94%	0,71%	1,47%	1,73%	0,55%	0,86%	43,90%	1,29%
	Mondim de Bastos	26,01%	1,69%	0,16%	62,44%	0,02%	4,26%	1,70%	2,50%	0,12%	1,09%	0,00%	3,77%
	Paredes	2,57%	0,36%	1,52%	0,18%	0,67%	1,71%	16,18%	3,60%	25,59%	47,51%	0,10%	36,93%
	Peso da Régua	13,91%	30,30%	0,33%	0,97%	4,18%	3,30%	6,24%	6,14%	0,55%	3,38%	30,71%	1,60%
	Santa Marta de Penaguião	20,33%	19,64%	0,26%	1,14%	5,97%	7,24%	6,70%	11,01%	0,69%	7,75%	19,27%	0,50%
	Santo Tirso	1,99%	0,20%	1,05%	1,26%	0,18%	8,46%	15,95%	1,84%	61,05%	8,03%	0,00%	6,47%
	Valongo	10,32%	1,15%	1,90%	0,51%	1,22%	6,17%	15,93%	10,54%	21,27%	30,30%	0,68%	5,74%
	Vila Real	38,26%	13,19%	0,67%	4,28%	2,39%	5,19%	7,33%	10,47%	1,32%	10,44%	6,47%	2,65%
	Vizela	4,93%	0,47%	0,06%	0,93%	0,02%	44,19%	39,90%	1,60%	4,57%	3,28%	0,07%	4,77%
	Total Destination	6,62%	2,81%	3,37%	10,37%	2,61%	13,71%	13,91%	4,27%	16,10%	23,57%	2,67%	100,00%

Fig. 4.20 – Proportional distribution of morning rush hour average periphery zone originated trips by municipalities

Source: (complied by author)

We can see that the biggest number of trips were made from the municipalities of Paredes, Guimarães and Fafe which collectively represent 60,35% of the trips originated from chosen set of municipalities of periphery zone, which is corelates with their total population and geographical proximity to the Tâmega e Sousa region, as these municipalities are the most populated administrative subjects that surrender core zone. It is worth to mention that municipality of Paredes which represents by far the biggest number of originated trips is used to be part of Tâmega e Sousa region and represented around fifth of its' population before in 2013 it was incorporated into the Porto Metropolitan area which can explain why it is represent more than a third of trips originated from analyzed municipalities of a periphery zone. Distribution of trips throughout destination municipalities of core zone corresponds with their population numbers and proximity to the main areas of trips origin and is in conformity with general "west-east" population trend that is globally observed in Portugal. One "anomaly" should be highlighted. It happens in the core region municipality of Celorico de Basto which attracts mostly from surrounding regions in total 10,37% of trips while representing only 4,32% of core zone population. Next, we can look at the values for evening rush hour pattern which are shown in the Figures 4.21 and 4.22.

		Destination												
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin	Total OD
Origin	Arouca	11	22	352	2	210	11	23	112	5	184	5	938	
	Cabeceiras de Basto	55	8	0	345	1	29	15	15	0	19	0	488	
	Castro Daire	3	7	3	15	82	0	2	22	0	7	16	157	
	Fafe	131	8	1	1273	2	1116	142	27	22	48	4	2774	
	Gondomar	99	17	172	8	22	53	79	84	120	306	2	961	
	Guimarães	131	11	2	170	2	1584	687	77	284	117	1	3066	
	Lamego	52	199	2	1	66	7	16	26	2	7	272	650	
	Mesão Frio	34	168	0	1	10	1	1	7	1	3	155	380	
	Mondim de Bastos	345	24	0	1066	2	66	16	32	2	15	0	1568	
	Paredes	215	12	103	15	34	181	1874	288	2979	4232	1	9934	
	Peso da Régua	33	76	6	7	12	3	6	17	1	12	82	255	
	Santa Marta de Penaguião	24	29	1	3	5	1	3	6	1	3	21	98	
	Santo Tirso	46	3	6	14	3	158	267	25	1142	83	0	1747	
	Valongo	110	17	33	11	12	67	160	114	451	292	3	1270	
	Vila Real	210	100	4	17	19	12	25	59	4	23	52	525	
	Vizela	44	3	0	43	2	771	390	20	79	43	2	1397	
	Total Destination	1544	704	688	2990	482	4060	3706	929	5093	5395	616		26206

Fig. 4.21 – Evening rush hour average periphery zone originated trips

Source: (complied by author)

		Destination											
		Amarante	Baião	Castelo de Paiva	Celorico de Basto	Cinfães	Felgueiras	Lousada	Marco de Canaveses	Paços de Ferreira	Penafiel	Resende	Total Origin
Origin	Arouca	1,22%	2,33%	37,57%	0,17%	22,37%	1,16%	2,50%	11,96%	0,58%	19,62%	0,52%	3,58%
	Cabeceiras de Basto	11,34%	1,72%	0,00%	70,80%	0,12%	5,94%	3,06%	3,01%	0,01%	3,99%	0,00%	1,86%
	Castro Daire	1,83%	4,28%	2,15%	9,32%	52,29%	0,13%	1,20%	13,77%	0,00%	4,52%	10,51%	0,60%
	Fafe	4,71%	0,31%	0,03%	45,89%	0,08%	40,24%	5,13%	0,97%	0,79%	1,72%	0,13%	10,58%
	Gondomar	10,26%	1,78%	17,96%	0,82%	2,24%	5,49%	8,19%	8,77%	12,48%	31,85%	0,17%	3,67%
	Guimarães	4,27%	0,36%	0,08%	5,56%	0,05%	51,65%	22,41%	2,50%	9,27%	3,82%	0,04%	11,70%
	Lamego	8,01%	30,59%	0,23%	0,22%	10,20%	1,09%	2,53%	3,93%	0,25%	1,09%	41,84%	2,48%
	Mesão Frio	8,95%	44,18%	0,00%	0,18%	2,62%	0,14%	0,23%	1,76%	0,29%	0,78%	40,86%	1,45%
	Mondim de Bastos	22,00%	1,51%	0,02%	67,98%	0,10%	4,24%	1,02%	2,01%	0,13%	0,97%	0,02%	5,98%
	Paredes	2,17%	0,12%	1,04%	0,15%	0,34%	1,82%	18,87%	2,90%	29,99%	42,60%	0,01%	37,91%
	Peso da Régua	12,86%	29,73%	2,54%	2,89%	4,55%	1,07%	2,18%	6,74%	0,45%	4,80%	32,18%	0,97%
	Santa Marta de Penaguião	24,82%	30,09%	1,23%	3,00%	5,13%	1,53%	3,46%	5,98%	0,57%	2,75%	21,44%	0,37%
	Santo Tirso	2,66%	0,18%	0,36%	0,80%	0,17%	9,02%	15,26%	1,40%	65,37%	4,77%	0,01%	6,67%
	Valongo	8,64%	1,31%	2,63%	0,88%	0,92%	5,31%	12,57%	8,95%	35,51%	23,02%	0,26%	4,84%
	Vila Real	40,08%	18,97%	0,81%	3,18%	3,65%	2,32%	4,67%	11,25%	0,77%	4,46%	9,83%	2,00%
	Vizela	3,17%	0,21%	0,00%	3,06%	0,15%	55,21%	27,91%	1,41%	5,63%	3,10%	0,14%	5,33%
	Total Destination	5,89%	2,68%	2,62%	11,41%	1,84%	15,49%	14,14%	3,54%	19,43%	20,59%	2,35%	100,00%

Fig. 4.22 – Proportional distribution of evening rush hour average periphery zone originated trips by municipalities

Source: (complied by author)

Here we can see that the total periphery zone originated trips value is lower than in morning rush hour pattern. Overall, distribution of destination municipalities of core zone is generally similar. We can observe that the 3 previously mentioned municipalities represent 60,19% of total trips originated from chosen set of municipalities of periphery zone. Celorico de Basto "anomaly" appears again even on the larger scale. It is worth to mention that in both patterns Santa Marta de Penguião was the least represented municipality in proportion to total trips originated in 16 municipalities surrounding the core zone with values of 0,50% and 0,47% for morning and evening rush hour patterns accordingly.

In the case of morning rush hour pattern selected set of municipalities represents 65,46% of total periphery zone originated trips, while in the case of evening rush hour pattern it represents 75,81% of total periphery zone originated trips. Taking into account that morning rush hour is characterized by a bigger proportion of overall periphery zone originated trips it could mean that more people come to core zone outside of borders of periphery zone municipalities that surrender core zone in the time of morning rush hour compared to the time of evening rush hour.

4.3.4. Map analysis

In the last section of fourth chapter, we would like to do a graphical presentation and analysis of obtained results during the mobile data analysis for Tâmega e Sousa region by developing and analyzing maps with average trip values for two already observed movement patterns: morning rush hour (7-10 a.m.) and evening rush hour (5-8 p.m.). Results for both core (shown in the Figures 4.23 and 4.24) and periphery zones (shown in the Figures 4.25 and 4. 26) would be presented. Before analyzing the obtained maps, it is worth to mention that only inter-municipal trips would be presented, for each municipality only the OD movements with biggest values are presented which as a rule mostly happened to be the movements between neighbouring municipalities.

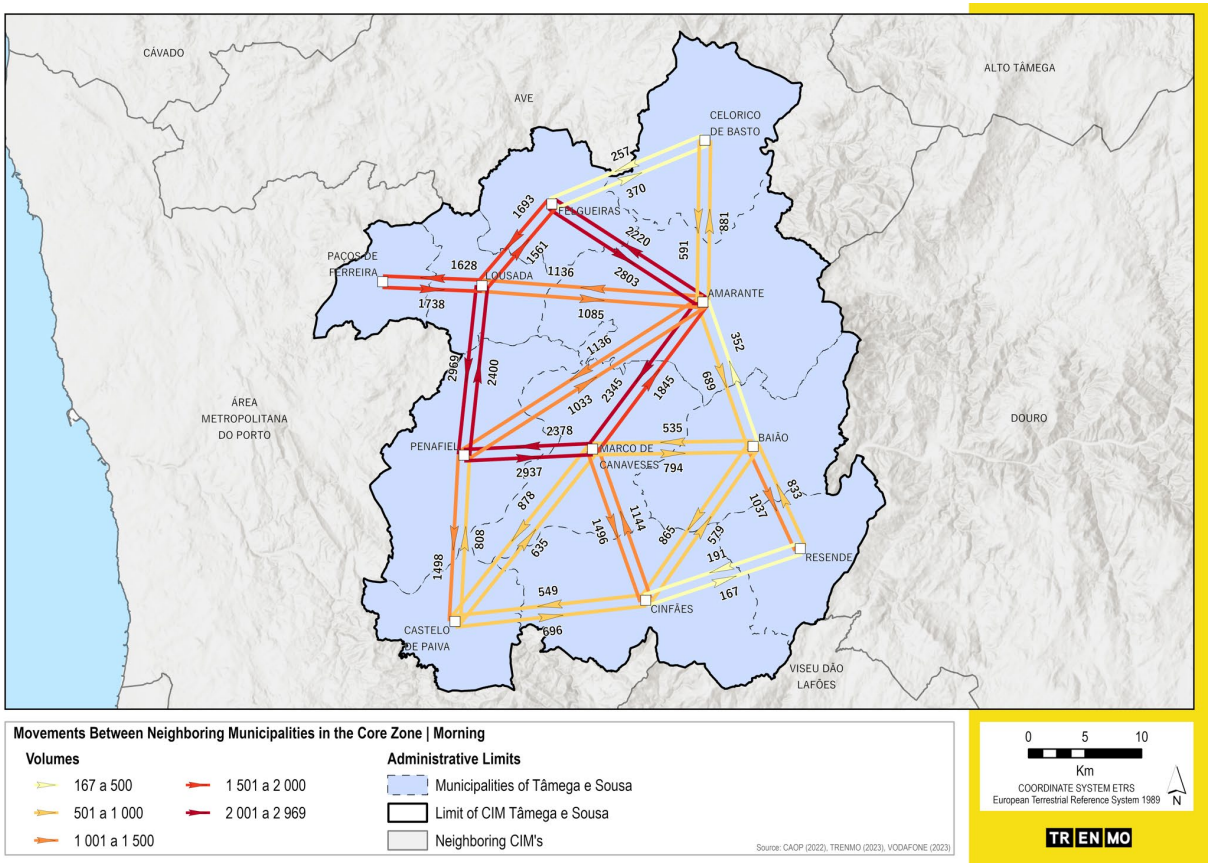


Fig. 4.23 – Map of morning rush hour average OD trips in the core zone

Source: (complied by author)

We can see that overall bigger number of inter-municipal trips is presented in the north-west part of the core region and they are distributed between 6 municipalities: Pacos de Ferreira, Lousada, Penafiel, Marco de Canaveses, Amarante and Felgueiras. It corelates with the demographic distribution of Tâmega e Sousa as combined these parts of the region represent 80,77% of its total population. The sum of the trips between these 6 municipalities accounts for 66,11% of all presented trips in the core zone. To finish up analyzing results for the core zone we can look at the map produced for the evening rush hour pattern.

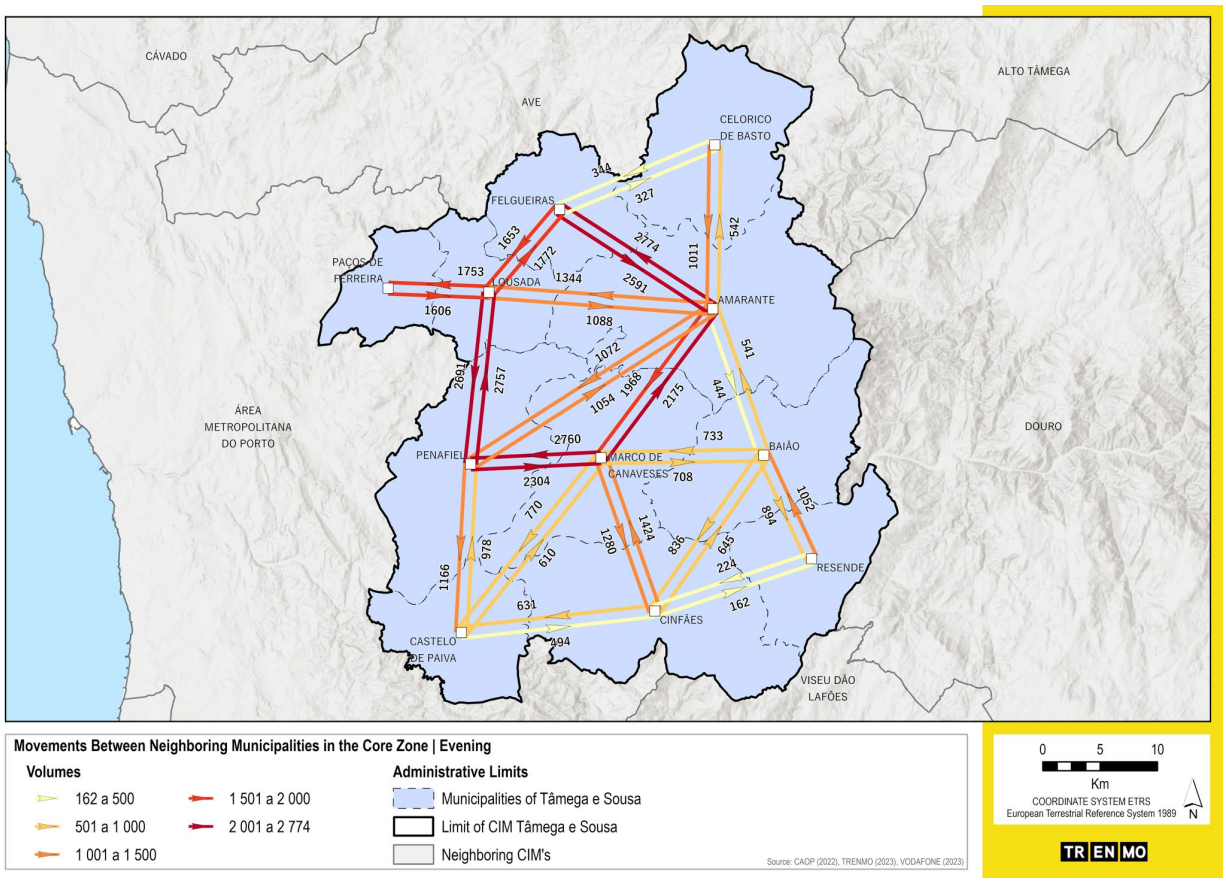


Fig. 4.24 – Map of evening rush hour average OD trips in core zone

Source: (complied by author)

Slight increase in intensity of inter-municipalities trips of 1% is observed compared to morning rush hour average values. 6 most populated municipalities that form north-western part of the region account for 66,48% of total presented trips. These results again generally corelates with values from Table 4.2, where evening rush hour was shown to have the biggest proportion of inter-municipalities trips out of analyzed movements patterns. Now after we have analyzed results for maps of the core zone, we can proceed with the analysis of maps for periphery zone originated trips again for two patterns that were already mentioned. For them three biggest OD pairs in terms of value of total trips directed for each municipality of core region were taken into account.

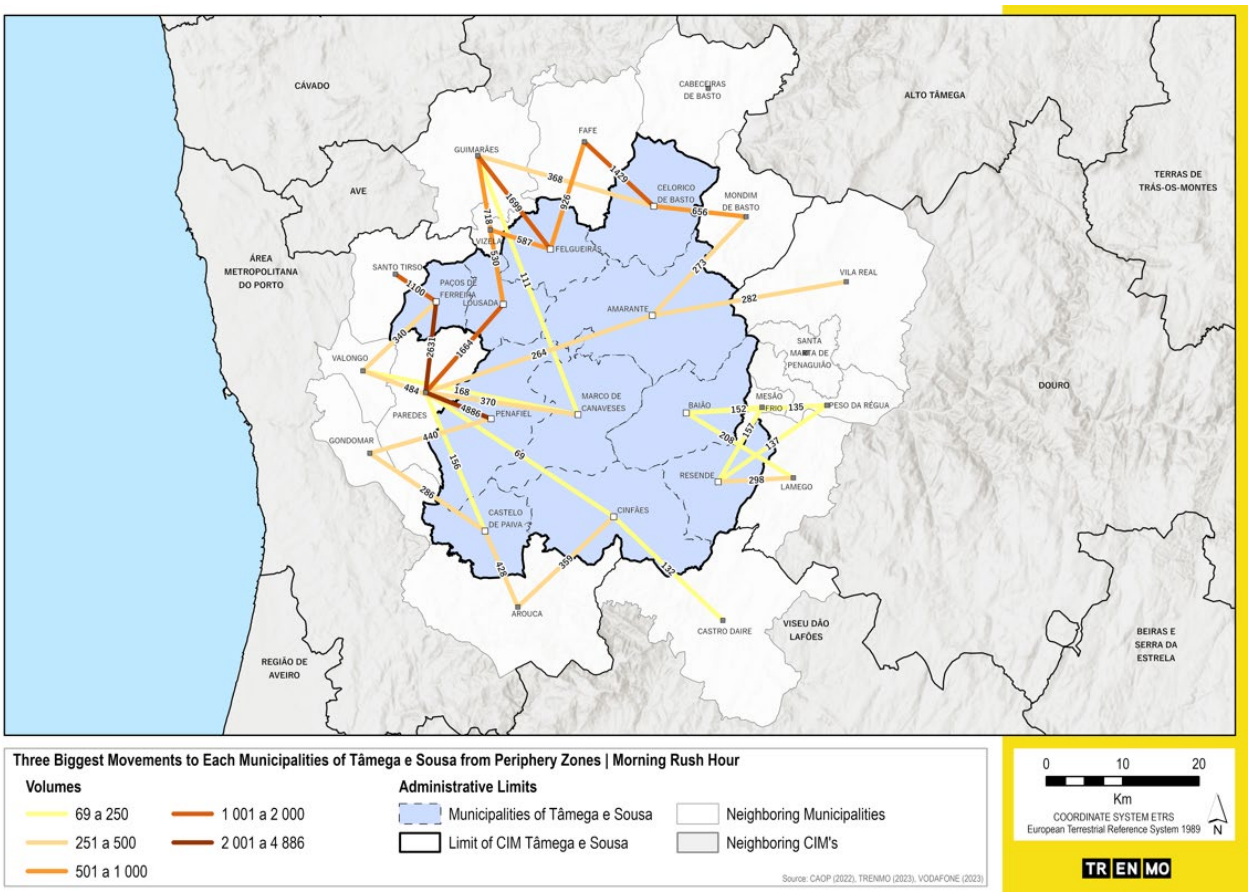


Fig. 4.25 – Map of morning rush hour average periphery zone originated trips

Source: (complied by author)

Here we can see that the biggest value of periphery zone originated movements are concentrated north-western part of periphery zone that surrenders the core zone which corelates with population distribution of the area as there are located the most populated municipalities. As a rule, the biggest movements to destination municipalities of core zone originate from the closest municipalities of observed periphery zone. As was mentioned in the section 4.3.3 devoted to analysis of the periphery zone the municipality of Paredes produces by far the biggest number of trips out of all other municipalities of periphery zone. At the same time the municipalities of Santa Maria de Penaguião and Cabeceiras de Basto don't have any numeric trip value on the map as they haven't passed the condition of representing at least one of the three biggest movements to any municipality of core zone which corelates to their proportional values of total periphery zone originated trips which were obtained in the previous section (0,5% and 1,44% accordingly). Next, we can analyze map for the evening rush hour.

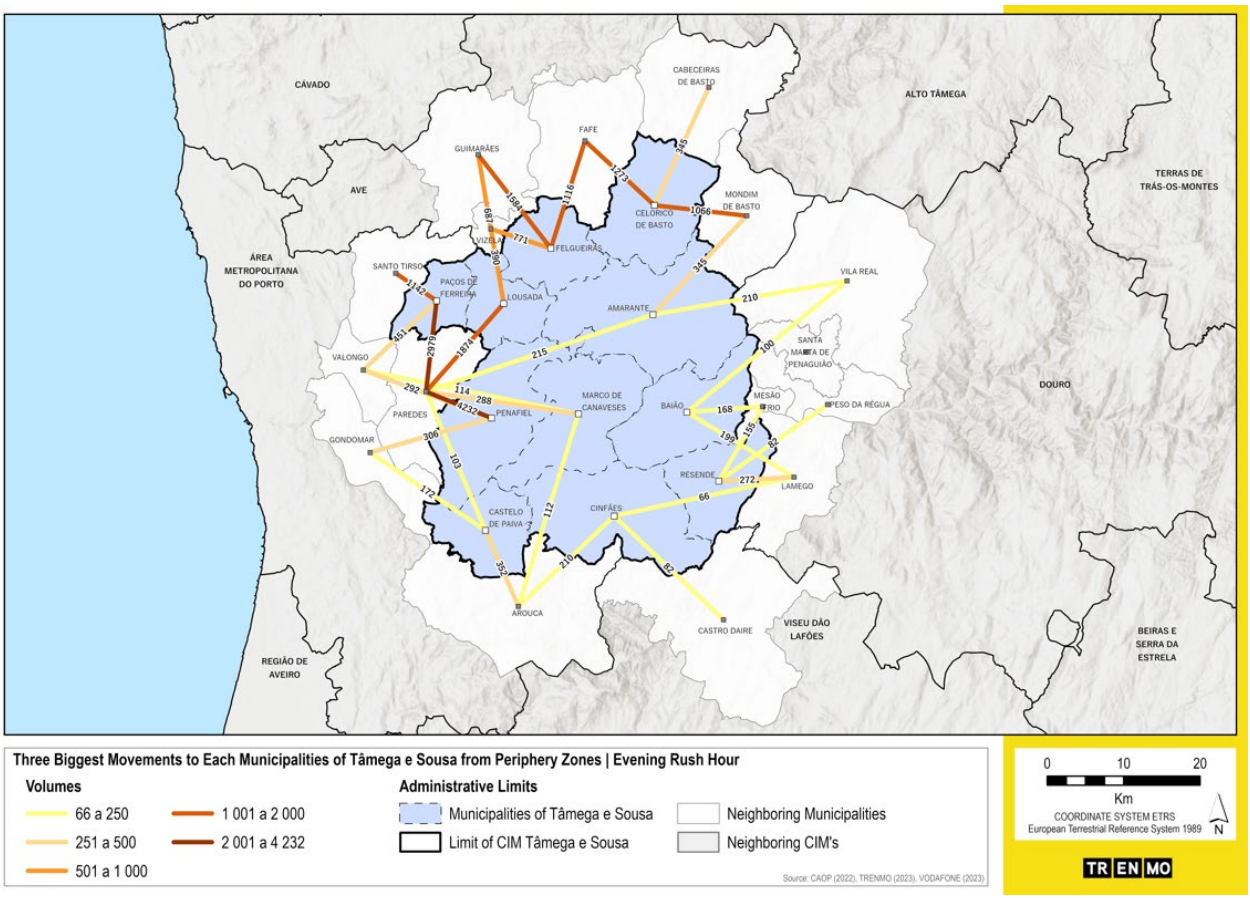


Fig. 4.26 – Map of morning rush hour average periphery zone originated trips

Source: (complied by author)

While overall distribution of movements remains pretty similar compared to the morning rush hour there are some changes on the level of OD pairs of municipalities which means that some municipalities are no longer represent one of the three biggest movements to the municipality of the core zone and at the same time some other OD pairs are formed. As an example, municipality of Cabeceiras de Basto have formed OD pair with the municipality of Celorico de Basto while it didn't have any representation in the morning rush hour map which also corelates with the results of the periphery zone analysis as the proportional value out of all periphery zone originated trips for this municipality is higher compared to the evening rush hour. Same can be said to the other periphery municipalities of the new formed OD pairs such as Lamego-Cinfaes which didn't have representation in the morning rush hour. At the same time municipality of the Santa Maria de Penaguião again don't have any representation as it accounts to the 0,37% of total periphery zone originated trips which is even less than in the morning rush hour pattern.

Summarizing this paragraph, the results of analysis of developed maps generally corelate with the results of analysis of core and periphery zones that were presented beforehand.

5

CONCLUSION

American scientist D. Hellerstein coined the term "industrial data revolution" to characterize the advent of Big Data in 2008 (Hellerstein, 2008). The 21st century witnessed a technological revolution in the acquisition and analysis of data. This revolution introduced new data sources and enhanced data processing methods, affording researchers numerous opportunities to supplement and improve traditional statistical knowledge.

Mobile communication has become an integral element of modern society, with penetration level of 96.4 subscribers per 100 inhabitants worldwide. Nearly every person on earth lives within the range of a mobile cell signal, indicating its ubiquitous use and providing highly representative data with sample of almost 100% of population. This data source is particularly useful for developing countries and regions that struggle with conventional statistical data collection.

The rise of big data has ushered in a data revolution in the industrial sector, creating new opportunities for socioeconomic analysis. Specifically, mobile phone data is a valuable resource, casting light on human movement patterns and supplementing conventional statistics. With their global reach and accessibility, mobile communication data hold great potential for advancing research and gaining a global understanding of socioeconomic processes. As we continue to leverage big data, researchers, policymakers, and businesses will be able to gain deeper insights and make more informed decisions to effectively address societal challenges.

In the framework of master's dissertation, by performing a comprehended review of existing scientific literature it was possible to shed light on the main countries, institutions and areas of studies that have showed interest in use of mobile operators' data. It was concluded that research is primary focused on USA and Europe and that main areas of studies include analysis of statistics, social and behavioural studies, analysis of settlement systems and mobility studies. It was also underlined that mobile data play special role in international projects in developing countries as it can take the place of traditional sources of statistics due to reason that developing countries often have difficulties with access to these types of information as national surveys could have been not implemented in many years. The first work on this topic could be considered "Real Time Rome" project conducted by MIT in 2006. In the framework of that study researchers were able to analyze how two different events in Rome affected the existing settlement system. By summarizing the insights obtained during performance of literature review it was possible to produce a SWOT analysis of mobile operators' data. Overall, it can be concluded that mobile operators' data serves

as a promising source of information not only in field of transport planning but for vast area of studies. At the same, it is important to acknowledge the challenges associated with the use of mobile operators' data, including issues of privacy and data protection.

Second part of the master's dissertation were dedicated to analysis of existing techniques, mechanisms and methods in regard of obtaining and processing of mobile data. It was observed that mobile operators store information about all actions of interaction between mobile subscribers and telecommunication network in form of Call Data Records for the main reason of assessment of their own business processes. Structure of types of information collected in Call Data Records as well as an example of rows from them were illustrated. It was concluded that telecommunication operators use various mechanisms for anonymization of data about subscribers before passing information samples to third parties in accordance to existing worldwide legislation on personal data privacy. Two types of Call Data Records were identified, active and passive ones. While active records require ongoing information exchange between subscriber and network (voice message, SMS, etc.), passive records could be obtained even if the phone is not active. Passive records are more desirable for a research purpose due to their higher frequency of data collection and positioning accuracy.

There are numerous methods of obtaining subscribers' geolocation with GPS-based being one of the most accurate. But mobile operators do not have access to most of them. While analyzing the mechanisms of location subscriber in space available to operators it was observed that all of them are tied to getting location based on telecommunication cell network. Various methods of geopositoning were overviewed such as Cell ID; triangulation, timing advance and Location Based Service, which includes few of mentioned methods and is the most accurate out of them (Resch, B. et al., 2005). In the end of analysis of techniques of mobile data mining review of several studies that implemented developing of Origin-Destination matrices for mobility research was conducted, which shed a light on existing practices of utilizing mobile data in order to understand behavioural patterns of people in terms of their daily performed trips.

In the second part of the chapter devoted to processing of mobile data analysis of various methods of data pre-processing was made. It was noted that raw mobile data is very unreliable and "noisy" and that there exists various of pre-procession methods to combat that issue, main one being Noise Reduction and Trajectory Segmentation. Main type of noise are oscillations that happen due to the nature of geopositioning process so the main focus of Noise Reduction is on Oscillation Filtering. Various methods of Oscillation Filtering were overviewed such as Kalman filter and Gaussian processes, as well as existing literature with examples of utilizing those methods in mobility research. It was observed that using these methods could help to detect and delete Oscillation Pairs from arrays of data, which could represent up to 6% of total number of records (Wu et al., 2014). Review of Trajectory methods, which are developed to divide continuous movement data of CDR into meaningful segments, such as Triangulation and K-means algorithm as well as example of studies that implemented them in mobility research was presented. Overall, it helped to conclude that right choice and execution of various data pre-procession methods is crucial part of utilization of mobile data in transport and mobile studies, that sometimes need from researchers a certain level of creativity as well as possibility to develop new approaches due to the lack of well-developed algorithms of actions and continuous emergence of a new challenges in mobile data analysis.

Final part of the third chapter involves evaluation of data processing methods which include Itinerary Reconstruction, Transport mode detection and Location patterns recognition. All three of these methods aimed to achieve better representation of people's behaviour in their daily commute and serve as tools to

meet the objectives of researchers. Overview of existing literature on use of these methods helped shed a light on their general performance in studies and usefulness with relation to master's dissertation.

In the last chapter of main body of master's dissertation results of the case study project on a Tâmega e Sousa region in Portugal, that was implemented within the TRENMO company, are presented. First part of the chapter is devoted to the overview of the region and helps to understand the context of a study area and zoning. Second part of the chapter characterizes raw mobile data sample obtained from a mobile provider company Vodafone and exhibits process of data extraction and pre-procession, which included selection of mobile records needed for analysis as well as separation of unnecessary records. First part of case study analysis was devoted to understanding of two movements patterns in core region of study area. For analysis were chosen morning rush hour and evening rush hour patterns as they would ensure representativeness with most common population movements patterns, such as morning "home-daily place of residence" commute and evening "place of daily residence-home" commute patterns. It was observed that in all chosen patterns inner-municipalities movements represent the biggest proportion of all movements in study area. Next part of analysis was focused on comparing obtained data with data from INE. It was concluded that data from INE corelate with mobile data in terms of proportion of inner-municipalities movements as well as in terms of distribution of trips throughout the municipalities of core zone. Simultaneously it should be noted that there have been observed considerably high percentage "gaps" in inter-municipalities OD pairs which could be explained by a relatively small values for those pairs in both data sources which could lead to such distortions. It can be summarized that generally mobile data corelates with data from INE but at the same appeared "gaps" do not allow to say that both of this data sources are interchangeable for mobility research.

Third part of analysis was focused on periphery zone originated trips for two main patterns of morning and evening rush hours. It was shown that selected set of municipalities represent for morning rush hour 75% and for evening rush hour 65% of all periphery zone originated trips and that municipality of Paredes represent biggest proportion of periphery zone originated trips from all municipalities of the zone. It is also worth to mention an "anomaly" appeared during the analysis which are municipality of Celorico de Basto. In the last part of case study analysis geographical representation of obtained results in form of maps for both core and periphery zones for two movements patterns were presented. It was concluded that results of maps analysis corelate with the conclusion from data analysis. Overall, it can be summarized that all three objectives of master's dissertation set out in the 1 chapter were met.

In conclusion, this research has demonstrated that mobile operators' data has the potential to transform the landscape of transport planning. It has obvious advantages over the traditional data sources and could replace them in conditions when they are scarce. Nonetheless, it is important to acknowledge the challenges associated with the use of mobile operators' data, including issues of privacy and data protection. It cannot be said that at this moment mobile data can totally take a place of traditional statistics sources. they both have their application and better work symbiotically. But in the end, with a concerted effort from researchers, mobile operators, and policymakers, the integration of mobile operators' data in transport planning can pave the way for more resilient urban mobility solutions.

BIBLIOGRAPHY

Ahas, R.; Aasa, A.; Silm, S.; Tiru, M. Mobile (2007). *Positioning Data in Tourism Studies and Monitoring: Case Study in Tartu, Estonia*. In Information and Communication Technologies in Tourism; Springer: Berlin/Heidelberg, Germany; pp. 119–128.

Ahas, R.; Aasa, A.; Yuan, Y.; Raubal, M.; Smoreda, Z.; Liu, Y.; Ziemlicki, C.; Tiru, M., Zook, M. (2015). *Everyday space–time geographies: Using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn*. Int. J. Geogr. Inf. Sci. 29, 2017–2039

Ahas, R.; Silm, S.; Järv, O.; Saluveer, E.; Tiru, M. (2010). *Using mobile positioning data to model locations meaningful to users of mobile phones*. J. Urban Technol., 1, 3–27

Ahas, R.; Silm, S.; Saluveer, E.; Järv, O. (2009). *Modelling home and work locations of populations using passive mobile positioning data. In Location Based Services and TeleCartography* II; Springer: Berlin/Heidelberg, Germany; pp. 301–315.

Alexander, Lauren, Shan Jiang, Mikel Murga et Marta C Gonzales (2015). *Origin-destination trips by purpose and time of day inferred from mobile phone data*. In: Transportation Research Part C: Emerging Technologies 58, p. 240–250

Alharbi, S. S., Bao, J., Li, X., & Wang, G. (2019). *Itinerary reconstruction from location-based social networks data*. ISPRS International Journal of Geo-Information, 8(1), 27.

Aurenhammer, Franz, and Herbert Edelsbrunner. (1984). *An optimal algorithm for constructing the weighted Voronoi diagram in the plane. Pattern Recognition* 17.2: 251-257.

Bajardi, P.; Delfino, M.; Panisson, A.; Petri, G.; Tizzoni, M. (2015). *Unveiling patterns of international communities in a global city using mobile phone data*. Data Sci., 4, 3.

Barth, M. (2018). *Sustainable Transport Planning: Ethical Dimensions. In Handbook of Ethics and Planning Research*, pp. 405-418. Edward Elgar Publishing.

Basiri, A., Winstanley, A., Moore, T., Biehl, M., & Amirian, P. (2018). *Urban Mobility Analytics: Overview and Future Directions*. IEEE Transactions on Intelligent Transportation Systems, 19(5), 1575-1597.

Bayir, Murat Ali, Murat Demirbas et Nathan Eagle (2009). *Discovering spatiotemporal mobility profiles of cellphone users.* In: World of Wireless, Mobile and Multimedia Networks & Workshops, WoWMoM. IEEE International Symposium on a. IEEE, p. 1–9 (cf. p. 7)

Bekhor, S.; Cohen, Y.; Solomon, C. (2013) *Evaluating Long Distance Travel Patterns in Israel by Tracking Cellular Phone Positions*. J. Adv. Transp., 47, 435–446.

Berlingerio, M.; Calabrese, F.; Lorenzo, G.; Nair, R.; Pinelli, F.; Sbodio, M. (2013). *All Aboard: A system for exploring urban mobility and optimizing public transport using cellphone data*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases; Springer: Berlin/Heidelberg, Germany; pp. 663–666.

Bhattacharya, S., Dey, A., Jain, R., & Bhaumik, C. (2015). *Understanding mobile app usage patterns using time-series segmentation*. In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services (pp. 163-172).

Birkin, M.; Clarke, C.; Clarke, M. (2017) *Retail Location Planning in an Era of Multi-Channel Growth*. Routledge: London, UK; 245p.

Blat, J., Calabrese, F., Ferrari, L., Soto, V., & Noulas, A. (2016). *Anonymity preserving and high precision processing of human mobility data for understanding city dynamics.* In Proceedings of the 25th International Conference on World Wide Web Companion (pp. 119-120).

Blondel, V.; Krings, G.; Thomas, I. (2010). *Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone*. Bruss. Stud., 42, 1–12.

Bolla R, Davoli F. (2000). *Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network*, Proc. IEEE WCNC, Chicago, USA.

Bulling, A., Blanke, U., & Schiele, B. (2014). *A tutorial on human activity recognition using body-worn inertial sensors*. ACM Computing Surveys (CSUR), 46(3), 33.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira Jr, J., Ratti, C. (2013*). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example*. Transportation Research Part C: Emerging Technologies, 26, 301-313.

Chen, H., Chen, C., & Chen, L. (2018). *Correction of GPS positioning errors caused by oscillation on curves*. IEEE Transactions on Intelligent Transportation Systems, 19(6), 1770-1782.

Csaji, Balázs Cs, Arnaud Browet, Vincent A Traag et al. (2013*). Exploring the mobility of mobile phone users*. In: Physica A: Statistical Mechanics and its Applications 392.6, p. 1459–1473 (cf. p. 9, 17, 24).

Cáceres, I., Ruiz, M. L., & Saladié, P. (2007). *Evidence for bronze age cannibalism in El Mirador Cave (Sierra de Atapuerca, Burgos, Spain).* American Journal of Physical Anthropology, 133(3), 899–917.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013*). Unique in the Crowd: The Privacy Bounds of Human Mobility*. Scientific Reports, 3, 1376.

De Nadai, M., Staiano, J., Larcher, R., Sebe, N., & Quercia, D. (2016). *The death and life of great Italian cities: A mobile phone data perspective.* In Proceedings of the 25th International Conference on World Wide Web (pp. 413-423).

Deville, P.; Linard, C.; Martine, S.; Gilbert, M.; Steven, F.; Gaughan, A.; Blondel, V.; Tatem, A. (2014). *Dynamic population mapping using mobile phone data.* Proc. Natl. Acad. Sci. USA, 111, 88–93.

Eagle, N.; Macy, M.; Claxton, R. (2010). *Network diversity and economic development*. Science, 328, 1029–1031

Elias and Daniel. (2016). *SOMOBIL–improving public transport planning through mobile phone data analysis.* Transportation Research Procedia 14: 4478-4485

Eurostat. (2014). *Feasibility study of the use of mobile positioning data for tourism statistics*. In Consolidated Report Eurostat; Contract No 30501.2012.001–2012.452; Eurostat: Luxembourg; 34p

Gao, S., Li, L., Zhu, Y., & Zhao, D. (2014). *Exploring urban human mobility patterns: A study using large-scale taxi GPS data in Shanghai*. PloS one, 9(12), e113623.

Gao, S., Ye, X., & Yin, J. (2012). *Trajectory segmentation of mobile phone users for urban mobility analysis.* Proceedings of the 10th International Conference on Ubiquitous Intelligence and Computing (UIC 2012), 498-512.

Gordon, Mason B. Haris N. Koutsopoulos, Nigel H.M. Wilson. (2018). *Estimation of population origin–interchange–destination flows on multimodal transit networks,* Transportation Research Part C: Emerging Technologies, 90, pp. 350-365.

Graells-Garrido, E., & Saez-Trumper, D. (2016). *A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow*. Conference: The Second International Conference.

Han, J., Pei, J., Kamber, M., & Dong, G. (2011). *Data Mining: Concepts and Techniques* (3rd ed.)

R. G. Lyons. (2010). Understanding Digital Signal Processing, 3rd ed., Prentice Hall.

Horak R. (2007). *Telecommunications and data communications handbook*. – John Wiley & Sons

Huang, J., Cheng, C., Xie, X., & Li, X. (2013). *Understanding individual mobility patterns from urban sensing data: A mobile phone trace example*. Transportation Research Part C: Emerging Technologies, 26, 301-313.

Jiang, S.; Ferreira, J.; Gonzalez, J.; Gonzalez, M. (2017). *Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore*. IEEE Trans. Big Data, 2, 208–219.

Joana Hora, Teresa Galvão Dias, Ana Camanho, Thiago Sobral, (2017). *Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system,* Transportation Research Procedia, 27, 2017, pp. 664-671.

Järv, O. (2013). *Mobile Phone Based Data in Human Travel Behavior Studies: New Insights from a Longitudinal Perspective*; University of Tartu Press: Tartu, Estonia; 63p.

Katayoun Farrahi and Daniel Gatica-Perez. (2011). *Discovering routines from large-scale human locations using probabilistic topic models*. In: ACM Transactions on Intelligent Systems and Technology (TIST) 2.1, p. 3.

Kim, S., Song, S., & Kim, Y. (2017). *A Smoothing Method for Reducing GPS Positioning Noise in Smartphone-Based Travel Surveys*. Sensors (Basel, Switzerland), 17(7), 1677.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014*). Exploring Universal Patterns in Human Home-Work Commuting from Mobile Phone Data*. PLoS ONE, 9(6), e96180.

Lenormand M, Picornell M, Cantú-Ros OG, Tugores A, Louail T, Herranz R, et al. (2014). *Cross-Checking Different Sources of Mobility Information*. PLoS ONE 9(8): e105184.

Li, Y., Yu, C., & Song, W. (2021). *An Intelligent Compensation Method of MEMS Inertial Sensors for Improving Activity Recognition in a Smartphone*. Sensors, 21(4), 1230.

Louail, T.; Lenormand, M.; Ros, O.-K.; Picornell, M.; Herranz, R.; Frias-Martinez, E.; Ramasco, J.; Barthelemy, M. (2014). *From mobile phone data to the spatial structure of cities*. Sci. Rep., 4, 5276.

Lu, X.; Bengtsson, L.; Holme, P. (2012). *Predictability of population displacement after the 2010 Haiti earthquake*. Proc. Natl. Acad. Sci. USA, 29, 11576–11581

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2008). *L-Diversity: Privacy Beyond K-Anonymity. ACM Transactions on Knowledge Discovery from Data*, 1(1), 3.

Masso, A.; Silm, S.; Ahas, R. (2018). *Generational differences in spatial mobility: A study with mobile phone data*. Popul. Space Place, 25, e2210

Namiot D. (2018). *On the assessment of socio-economic effects of the city railway*. International Journal of Open Information Technologies. – T. 6. – #. 1. – S. 92-103

Nemeškal, J.; Ouˇrednícˇek, M.; Pospíšilová, L. (2020). *Temporality of urban space: Daily rhythms of a typical week day in the Prague metropolitan area.* J. Maps, 1, 30–39.

Novak, J.; Ahas, R.; Aasa, A.; Silm, S. (2013) *Application of mobile phone location data in mapping of commuting patterns and functional regionalization: A pilot study of Estonia*. J. Maps, 1, 10–15.

Novak, J.; Temelova, J. (2012). *Everyday Life and Spatial Mobility of Young People in Prague: A Pilot Study Using Mobile Phone Location Data*. Sociol. Casopis, 5, 911–938

Olsson, G. (1970) *Explanation, Prediction and Meaning Variance: An Assessment of Distance Interaction Models*. Econ. Geogr., 46, 223–233.

Ourednícek, M.; Nemeškal, J.; Pospíšilova, L.; Hampl, M. (2019). *The Delimitation of Metropolitan Areas for the Integrated Territorial Investments: Technical Methodology*; Ministry of Regional Development: Prague, Czech; 70p.

Pithakkitnukoon, S., Smoreda, Z., Olivier, P. (2012). *Socio-Geography of Human Mobility: A Study Using Longitudinal Mobile Phone Data*. PLoS One, 7(6).

Ratti, C. (2005). *Mobile Landscape—Graz in real time*. In Proceedings of the 3rd Symposium on TeleCartography in Vienna University of Technology, Vienna, Austria, 28–30; pp. 28–30.

Ratti, C.; Sobolevsky, S.; Calabrese, F.; Andris, C.; Reades, J.; Martino, M.; Claxton, R.; Strogatz, S. (2010). *Redrawing the Map of Great Britain from a Network of Human Interactions*. PLoS ONE, 5, e14248

Ratti, Carlo, Dennis Frenchman, Riccardo Maria Pulselli et Sarah Williams (2006). *Mobile landscapes: using location data from cell phones for urban analysis*. In: Environment and Planning B: Planning and Design 33.5, p. 727–748

Reades, J.; Calabrese, F.; Sevtsuk, A.; Ratti, C. (2007). *Cellular census: Explorations in urban data collection*. IEEE Pervas. Comput., 6, 30–38.

Resch, B.; Romirer-Maierhofer, P. (2005). *Global Positioning in Harsh Environments*; Technical Report, IDE0504; School of Information Science, Computer and Electrical Engineering Halmstad University: Halmstad, Sweden; pp. 22–40.

Schlich, R., Schönfelder, S., Hanson, S. and Axhausen, K. (2004). *Structures of Leisure Travel: Temporal and Spatial Variability*. Transport Review, 24, 219–237

Schneider, Christian M. (2013) *Unravelling daily human mobility motifs*. Journal of The Royal Society Interface 10.84: 20130246.

Silm, S.; Ahas, R. (2014). *The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset*. Soc. Sci. Res., 47, 30–43.

Silm, S.; Ahas, R.; Nuga, M. (2013). *Gender differences in space-time mobility patterns in a post-communist city: A case study based on mobile positioning in the suburbs of Tallinn*. Environ. Plan. B Plannning Des., 40, 814–828

Silva, T. H., Cesário, E., & Pereira, F. C. (2017). *Urban mobility analysis and itinerary reconstruction using call detail records*. Transportation Research Part C: Emerging Technologies, 78, 418-435.

Smith-Clarke, C.; Mashhadi, A.; Capra, L. (2014). *Poverty on the Cheap: Estimating Poverty Maps Using Aggregated Mobile Communication Networks.* In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May; pp. 511–520.

Srinivasan, S., Krause, A., Kakade, S., & Seeger, M. (2018). *Gaussian Process Smoothing of Noisy Time Series: A Case Study of Mobile Phone Positioning Data*. Journal of Machine Learning Research, 18(48), 1-41.

Stopher, Peter & Greaves, Stephen. (2007). *Household travel surveys: Where are we going?* Transportation Research Part A: Policy and Practice. 41. 367-381. 10.1016/j.tra.2006.09.005.

Sun, L., Axhausen, K. W., Lee, D. H., & Huang, X. (2016). *Understanding Metropolitan Patterns of Daily Encounters*. Proceedings of the National Academy of Sciences, 113(10), 2620-2625.

Sun, L., Axhausen, K. W., Lee, D.-H., Huang, X., & Chung, E. (2014). *Understanding travel behavior variability by clustering smart card data: A case study of London underground users.* Journal of Transport Geography, 34, 146-156.

Tamblay, S., Muñoz, J. C., & de Dios Ortúzar, J. (2018). *Extended Methodology for the Estimation of a Zonal Origin-Destination Matrix: A Planning Software Application Based on Smartcard Trip Data.* Transportation Research Record, 2672(8), 859–869.

Tiru, M. (2014). *Overview of the sources and challenges of mobile positioning data for statistics*. In Proceedings of the International Conference on Big Data for Official Statistics, Beijing, China, 28–30 October 2014; pp. 1–26

Tizzoni, M.; Bajardi, P.; Decuyper, A.; King, G.; Schneider, C.; Blondel, V.; Smoreda, Z.; González, M.; Colizza, V. (2014). *On the use of human mobility proxies for modeling epidemics*. PLoS Comput. Biol., 7, e1003716

Tobler, W.R. (1970) *A Computer Movie Simulating Urban Growth in the Detroit Region.* Econ. Geogr., 46, 234–240

Toole, J. L., Colak, S., Sturt, B., & Alexander, L. P. (2015). PathScan: *A mapping-based system for understanding and improving the walkability of urban environments.* In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 477-486).

Toole, J. L., Colak, S., Sturt, B., & Alexander, L. P. (2015). *The Path Most Traveled: Travel Demand Estimation Using Big Data Resources*. Transportation Research Part C: Emerging Technologies, 58, 162-177.

Tseng, Y. H., Kolokolova, A., & Shahabi, C. (2017). *Enabling Public Transit Systems with Big Data: A Case Study of Mobile Ticketing Data for Transit Planning*. IEEE Transactions on Intelligent Transportation Systems, 18(12), 3318-3327.

Versichele, M.; Neutens, T.; Goudeseune, S.; Van Bossche, F.; Van de Weghe, N. (2012). *Mobile Mapping of Sporting Event Spectators Using Bluetooth Sensors, Tour of Flanders Sensors*. Sensors, 12, 14196–14213.

Vogelova, M (2012). *Using Residual Positioning Data from Mobile Networks for Tourism Monitoring.* Czech Touristic Authority—Czech Tourism. Tourism Statistics in the 21st Century. Session Paper in 11th Global Forum on Tourism Statistics.

Wang, Pu, Timothy Hunter, Alexandre M Bayen, Katja Schechtner et Marta C Gonzales (2012). *Understanding road usage patterns in urban areas*. In: Scientific reports 2, p. 1001 (cf. p. 14)

Widhalm, Peter, Yingxiang Yang, Michael Ulm, Shounak Athavale et Marta C Gonzales (2015). *Discovering urban activity patterns in cell phone data*. In: Transportation 42.4, p. 597–623

Wu, Wei, Yue Wabg, Joao Bartolo Gomes et al. (2014). *Oscillation resolution for mobile phone cellular tower data to enable mobility modelling*. In: Mobile Data Management (MDM), IEEE 15th International Conference on. T. 1. IEEE, p. 321–328 (cf. p. 8)

Zhang, X., Zhou, X., & Lin, F. (2017). *Transport mode detection using smartphone sensors*. IEEE Transactions on Intelligent Transportation Systems, 18(11), 2988-2999.

Zheng, Y., Liu, Y., Yuan, J., Xie, X., & Sun, G. (2015). *Learning transportation mode from raw GPS data for geographic applications on the web.* Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1755-1764.

Zhu, Y., Zheng, Y., & Zhou, X. (2014). *Inferring transportation mode and travel purpose from mobile phone data for itinerary reconstruction. ACM Transactions on Intelligent Systems and Technology* (TIST), 5(3), 1-24.

Çolak, Serdar, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta et Marta C Gonzales (2015). *Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities*. In: Transportation research record: Journal of the transportation research board 2526, p. 126–135 (cf. p. 2, 72).e

Ščepanovic, S.; Mishkovski, I.; Hui, P.; Nurminen, J. (2015). *Mobile Phone Call Data as a Regional Socio-Economic Proxy Indicator*. PLoS ONE, 4, e0124160Ewing, R., & Cervero, R. (2010). Travel and the Built Environment: A Meta-Analysis. Journal of the American Planning Association, 76(3), 265-294.

Andreas Schmidt-Dannert. Positioning Technologies and Mechanisms for mobile Devices. (https://www.snet.tu-berlin.de/fileadmin/fg220/courses/SS10/snetproject/positioning-technologies_schmidt-dannert.pdf) Accessed on 1 January 2023.

European Commission. ESSnet Big Data. 2020 (https://ec.europa.eu/eurostat/cros/content/essnet-big-data_en) Accessed on 1 January 2023.

Hellerstein, J. (2008). *The Commoditization of Massive Data Analysis*. O'reilly Radar. (http://radar.oreilly.com/ 2008/11/the-commoditization-of-massive.html) Accessed on 1 January 2023.

Instituto Nacional de Estatística (https://www.ine.pt) Acessed on 1 Janoary 2023.

Real Time Rome. MIT Senseable City Lab. 2006. (http://senseable.mit.edu/realtimerome/) Aaccessed on 1 January 2023.

Stelman, T. Trendit Mapping Population Movements through Mobile Signals. NoCamels. (https: //nocamels.com/2012/03/trendit-mapping-population-movements-through-mobile-signals/) Accessed on 1 January 2023.

Wilipedira.org (https://wikipedia.org/) Accessed on 1 January 2023