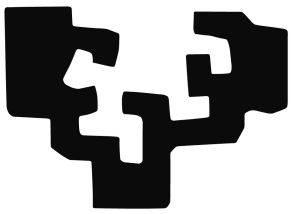


On the use of multi-sensor digital traces to discover spatio-temporal human behavioral patterns

By

Ricardo Luis Muñoz Cancino

PhD Advisors:

Prof. Dr. Manuel Graña Romay

Dr. Sebastián Ríos Pérez

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Donostia - San Sebastián 2023

ON THE USE OF MULTI-SENSOR DIGITAL TRACES TO DISCOVER SPATIO-TEMPORAL HUMAN BEHAVIORAL PATTERNS

by

Ricardo Luis Muñoz Cancino

Submitted to the department of Computer Science and Artificial Intelligence of the UPV/EHU in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

We live in the information era, which presents an unprecedented opportunity for analyzing human behavior. Technology has penetrated our lives in such a way that a large part of our day to day is surrounded by technological devices that record the activities we carry out every day. Whether using the cell phone to call friends, send a text message, post on a social network, or pay through electronic devices, it leaves a digital trace that is stored and allows us to analyze human behavior. A subset of great interest is those digital traces that contain the geographic location because it allows us to understand the interaction between people and the urban infrastructure. The use of digital traces covers multiple disciplines associated with human mobility and its interaction with the city, such as urban planning, infrastructure management, public transportation management, and public policies. This information is also widely used to target responses to unfortunate events such as natural disasters or terrorist attacks and study biological viruses' spread and contagion. This work addresses three gaps in detecting human behavioral patterns using digital traces. The first gap is related to the algorithms used for detection, where we challenge the traditional approach relying on distance-based clustering algorithms like K-Means. The second gap is associated with the pattern validation process, which demands extensive expert knowledge about the geographical areas studied. The third gap is the lack of consideration for the temporal dimension, as classical approaches focus on finding static patterns, but behavioral patterns change over time. Therefore it is necessary to find a proper approach to analyze how human behavioral

patterns change over time. We propose a methodology that adapts latent semantic models to identify human behavioral patterns. In addition, quantitative metrics are proposed to assess the quality of the patterns obtained. Moreover, two methods are presented to study the long-term changes in the detected patterns. The methodology is applied to different types of digital traces, which were grouped into three large datasets: the telecom dataset containing 880 million call detail records; the banking dataset with 85 million geo-tagged credit card purchases; and the social media dataset, a collection of 32 million geo-tagged urban activities like tweets, check-ins, and social media comments. The results show that latent semantic models detect human behavioral patterns and identify new behaviors not observed by distance-based clustering algorithms. Latent Dirichlet Allocation models performed better than traditional models for the static detection problem, while Dynamic Topic Models overperformed in the task of detecting spatiotemporal patterns. Moreover, the proposed metrics allow us to compare human behavioral patterns and thus select the one that best describes the kinds of actions developed by the individuals while interacting with the city. In future work, we want to study if there is a hierarchical relationship between the patterns that can be obtained from the single-cities analysis with other types of spatial aggregation, such as cities or countries. On the other hand, we want to delve into the mathematical properties of the proposed metrics, analyze them and test new scenarios for identifying activity patterns.

Keywords: Human Behavioral Patterns ; Topics Models ; Geo-tagged Data ; Multi Sensor ; Multi Temporal

.

Saludos

Acknowledgements

This work would not have been accomplished without the financial support of CONICYT-PFCHA/DOCTORADO BECAS CHILE/2019-21190345. The second advisor received research funds from the Basque Government as the head of the Grupo de Inteligencia Computacional, Universidad del Pais Vasco, UPV/EHU, from 2007 until 2025. The current code for the grant is IT1689-22. Additionally, he participates in Elkartek projects KK-2022/00051 and KK-2021/00070. The Spanish MCIN has also granted him a research project under code PID2020-116346GB-I00.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This introductory chapter presents the motivation to study land use patterns generated from human mobility. Section 1.1 starts with the research's motivation and introduces the research area addressed in this thesis. Section 1.2 presents the research problem and the thesis's general and specific objectives. Then, in Section 1.3, we show the research methodology to continue with the contributions of this research in Section 1.4, and its results, in the form of publications, are detailed in the Section 1.5. Finally, the structure of the thesis is presented in Section 1.6.

1.1. Motivation

Today we have an unprecedented opportunity to analyze human behavior. The increase in technology penetration means that each person is connected or related to some technological device for a large part of their day. Whether using the cell phone to call friends, sending a text message, posting on a social network, or paying through electronic devices, it leaves a digital trace that is stored, and allows us to analyze human behavior. When individuals interact with these technological devices, which we will call sensors, they store not only the interaction details but also include metadata that further enrich the knowledge that can be extracted. This leads to an unprecedented collection of information. Additionally, most of the activities mentioned above provide individual geo-referencing information when performing this action, allowing us to enhance human behavior analysis using geo-spatial data.

Although crowdsourcing and geo-crowdsourcing were coined more than a decade ago to denote the storage and subsequent analysis of digital traces, using digital traces to analyze

human behavior is relatively new. In many cases, surveys are still used to characterize the behavior. However, this alternative does not allow for making quick decisions nor allows us to examine the user's behavior in detail. Using surveys to analyze the users' interaction with their surroundings requires a high logistical and human deployment to gather the information and therefore carries a high cost. Due to the high cost involved, surveys only allow gathering information from precise strategic points, limiting the granularity of the insights obtained, and also have a time limitation since data will only be available for the specific time window over which the survey lasted. Furthermore, these drawbacks restrain the behavior analysis when the information needs to be collected quickly, such as responding to a catastrophe.

Digital traces have a wide range of uses and provide practical and valuable insights into different research areas such as urban planning and infrastructure management [1–4]. In the event of a natural disaster, this information allows identifying areas to focus the aid efforts and planning the response [5–12]. From a health and environmental perspective, this information allows us to quantify the effect of climate change [13–17], to analyze the response and impact of a terrorist attack [18–20] and to study the spread and contagion of biological viruses like dengue [21], zika [22], ebola [23], HIV [24], and SARS-CoV-2 [25–27]. In transport, digital traces enable the forecast of vehicle traffic [28–33], people's crowd flows [34–37], and public transportation management and public policies [38–40]. Additionally, digital traces have the potential to significantly improve our knowledge of human behavior and help organizations make data-driven decisions. Organizations use this information to improve customer analytics [41–43], marketing decision-making [44, 45], and optimal business facility location [46, 47].

Different types of digital traces record different behaviors. Phone calls allow us to know the social and relational behavior of people. In the case of credit and debit cards, they detail the purchasing behavior and the registered events of social networks, allowing us to know opinions, hobbies, and social life. These digital traces are often associated with a location that also allows us to know how people interact with their environment. This location can be exact, like the one delivered by the GPS on the phone or the point of sale associated with the commerce where the individual bought, as well as being approximate, as is the case of knowing the nearest cellphone tower that processes the call information. In particular, for this study, we are interested in those digital traces that allow us to relate an individual's behavior with their environment. This information is valuable as it allows decisions that impact entire communities, whether at the local level of a neighborhood or even at the level

of an entire country.

In this thesis work, we will address three main gaps identified in the field of identifying human behavioral patterns.

- **Algorithms**: The first gap, related to the algorithms used to detect human behavioral patterns, is addressed by exploring and evaluating alternative algorithms. The traditional process to detect human behavioral patterns is made using distance-based clustering algorithms such as K-Means. K-means is the most used classification algorithm in an endless number of applications. However, there is still room to improve the detected patterns. In fact, when analyzing the patterns obtained by the K-Means algorithm, the question remains whether these algorithms manage to identify all the types of behavior that should be observed.

- **Patterns validation**: The second gap corresponds to how the obtained patterns are validated. Choosing the set of patterns that best describes the behavior observed through the digital traces is fundamental. The success of this task depends heavily on expert knowledge about the study population, the area in which the information was collected, and the behavior that should be observed through the data collection sensor. Inadequate selection and validation of patterns can lead to biased conclusions or errors, so it is crucial to have tools to reduce dependence on domain-specific knowledge to validate the patterns obtained successfully.

- **Time-dependent patterns**: The third gap in the behavioral pattern discovery process is the consideration of the temporal dimension. The classical approaches often focus on finding static patterns in the data and do not account for the long-term evolution of behavioral patterns. It is known that behavioral patterns change over time, so finding ways to incorporate the temporal information in extracting behavioral patterns and thus consider the long-term effects and changes of these patterns over time is essential.

1.2. Research Problem

1.2.1. General Objective

The main objective of this thesis is to extend the general knowledge in the use of digital traces to study the patterns of human activity and its interaction with the environment.

1.2.2. Specific Aims

This research project aims to explore innovative methods for detecting human behavioral patterns, addressing the gaps mentioned above, and considering the temporal and spatial dimensions. The thesis will be structured into a series of three related studies to achieve this goal. Each study will build on the findings of the previous studies, with the overall aim of providing a comprehensive answer to the previously identified gaps. In particular, we introduce the research questions we seek to answer in this study.

1.2.2.1. Aim 1: Formalizing the human behavioral patterns validation

- **Research Questions:** Is it possible to reduce the dependence on expert evaluators to identify human behavioral patterns? What should be the main metrics to evaluate and choose between a set of patterns?
- **Broader Implications:** Formalizing the evaluation and selection of human behavioral patterns will facilitate research on this topic because it is no longer dependent on experts in the geographical areas of study.

1.2.2.2. Aim 2: Modeling human behavioral patterns

- **Research Questions:** Is it possible to detect human behavioral patterns? What is the best way to represent digital traces to detect these patterns? What relationship do these patterns have with the geographic area of data collection?
- **Broader Implications:** Surveys are the most widespread way to learn about people's mobility patterns. A method that allows us to delve into human behavioral patterns through passively collected information will reduce costs and increase the speed and frequency with which insights can be obtained.

1.2.2.3. Aim 3: Temporal and spatial human behavioral patterns

- **Research Questions:** Are the patterns detected stable over time? How can we incorporate the temporal dimension in identifying human behavioral patterns?
- **Broader Implications:** Human behavior changes, as does their interaction with their neighborhoods and cities. For this reason, having a methodology that allows us to observe these changes in behavior is of great help for decision-making in public policies and all other uses given to this information.

1.3. Proposed Methodology

To address the objectives set out in this thesis, we will use a methodology based on the KDD (knowledge discovery and data mining) process [48]. To achieve this, we adapt the methodology to allow insights into human activity patterns and their interaction with the environment. Our methodology is presented in Figure 1.1.

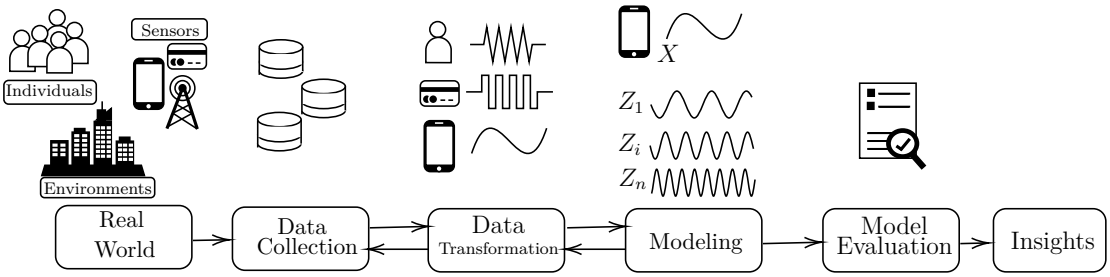


Figure 1.1: Methodology

The methodology begins in the real world, where people carry out their day-to-day activities, leaving digital traces due to their interaction with technological devices or sensors that record and store them. People interact with many sensors during their day; however, for this study, we are interested in those sensors that associate and record the location of the individual at the time of performing an action. Once the data has been collected, it is necessary to transform this information and represent it in a way that facilitates its use by traditional machine learning algorithms. Then, a series of models and techniques are applied to detect behavior patterns from the digital traces generated by each individual. The patterns obtained by using several methods are evaluated in two large dimensions. The temporal dimension allows us to determine the cycle of time in which individuals carry out repetitive

actions and the spatial dimension in which it is analyzed whether these patterns relate to surrounding areas from where the information was gathered.

1.4. Contributions and outline

This thesis aims to address the problem raised above, and within that scope, this work contributes with:

- A review of the state-of-the-art using digital traces to study human interaction with the environment and the city.
- A methodological framework for the evaluation and interpretation of behavioral patterns.
- A novel static method to extract patterns of human behavior and its interaction with the environment.
- An extension of the method mentioned above to incorporate the temporal dimension of behavior patterns and thus analyze their changes over long periods.

The solutions proposed in this study complement and challenge how to use digital traces to identify human behavioral patterns. As will be presented in Chapter 2, traditional approaches focus on detecting static patterns with clustering algorithms such as K-means and also rely on exhaustive knowledge of the geographic area where the digital traces were collected. This work will extend the existing knowledge in this field, enable new studies on the historical analysis of human behavioral patterns, and strengthen the quantitative evaluation of these patterns.

1.5. Publications

The following publications are the direct result of the reported work in this Thesis and they were also developed during the Ph.D. study period:

- Muñoz-Cancino, R., Rios, S. A., Goic, M., & Graña, M. (2021). Non-Intrusive Assessment of COVID-19 Lockdown Follow-Up and Impact Using Credit Card Information: Case Study in Chile. International Journal of Environmental Research and Public Health, 18(11), 5507. [49]

- Muñoz-Cancino, R., Rios, S. A., & Graña, M. (2023). Multi-sensor and multi-temporal city activity patterns using dynamic topics models. Sensors, Under Review.

Articles related to this thesis that were published before starting doctoral studies.

- Sebastián A. Ríos, Ricardo Muñoz, Land Use detection with cell phone data using topic models: Case Santiago, Chile, Computers, Environment and Urban Systems, Volume 61, Part A, 2017, Pages 39-48, ISSN 0198-9715. [50]

Other articles were developed and published during the doctoral studies are indirect results of these doctoral studies.

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the combination of graph data for assessing thin-file borrowers' creditworthiness, Expert Systems with Applications, Volume 213, Part A, 2023, 118809, ISSN 0957-4174 [51]

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña, On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance, Expert Systems with Applications, Volume 218, 2023, 119599, ISSN 0957-4174 [52]

- Ricardo Muñoz-Cancino, Cristián Bravo, Sebastián A. Ríos, Manuel Graña (2022). Assessment of Creditworthiness Models Privacy-Preserving Training with Synthetic Data. In: , et al. Hybrid Artificial Intelligent Systems. HAIS 2022. Lecture Notes in Computer Science(), vol 13469. Springer, Cham. [53]

1.6. Structure of the thesis

This work is structured in chapters, which are detailed below:

- **Chapter 2**: Presents the previous work and delves into the background to understand the area of study and techniques used.

- **Chapter 3**: Details the three data sets used in this thesis: The telecom dataset containing 880 million call detail records; the banking dataset with 85 million geo-tagged credit card purchases; and the social media dataset, a collection of 32 million geo-tagged urban activities like tweets, check-ins, and social media comments.

- **Chapter 4**: Details the methodology proposed in this thesis
- **Chapter 5**: Presents the experimental setup designed to answer this work's research questions and objectives. Additionally, it introduces the results of the experiments designed, which are addressed in depth in **Chapter 6**, **Chapter 7**, and **Chapter 8**
- **Chapter 9**: Presents the conclusions and gives possible lines of research for the future.

Chapter 2

Background and Related Work

This chapter summarizes previous work in the area of study related to this thesis. To do this, we summarize the work done using digital traces to obtain insights into human mobility and its relationship with urban areas. Then, we dedicate a section showing how this information has been used to support decision-making in public policies.

2.1. Digital traces and volunteered geographic information

During the last decades, we have experienced unprecedented technological developments; technology is ubiquitous and is found in everything that surrounds us. People carry multiple technological devices that act as sensors and record different aspects of their lives throughout the day. Today, 91% of people in the world own a cell phone [54]; these devices record our phone call behavior, track our geolocation through GPS [55], and give us access to different social media sites where we can share thoughts, opinions, hobbies, and moments. Even it is possible to interact with other users. Additionally, we carry around credit cards that track what, when, and where we shop, and some people carry smartwatches that track our heart rate and monitor how well we sleep. When we collect and use this information to extract knowledge, we will speak of crowdsourcing [56], and when we restrict the information collected to that which refers to the spatial location of the individual, we will speak of geocrowdsourcing [57]; in both cases, people are understood as a sensor. Goodchild deepens this phenomenon and coins the term Citizens as Sensors [58]. In geography, Volunteered

Geographic Information (VGI) [4] refers to all the information collected from geocrowdsourcing. We are interested in VGI since it does not matter what people, technological devices, or sensors are recording; the combined interpretation of spatial and temporal dimensions allows us to understand and interpret human behavior in a much more comprehensive manner [59].

2.2. On the use of digital traces

Much research has been done on characterizing patterns in urban areas using social crowd-based resources like geo-tagged tweets or cell phone records. Fujisaka et al. [60] discovered regional characteristic patterns from movement histories using aggregation and dispersion models in order to understand the nature of human mobility. Similar work was developed by Wakamiya et al. [61], where they defined the geographic regularity of an urban area using daily crowd activity patterns and analyzing their changes over time. Also, Noulas et al. [62] applying spectral clustering, modeled crowd activity patterns in two cities using geolocated information provided by Foursquare. Crandall et al. [63] performed landmark location using data from geo-tagged photos on Flickr with the mean-shift algorithm. Additionally, Frias-Martinez et al. [64] evaluated the use of geo-located tweets as a complementary source of information for urban planning applications using SOM, Voronoi tessellation and K-means algorithm. Those authors also proposed a technique that automatically determines land uses in urban areas by clustering geographical regions with similar tweeting activity patterns [65].

The human Digital traces and VGI allow the study of human behavior from its interaction with urban infrastructure and urban planning. They also allow transportation management and management of natural disasters and terrorist attacks. The uses are varied and of significant impact; below, we detail the most important uses of human digital traces.

- Urban planning and infrastructure management [1–4].
- In the event of a natural disaster, this information allows identifying areas where to focus the aid efforts and plan the response [5–12]. Moreover, this methodology is also used to quantify the effect of climate change [13–17].
- Analyze the response and impact of a terrorist attack [18–20].
- Management, characterizing, and forecast of vehicle traffic [28–33]

- Public transportation management and public policies [38–40]
- Forecasting people's crowd flows [34–37]
- The spread and contagion of biological viruses like dengue [21], zika [22], ebola [23], HIV [24], and SARS-CoV-2 [25–27]
- Customer analytics [41–43], marketing decision-making [44, 45] , and optimal business facility location [46, 47].

2.2.1. Urban planning and infrastructure management

The population boom of the last century and the fact that most people live in urban areas make it necessary to understand human interaction with the infrastructure of cities to plan the city's development and understand the needs of infrastructure based on public spaces [1, 3]. In addition, the projections indicate that most of the growth will occur in urban cities. These reasons make it even more urgent to understand the complexity of the urban phenomenon and design the city through the construction of codes, public policies, and regulations. Noteworthy studies to understand this phenomenon date back to the 1970s with the work of Pushkarev [66] and Whyte [67], who laid the foundations for recording human activity and its interaction with public spaces, thus leveraging data-driven decisions in urban designers and urban planners.

In its beginnings, this discipline starts from the knowledge of experts, traditionally being data-scarce and going to depend heavily on demographic information [68] or statistics regarding public and private transport use [69]. More modern approaches use records of human activities from agreements with telecommunications companies to extract aggregated information from cell towers [50, 70, 71] or use satellite images [2, 72, 73]. Dembski et al. [4] raises the need for citizens to participate in this mission through digital traces and geo-crowdsourcing data and to promote information collection in cities. For this, they developed a prototype of a digital urban twin in the German city of Herrenberg.

2.2.2. Natural disaster management

The great flow of information generated from digital traces and VGI has become an important data source to characterize, visualize, analyze and predict natural disasters. This information

allows changing natural disaster management strategies to reduce the population's suffering and economic losses. One of the first uses is monitoring dangerous zones and identifying new hazard zones. It is also possible to design mitigation strategies for the effects of a natural disaster, plan assistance teams in disaster zones and manage humanitarian actions, plan relief efforts and contribute to the design of reconstruction plans [7].

Xin et al. [6] used data collected during the 2003 Cyclone Mahasen in Bangladesh to describe the change in human behavior after the natural disaster. Additionally, they use the information collected to analyze population migration patterns and correlate the insights obtained from mobile data with seasonal episodes of migration in Bangladesh. Podesta et al. [8] analyzed and studied the community resilience during Hurricane Harvey, which hit and caused significant damage in the Houston metropolitan area and Southeast Texas in 2017. To do this, they used the records of visits to different points of interest in the city as a proxy to understand and quantify the combined effects of disturbances on lifestyle, infrastructure, and the environment. Another analysis of the effects of Hurricane Harvey is presented by Farahmand et al. [12]. This time, researchers used Mapbox telemetry data for a quick assessment of the flood. In this study, a flood indicator is proposed to quantify the changes in the concentration of human activity. Observing that in the flooded areas, the indicator presents anomalous activity.

A more general approach is presented in Yabe et al. [10]. In this study, the effects of five major natural disasters that occurred in three countries will be analyzed, for which the trajectories of over 1.9 million mobile users will be analyzed. The authors found that despite the affected regions' socio-economic diversity, the recovery patterns after the disaster are similar. Furthermore, using information obtained from household survey data Dargin et al. [11] analyze how people seek information during disasters and study the perception of the reliability of social media platforms during these events. This survey collects information regarding the three major hurricanes that occurred in the United States between 2017 and 2018.

A detailed literature review related to the analysis of natural disaster management and the role of digital traces and VGI is presented by Yu et al. [7]. The study presents the leading big data sources, discoveries, associated achievements in each disaster management stage, and emerging technologies for natural disaster management with Big Data. Additionally, Fan et al. [9] present a general proposal for managing natural disasters based on collecting data

from multiple sensors, integrating data and analytics, and a multi-actor approach based on game theory to support decision-making.

2.2.3. Human digital traces to quantify the effect of climate change

The effects of climate change are already being observed in multiple areas. These changes bring an increase in natural disasters, human migration flows, effects on tourism, and even changes in the natural processes of flora and fauna. Several authors have proposed methods to quantify and understand the effects of climate change. Milojevic et al. [15] present a systematic review of studies based on machine learning to mitigate and understand the effects of climate change. They show the benefits of supporting decision-making based on analytical models and that the implications of the insights obtained from them can be at different scales, whether at the urban, city, building, or household levels. Finally, they propose a framework to optimize urban planning based on machine learning.

The effects of climate change cause people to migrate from environmentally affected areas. Xi et al. [6] studied the long-term and short-term effects on the migratory movement of the population. This study was designed using mobile phone data obtained during Cyclone Mahasen in Bangladesh, and in this way, they managed to characterize migration episodes after natural disasters.

Kubo et al. [14] propose a framework to analyze human welfare under conditions generated by climate change. His study focuses on challenging the calculation of human welfare on the tourist coasts of Japan. The incorporation of different scenarios generated by climate change projects losses in economic value at the national level that are much greater than those used to manage mitigation resources. Due to these findings, they propose changing the current ranking of beaches based on economic value, enabling decision-making under climate change.

Alampi et al. [16] study the possible effects of climate change on the tourist flow who visit rural areas motivated by wine tourism. To do this, they use measurements of precipitation, temperature, and sea level pressure in addition to VGI gathered from the Flickr photo-sharing social media platform. They anticipate demand changes for this type of tourism and a movement of the peak session from summer to spring and suggest that these findings be used to adapt the services supply and the planning of festivals and tours.

Funada et al. [17] study the effects of climate change through changes in the flowering phenomenon. Through a model that detects flowers from street-level photos, the authors

characterize the phenomenon of flowering and thus propose a semi-automatic framework that also reduces the tremendous economic cost of similar studies.

2.2.4. Response and impact of a terrorist attack

Analyzing collective reactions to traumatic events such as terrorist attacks is necessary because it allows us to understand the population's reaction and improve response plans to these regrettable events. Garcia et al. [18] studied the emotional changes during the terrorist attack in Paris in 2015 and thus validate Durkheim's theory on how the collective emotions resulting from these events lead to higher levels of solidarity in the affected population. The authors find an increase in negative emotions in response to the event and a long-term increase in lexical indicators related to solidarity. This study analyzed the digital traces of 62,114 Twitter users during a follow-up period after the terrorist attack in Paris. Another analysis of the terrorist attacks in France using digital traces is featured in Schafer et al. [19], where the authors analyze the millions of tweets stored by the National Library of France after the terrorist attacks. This study analyzes the collection process, the guidelines, and the tools necessary for collecting, storing, and analyzing this information.

Berube et al. [20] developed an unsupervised framework to study the reactions and effects of the 2017 Manchester Arena bombing. This analysis was developed by applying Latent Dirichlet Allocation (LDA) on millions of tweets obtained 24 hours after the event. The findings showed an improvement in social media monitoring compared to the tools used by law enforcement and other government agencies.

2.2.5. Management, characterizing, and forecast of vehicle traffic

Traffic congestion is a problem affecting citizens' lives, especially in large cities. Mathematical models that allow predicting situations and moments of high congestion are used every day to prevent the saturation of streets and highways. Using GPS information gathered from vehicles circulating in Tunisia, Elleuch et al. [30] generated a model based on neural networks to predict the state of congestion in freeways and highways, highlighting the non-linear behavior of traffic in the different types of roads studied.

Ashwini et al. [31] present a benchmark of multiple data sources used for traffic forecast-

ing, and each data source shows the value they add to the resolution of the problem. These data sources include VGI data such as cellphone network data and social media. Within its complete analysis, the added value of each data source is measured based on its precision, reliability, difficulty in obtaining and preprocessing, and infrastructure and maintenance costs.

Salazar et al. [33] propose a framework that allows them to collect geo-referenced Tweets. Using the collaborative volunteered geographic information, they generate a model to predict traffic congestion and a spatio-temporal analysis to characterize and describe the traffic behavior of specific city areas.

Another significant challenge in vehicular traffic management is reducing emissions and negative externalities. A studied alternative is using machine learning models to reduce public transportation emissions [74]. On the other hand, Alam et al. [75] propose a methodology to quantify, estimate and predict vehicle emissions and for this, they use VGI gathered from the GPS of smartphones used by drivers. Furthermore, Krause et al. [76] present a particular case of emission measurement, showing the quantification and forecast of emissions associated with the new German passenger car fleet for 2030.

2.2.6. Public transportation management and public policies

Perola et al. [38] designed a framework for the exploratory and visual analysis of mobility in the Helsinki metropolitan area. This analysis is carried out at the level of postal code areas and shows the mobility network between the different areas. This analysis is done through the aggregation of geotagged Tweets.

Waller et al. [39] designed software for transport demand analysis. This software visualizes roads, creates zones for traffic analysis, and implements a genetic algorithm to estimate origin-destination (OD) demand patterns. The authors propose to use this tool for rapid decision-making for long-term strategic planning. This approach does not directly use VGI but aggregated behavioral information from pervasive traffic data providers such as TomTom and Google.

Graff et al. [40] make available a python library to obtain Twitter data from different domains, which has been collected since December 2015. Although this information has many uses, such as studying natural disasters, health problems, and natural language processing, The most relevant case is the information on the number of trips between more than

200 countries or territories. This information is relevant since it enables the comparison of strategies for managing vehicular traffic and the OD demand forecast.

2.2.7. Forecasting people's crowd flows

The prediction of crowd flows is an important research topic due to the social cost it entails and the negative externalities it causes in citizens' lives. The crowd flow is directly related to the quality of service and infrastructure planning and plays a fundamental role in security and surveillance monitoring. Three primary data sources are identified for the analysis of crowd flows, video analysis, Spatio-temporal analysis, and the use of the VGI gathered from social media. Ebrahimpour et al. [36] present the state-of-the-art and the main data sources, techniques, and algorithms.

From the characterization of Twitter activity in Singapore City, Goh et al. [34] generated a model to predict the crowd flow throughout the city. The model shows that it is possible to have a model to accomplish this objective and complements it by incorporating the tweet tense and sentiment analysis. The authors use a deep-neural-network architecture that improves the prediction accuracy in some scenarios. Terroso et al. [37] use a similar approach to the one proposed by Goh et al. [34]. This similarity is at the level of data and models. The authors also use information from Twitter and combine it with a location-based mobility dataset provided by cellular networks, thereby generating a model to predict the number of trips nationwide. The results show the value of geotagged Twitter VGI as a complement and alternative to mobility attributes based on mobile phone location. The architecture that obtains the best results with Twitter data is based on deep-learning, Long Short-Term Memory (LSTM) models.

Zhao et al. [35] challenge the traditional use of taxis' GPS trajectory data and bike-sharing data taxis because they only present a partial view of crowd movement. To do this, they use cellphone data and convolution neural networks to forecast crowd flows, obtaining better results than an estimate based on time series regression models.

2.2.8. The spread and contagion of biological viruses

Tran et al. [23] analyze how the population shares news, thoughts, and concerns in the face of severe outbreaks. In particular, the authors analyze the diffusion in social media of information related to Ebola. The authors analyze the reactions regarding Ebola by

gathering and estimating the geolocation of around 2 billion Ebola-related tweets during the major Ebola outbreak between August 2014 and December 2014. This study allows us to understand the citizens' reactions to the outbreak and analyzes the spread of topic-based information that can be used for public health crises. In the same way, Masri et al. [22] used geotagged Twitter data to study the ZIKA virus (ZIKV) outbreak that caused severe public health consequences in 2016. The authors used two auto-regressive models to estimate the number of infected cases a week in advance, both for the state of Florida and the entire country (U.S. model). The resulting models showed a high level of reliability in estimating the number of cases and a high spatial correlation when contrasting with the confirmed cases across all 50 U.S. states. These results show the value of using VGI for disease surveillance. Kraemer et al. [77] also used geotagged Twitter data to study the dengue virus spread. In this study, they analyzed the virus incidence in Lahore, Pakistan. They showed that the highest incidence was in high-mobility sectors during the day, which would be explained since transmission is through a day-biting mosquito. After that, the same authors used the same methodology for studying dengue spread to study the spread of COVID-19 in China [26]. This article used geotagged Twitter data gathered before the coronavirus pandemic to establish a mobility baseline. Then, compare the expected mobility and the mobility during the coronavirus pandemic. Abdallah et al. [25] implemented a monitoring system capable of identifying central areas infected or with suspected infected people from the medical records of the new suspects integrated with cellphone records. The study's objective was to track the location's history of the infected people and then, by distance analysis, to identify other people who could have had contact with the infected person.

Another important study for dengue spread was proposed by Ramadona et al. [21]. They use geotagged information from Twitter to study and predict spatiotemporal patterns of disease spread. In particular, the authors study the intra-urban spread in Jakarta City, Indonesia, and for this, they build a mobility index. The mobility index is estimated with a Poisson regression model using lags of up to 6 months. The results show that the mobility index has the highest predictive power for dengue transmission and that geotagged Twitter data helps understand the direction and risk of the spread of dengue.

Brugh et al. [24] propose a method to estimate the risk of HIV and map the spatial heterogeneity of the population. This methodology aims to reduce HIV in adolescent girls and young women. The authors analyzed geotagged household surveys from Eswatini, Haiti,

and Mozambique. Using this data and satellite imagery, they applied models to predict the number and proportion of people at risk of HIV in the studied countries. This information can help planners design prevention programs in high-risk geographic areas and thus maximize the impact on reducing HIV incidence.

During the coronavirus pandemic, people's mobility was restrained and reduced by installing non-pharmaceutical measures to prevent, contain, and reduce contagion. Luca et al. [27] used cell phone records to monitor spatiotemporal patterns of international mobility and thus study changes in the flow of people during the introduction of non-pharmaceutical measures to control the pandemic. To include the effects of these measures, the authors developed a tailored gravity model to quantify the effects of non-pharmaceutical measures.

2.2.9. Customer analytics

Contextual information is an indispensable asset for optimizing and monetizing company information systems' insights. Ferro et al. [47] present an exhaustive literature review that comprehensively characterizes analytics systems based on location and geotagged information. This study depicts 168 articles and examines their contribution from business aspects, data sources, and the knowledge extraction process. A similar analysis and literature review are presented by Pachni et al. [45], where the authors focus the study on various mobile location-based techniques used by businesses and corporations to increase the value delivered to their customers and offer personalized experiences.

Competitive location problems are location models that explicitly incorporate the fact that other facilities already exist in the area or that new competitors will enter that market [78]. Wei et al. [46] present an approach to address the competitive location problem using social media data and customer evaluation and rating. Combining the Huff model and a geographically weighted regression (GWR), the authors assess local customers' sensitivities, testing their approach with the five most renowned retailers in Beijing. Relevant knowledge that can be used as input for the competitive location problem is that provided by Chen et al. In their study, they predict individuals' future location using geotagged social media data. They first use a hierarchical clustering model based on density to identify the most visited areas and then a multi-feature Bayesian model to forecast future spatiotemporal locations. The proposed approach outperforms a state-of-the-art method.

Miah et al. [44] use geotagged VGI gathered from social media to analyze attractions of

interest to tourists and characterize visitation patterns by visitor type. This approach was validated through a focus group using Australian outbound travelers. One of the advantages of this framework is being generic and can be applied in diverse contexts and geographies to provide valuable insights for strategic decisions in tourism companies or by government agencies in charge of promoting and stimulating local tourism. Similarly, Srinon et al[41] propose a mobility framework to improve the decision-making of tourism suppliers such as hotels, restaurants, and tourist attractions. To carry out their analysis, they collected geotagged social media data from the Bangkok metropolitan area, extracting insights into tourist preferences.

Fan et al. [43] present an analysis and characterization of customer clusters based on geotagged VGI collected from Sina Weibo and Dianping.com. The data gathered represents the catering consumer space of the Pearl River Delta region, Guangdong. Using a density-based clustering algorithm (DBSCAN), the authors are able to identify the location and characteristics of consumer clusters.

As it has been observed, using geo-tagged digital traces is very useful for multiple disciplines, where knowing the interaction between individuals and the urban infrastructure is essential. In summary, Table 2.1 shows a comparative analysis between several previously mentioned studies.

2.3. Digital Traces to evaluate the impact of COVID-19

Since the early days of 2020, the COVID-19 pandemic has been shocking [79] the world [80, 81] with several waves of COVID-19 outbreak hitting all countries across the world differently [82], even showing different incidence inside the administrative partitions of the countries [83]. The main tools that have been proposed [84] to control the damage of a viral infection outbreak are related to either non-pharmacological interventions (NPI) impeding the spread of the virus or pharmacological interventions aiming to treat the associated disease severity. Regarding pharmacological developments, many studies are trying to assess the benefits of several ancient and new molecules against the SARS-CoV-2 virus [85–88]. At the same time, immunization approaches are tested at large on the world population [89]. Regarding non-pharmacological interventions [90–92] many countries or their lower administrative divisions

(such as states, regions, or counties) have implemented quarantines and other restrictions of movement, while recommending social distancing, wearing masks and general prophylaxis measures. Concurrently, there is growing apprehension about the side effects of quarantines, curfews, and other restrictions of movement on the general and at-risk population, specifically children and young adults, because they can affect the quality of sleep and physical activity [93, 94], modify the alcohol consumption and eating habits [95], and increase the stress levels [96]. Additionally, the use of urban green spaces has significantly been affected by the pandemic, and it has been valued highly as a resource to overcome the mental burden of the situation [97].

There is emerging literature about the observation of the economic impact of COVID-19 through the lens of consumer spending obtained from credit card information. For instance, credit card digital traces from the second biggest bank in Spain shows a sharp v-shape in the aggregated consumption due to the strict lockdown measures imposed by the Spanish government [98], while Cakmakli et al. [79] used the aggregate information in a predictive epidemiological model of pandemic evolution in Turkey. A detailed study over March-August 2020 in the USA credit card market [99] found a sharp decrease in transactions and balances in mid-March with a slow, incomplete recovery from May onwards. Some economic sectors did experience sharp decreases in activity while others increased the volume of transactions [100]. The study in Akos et al. [99] compares the effect of NPI measures with the psychological pressure of the pandemic roughly measured by the number of cases in the surrounding areas, aka pandemic severity, finding that pandemic severity has a more substantial effect on the diminishing credit card transactions. Pandemic fatigue implies that this effect has weakened over time since the outburst. The effect of pandemic severity on over volume of consumer transactions was also observed in China since the pandemic outbreak in January 2020 [101]. A similar v-shape decrease in the volume of transactions was found in France, where the decrease in credit card transactions started a couple of weeks before the lockdown measures [102]. An additional confirmation of the effects of pandemic severity and mandated NPI on consumer spending is the comparison of the volume of all means of electronic transactions in Sweden and Denmark during the early months of 2020 [103].

However, more research at a microscopic level needs to be done about the impact on consumer behavior of stay-at-home mandates and other NPI measures. A relevant research question is whether the NPI anti-COVID-19 measures have induced behavioral changes in

the population and whether these changes can be observed by analyzing digital traces. This analysis is presented in Section 7.3, and the complete detail can be consulted in our previous research article Muñoz-Cancino et al. [49].

2.4. Data-driven policy-making using digital traces

Recent literature has shown that multiple data sources can be informative in characterizing the interaction between humans and their environments. This research focuses on three geo-tagged digital traces data sources, call detail records, credit and debit card physical purchases, and social media activity. In order to understand their ability to identify novel human behavioral patterns, it is important to distinguish their basic characteristics. On the one hand, transactional credit card data have been available for several years, but the penetration of this payment method has been relatively slow. In many countries, cash is still the most widely used payment method, but the share of cash payments, in terms of volume and value relative to credit and debit card payments, has been decreasing [104]. For the case we analyze in this research, 65% of households use debit cards routinely while 25% use credit cards [105]. On the other hand, mobile phones are a more recent technology but with a much faster adoption. According to [54, 106], even in developing countries, mobile phone ownership spans more than 91% of the population.

Different data sources also present their own strengths and challenges to support policy-making about human behavioral patterns. For example, credit card data is lighter than mobile phone data because it only stores a location when a customer is engaged in a commercial transaction. Simultaneously, call detail records are much more massive because they can create a new location record every time a cell phone connects to an antenna (hundreds of data points per cell phone per day). In this same sense, interaction with social media generates high volumes of information because people are constantly publishing their daily activities.

When analyzing dynamic patterns, credit card data can be less prone to measurement errors and might provide a more reliable signal of each user's relevant economic activities. In this regard, credit card data can provide a robust estimation of the location where citizens live, work, and conduct their frequent activities. However, mobile phone and social media data might be more appropriate to characterize other activities not directly associated with

an economic transaction, such as transportation or outdoor sports. Finally, as transactions from credit card data are associated with specific vendors, it is relatively easy to assign a label to characterize the type of transaction; for mobile phones, this association to an economic area is indirect, requires external information, and presents considerable noise.

Despite the tremendous potential of these new data sources in supporting policy making, there are important methodological challenges to translating massive data sets into valuable insights. An important issue is how we translate the raw data into a few aggregated classes of mobility patterns in the city. For example, this information is the primary source when developing an origin-destination matrix used worldwide to design the public transportation system of a city [107]. Similarly, in the present COVID-19 pandemic, this data can help generate a mobility index to monitor a city's quarantine policies and later evaluate the impact of these policies. In Section 7.3 of this article, we precisely illustrate using a massive digital traces dataset to describe the impact of lock-downs in the context of the COVID-19 pandemic).

Understanding complex human activity patterns using geo-crowdsourced data has been an active research area in the last decade. There are studies using data from multiples resources like Foursquare [62], geo-tagged tweets [60, 61, 64], cell phone records [50, 108–111], geo-tagged Flickr photos [63], and geo-tagged Chinese social media messages [112].

The patterns derived from digital traces are an essential input in decisions associated with public policies. Many authors seek to explain urban phenomena related to public transport [113] and traffic flow [114], where human behavioral patterns are one of the factors that influence them. Hu et al. [115] studied the impact of land use and amenities on the use of public transport. Other scholars propose new approaches to assess Origin-Destination matrices through big data analysis, providing faster, more flexible, and more affordable instruments to adjust public transport policies [107, 116–118]. Besides, human mobility patterns detected from digital traces have been used in many other public policy domains such as flood risk management and urban planning [119, 120]. Previous literature also uses new sensors to describe the relationship between floods and poverty [121], and to study the decrease in agricultural land use utilizing surveys, high-resolution satellite images [122, 123] and deep learning techniques [124]. Finally, human behavioral patterns are used to measure the effectiveness of curfew and lockdown policies during the pandemic caused by the SARS-CoV-2 virus [49, 125].

A detailed description of human behavioral and mobility patterns has proven helpful in supporting policymaking. Despite the extensive research work devoted to this matter, in most cases, these patterns are taken as given, and they are not obtained from observed behavior from the citizens. For example, Lang et al. [126] present a methodology to combine different sources characterizing land use patterns arising from digital traces in several cities in China. They identify three types of cities (economically led, governmentally led, and geographically restricted), but in their research, each cell is described by only one land use type. In contrast, we represent each cell as a mixture over a set of different human behavioral patterns in this study. Our approach can be advantageous in describing human activities in densely populated cities where multiple land uses can coexist in the same cell. Hu et al. [115] developed a framework that can be applied to urban planning for transit-oriented development by studying the impact of land-use features on public transportation in both time and space. They analyze land-use features at two levels: the general land use by sector type and the compositions of services in each zone. They find that more granular land usage information increases predictive power. They conclude that high-resolution data can be more insightful in describing the interdependence of public transportation and land use. The analysis of call detail records, credit and debit card data, and social media urban activities we use in this research provide an even better resolution on the information needed to design urban territories and evaluate land use policies.

Additional interesting applications of human behavioral patterns explaining the interaction between individuals and the city to support public policies is the research of Darabi et al.[120]. They compare various decision tree-based machine learning techniques and genetic algorithms to estimate an urban flood risk mapping based on records and surveys for Sari City, Iran. One of its main conclusions is the role of land use characterization in determining flood hazards. Hong et al. [119] analyzed communication behavior among people during floods using call detail records data. The results revealed higher activity than regular during the crisis, indicating a search for peer support.

Although call detail records and geo-tagged social activities have been used extensively to study individual mobility patterns and land-use research [50, 108–111], the use of credit and debit card transactional data have been scarcely investigated Most of the previous work has been devoted to analyzing human mobility or spending patterns. Brockmann et al. and Gonzalez et al. [127, 128] studied mobility patterns using credit card and cell phone data.

The research showed that Levy Law can approximate the distribution of people's movement. Clemente et al. [129] present a comprehensive analysis to comprehend lifestyles from credit card records; this enables them to have a geospatial characterization of peoples' purchasing and spending patterns. They analyzed the clients' purchase sequences using text mining techniques such as TD– IDF. However, this geospatial characterization is only accomplished by enhancing their analysis with call detail records. Therefore, the geospatial characteristics of its analysis are not directly provided by credit card records. Our approach requires only one data set (credit card records) instead of two massive datasets (credit card records and call detail records) to derive meaningful patterns. Since we are processing data directly from the point of sale, we have the geospatial location of the transaction performed.

More needs to be documented regarding the spatial and temporal stability of human behavioral patterns using digital traces. Lenormand et al. [110] present one of the few studies on multi-sensor stability. They performed a cross-check analysis contrasting three data sources: Twitter, census, and cell phone data. They also examined the correlation between patterns obtained from these data sets in Madrid and Barcelona. However, the call detail records were the only helpful information to characterize land use patterns. Additionally, they extended their research in [130] by applying a functional network approach to detect four major land uses corresponding to different temporal patterns and compared these patterns over multiple cities in Spain.

Tabla 2.1: Main studies on different use cases using digital traces

Objective	Topic	Article	Data	Study Area
Urban planning and infrastructure management	Land use change trajectories	[2]	Remote sensing images	Tianshui, Gansu
	Sustainable Urban Management	[3]	Flickr	Manchester
Management, characterizing, and forecast of vehicle traffic	Traffic Forecast	[33]	Geo-tagged Tweets	Mexico City
	Traffic Forecast	[30]	GPS traces	Tunisia
	Literature review	[31]		
Public transportation management and public policies	Origin-destination demand	[38]	Tweets	Helsinki
	Origin-destination demand	[39]	Travel time data	Sydney
Forecasting people's crowd flows	Crowd flow prediction	[34]	Tweets	Singapore city
	Crowd flow prediction	[35]	Cell phone data	North-east of China
	Literature review	[36]		
	Number of trips prediction - nationwide	[37]	Tweets	Spain
The spread and contagion of biological viruses	Dengue spread	[21]	Tweets	Jakarta, Indonesia
	Dengue spread	[77]	Tweets	Lahore, Pakistan
	Covid Spread	[26]	Tweets	China
	HIV	[24]	Surveys	Eswatini, Haiti and Mozambique
	Ebola	[23]	Tweets	
	ZIKA	[22]	Tweets	Florida, USA
Climate Change	Human migration	[14]	Cell phone Data	Japan
	Human welfare under climate change	[6]	Cell phone data	
	Literature review	[15]		
	Wine Tourism	[16]	E-OBS project and Flickr	
	Cherry blossom flowering	[17]	Geo-tagged images	
Terrorist Attacks	Paris terrorist attacks	[18]	Tweets	Paris
	Paris terrorist attacks	[19]	Tweets and webpages	Paris
	2017 Manchester Arena bombing	[20]	Tweets	Manchester

Chapter 3

Data Description

In this chapter, we detail the data sources used for this thesis. These data sources were grouped according to the sensor type that records the individuals' digital traces. For each source, we explain the information generation process, the volumetry, the data recorded by each sensor, and what was used in this studio.

3.1. Privacy protection and ethical guidelines

All data sources used in this thesis do not compromise any individual's identity or personal information. The telecom and banking datasets were used at antenna and point of sales levels, respectively, so they only have aggregate information at the sensor level and in no case of individuals. On the other hand, the information on online activity was obtained from public datasets, taking all the safeguards for its storage and processing. They were aggregated at the city level, so the user or individual performing the action is not considered for analysis. In addition, any final data produced as a result of this research does not compromise customers' privacy, and there is no possibility that this investigation can leak any personal private information.

3.2. The telecom dataset: Call detail records

3.2.1. Overview

The Base Transceiver Stations (BTSs), also known as antennas or cell tower are responsible for broadcasting and transmitting between the radio network and mobile phones [131]. The

antenna tower and the equipment can identify mobile phones. To manage and improve network efficiency, the BTSs are grouped hierarchically into Location Areas (LA) and are centrally controlled by a Base Station Controller (BSC). The BSC is responsible for the assignment and changes at BTS levels within the same LA. Regarding their activity on the network, mobile phones have two states, active and idle. When a mobile phone is idle, it does not consume any radio resources. However, it constantly evaluates the signal strength and decides whether to switch to another cell to improve this signal strength. A mobile phone spends most of its time idle and only switches to the active state during a phone call or data transaction. Every time a mobile phone becomes active, that is, during a phone call, a data transaction, or sending of SMS, a Call Detail Record (CDR) is generated. The CDRs are generated and collected by telecommunication companies mainly for billing purposes. They are the ones that store the recorded activity of a mobile phone user and allow them to determine the temporal and spatial location through the BTS that manages the user's activity [132].

A CDR provides information on when, from where, to where, with whom, and for how long a user communicates [133]. The CDRs contain the BTS information of the caller and the callee party, a timestamp, and the call duration. Although the structure of the CDRs can change and is not always standardized, the attributes mentioned above allow the spatiotemporal analysis of behavior from mobile cell phone data. Table 3.1 shows, as an example, the information contained in a CDR. Although each record contains several fields, the example contains the necessary data for our study. In addition to the information provided in this table, each BTS has a geographical location described as latitude and longitude.

Tabla 3.1: Sample of a typical Call Detail Record

Caller Phone Number	Callee Phone Number	Caller BTS	Callee BTS	Timestamp	Call duration (seconds)
0012542872	0042478890	BTS00001	BTS00234	2022-04-10 23:15:18	230
0012542872	0056978425	BTS00041	BTS00934	2022-08-17 08:25:13	37
...	...	...	...	...	...
...	...	...	...	...	...
0085967423	0012457784	BTS02477	BTS00065	2022-09-30 09:55:23	108

In particular, the telecom dataset used in this research was originally collected by the major telecommunications company in Chile, the dataset consists of 880 million phone calls recorded over a 77 day period for approximately 3 million anonymized mobile phone users. It contains the information about the phone call: date, time, duration and coordinates

(latitude and longitude) of the BTS (base transceiver station) routing the communication for each phone call. Furthermore, we only know the coordinates of the BTS routing the communication, hence exact location of users are not known within a tower's service area.

3.2.2. Study Area

The study area of this research is the dynamic area of Santiago Metropolitan Region, one of Chile's sixteen administrative divisions. Administratively, this region is divided into 52 communes whose borders are shown in Fig.3.1, covering an area of $15,403.2km^2$ and a density of nearly 470 inhabitants per square kilometer. The total population of the region is approximately 7.3 million inhabitants. This region only covers 2.0% of the country's total surface area but represents approximately 40% of the country's inhabitants.

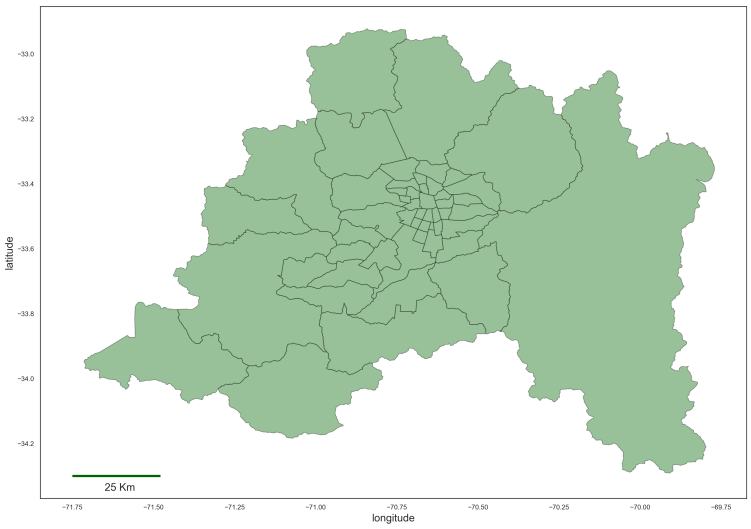
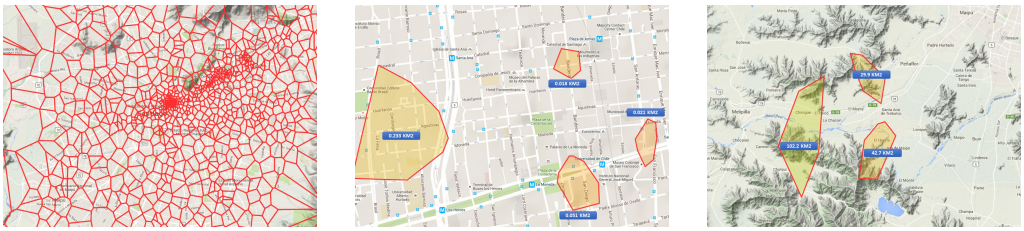


Figure 3.1: Santiago Metropolitan Region

This dataset only provides an approximate geographic location of the caller and the callee because the BTS is in charge of broadcasting and transmitting the call. The location of the BTS is fixed, and each one attends the calls of a respective location area. A Voronoi tessellation is used to define the area covered by each BTS and thus establish a relationship between the call behavior and the geographic area where it is developed. The Voronoi tessellation treats each BTS as the centroid of a region or area called the Voronoi cell. This region is assigned so that any point inside the Voronoi cell is closer to the centroid of that cell than to any other centroid of the other Voronoi cells [134].

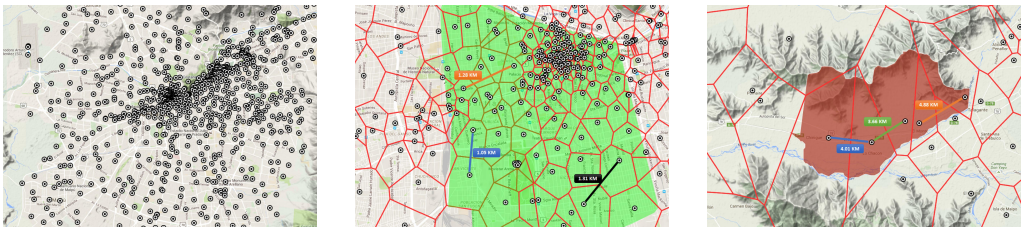
Figure 3.2(a) shows the result of applying a Voronoi tesselation on the Santiago Metropoli-

tan Region using the BTS as centroids. In urban areas (see Fig. 3.2 (b)) each BTS server an area of approximately 0.021 km2 and 74.6 km2 in rural areas (see Fig. 3.2 (c)). There are 1183 BTS towers routing the communication in Santiago (see Fig. 3.3(a)), the distance between BTS's can be a few meters (see Fig. 3.3(b)) in areas up to several kilometers in rural areas (see Fig. 3.3(c)).



(a) Santiago Tessellation (b) Urban Areas (c) Rural Areas

Figure 3.2: Voronoi Tessellation



(a) Santiago BTS Positions (b) Urban Areas (c) Rural Areas

Figure 3.3: Voronoi tessellation showing antennas positions

3.3. The banking dataset: Credit and Debit card records

3.3.1. Overview

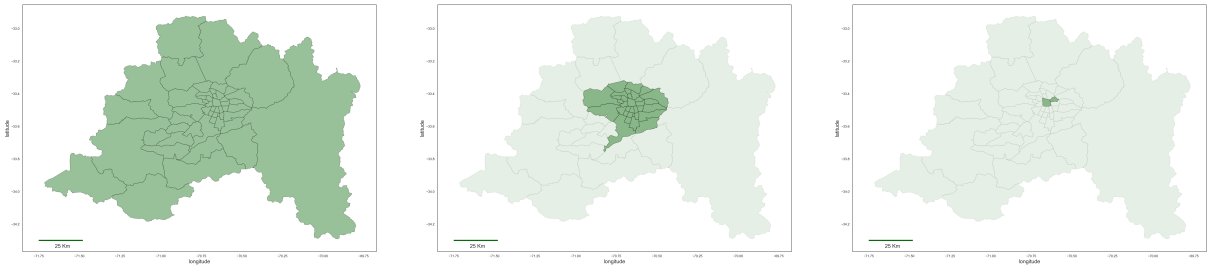
Credit and debit cards are small plastic or metal cards issued to clients of financial institutions to be used as a payment method. Credit cards allow cardholders to borrow funds to pay for goods and services in exchange for a promise of future payment to the institution that issued the card. On the other hand, debit cards are associated with a checking account. When purchasing goods or services, their cost is immediately deducted from the amount available in the cardholder's check account. Both types of cards allow making physical and virtual purchases. When the purchase is made physically, it is done through a device called a Point of Sales (POS) which records the payment and validates the information against the financial institution. Each POS is georeferenced to the business point of sale, so it is possible

to establish the exact location where the purchase was made.

In particular, the banking dataset stores purchasing activity records for customers of a financial institution. Each record in the dataset contains the card type (Credit or Debit), purchase id, latitude and longitude of the POS terminal where the transaction was made, and the day and hour of the transaction. These fields are the minimal information required for this research analysis. Furthermore, the dataset contains the business sector related to the company where the transaction was made. The dataset summarizes transactions made between $2017-01-01$ and $2017-12-31$ and contains 85 million registers associated with more than 80 thousand terminals.

3.3.2. Study Area

As with the telecom dataset presented in Section 3.2.2, the study area will be the Santiago Metropolitan Region. However, due to the particularities of this dataset and the fact that it has numerous geographical locations due to the large number of deployed POS, the analysis incorporates new aggregations for the same region. The Santiago Metropolitan Region can be defined at three geographic levels. Administratively, the city of Santiago is organized into communes whose borders are shown in Fig.3.4(a), with a high degree of autonomy. First, we may consider the Metropolitan Region where the city of Santiago is located. The Metropolitan Region has a population of almost 7 million inhabitants and a total area of $15{,}400 km^2$. Second, we may consider Santiago City (Fig.3.4(b)), a smaller area of the greater Metropolitan Region that excludes rural areas. Finally, we may consider Santiago downtown, where a large fraction of business offices are concentrated, corresponding to Santiago's two most crowded communes. These three areas are illustrated in Figure 3.4.

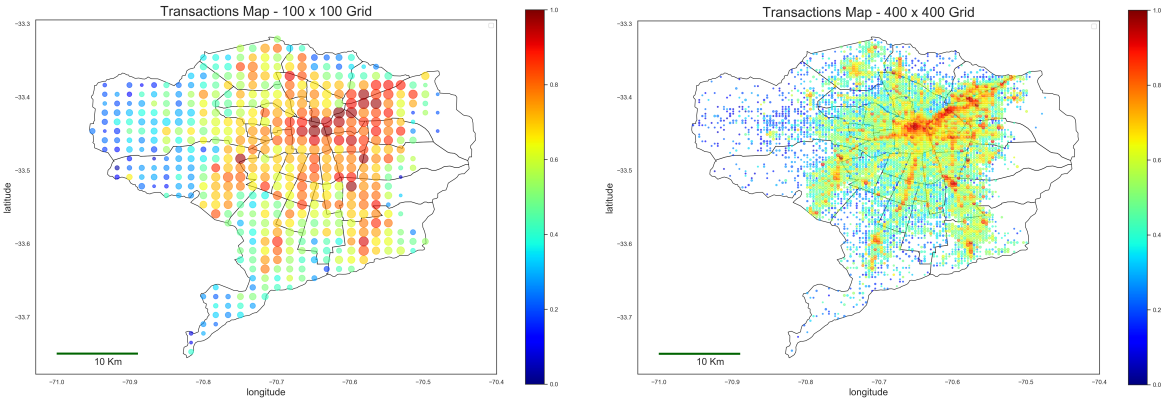


(a) Greater Santiago (b) Santiago City (c) Downtown Santiago

Figure 3.4: Santiago Metropolitan Region

This dataset provides each cardholder's exact location at the purchase time. However,

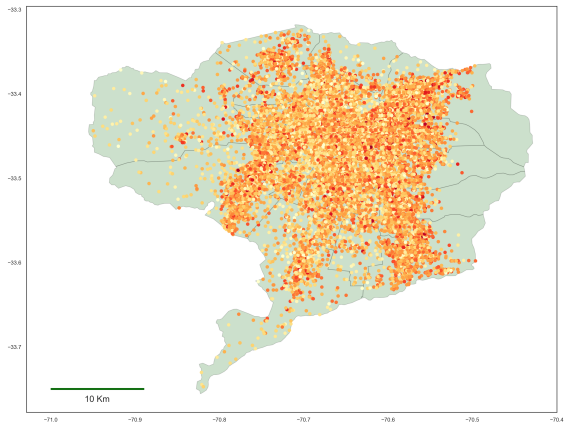
the number of different locations is given by the number of merchants' points of sale. For further analyses, we assume that the purchasing activity of cardholders related to the POS Terminal is similar to other POS terminals in their neighborhood. Therefore we can represent the set of POS Terminals through a spatial aggregation. This study's spatial aggregation results in dividing the city using $n \times n$ uniform grids. In this work, we considered two grid configurations: a 100×100 grid (10,000 cells) and a 400×400 grid (160,000 cells). Figure 3.5 shows the Spatial Aggregation grids used in this research for Santiago city. These figures show the total number of credit and debit card transactions made in each cell on a logarithmic scale.



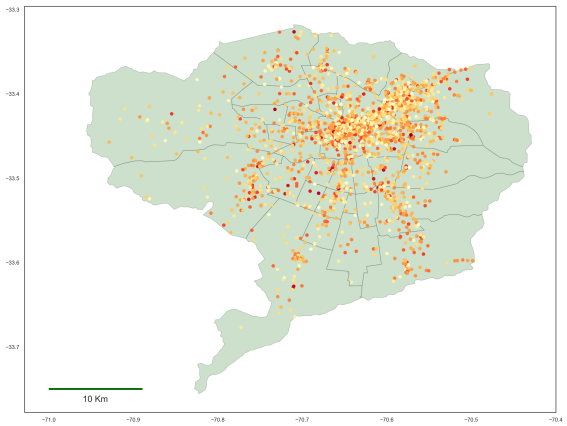
(a) Santiago City 100×100 grid (b) Santiago City 400×400 grid

Figure 3.5: Spatial Aggregation Grids

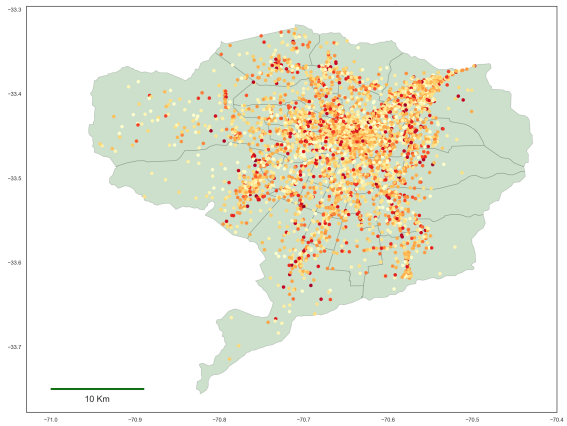
To better illustrate how different economic sectors are geographically distributed in the area under study, in Figure 3.6 we show the number of transactions *per* POST for a selected number of economic sectors in Santiago City. The uneven distribution through different economic sectors in different areas of the city, illustrates the ability of cardholders' digital traces to inform spatiotemporal economic patterns in urban areas.



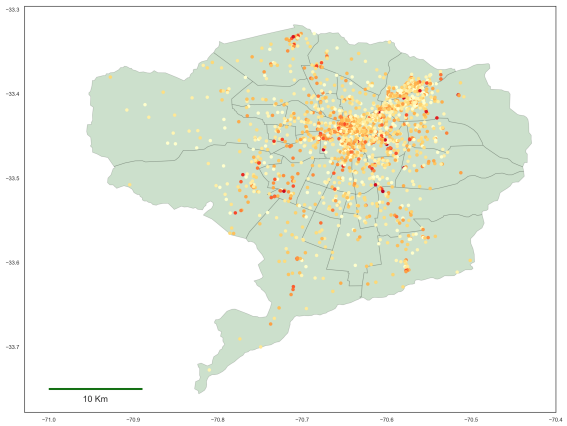
(a) Retail-Food Stores



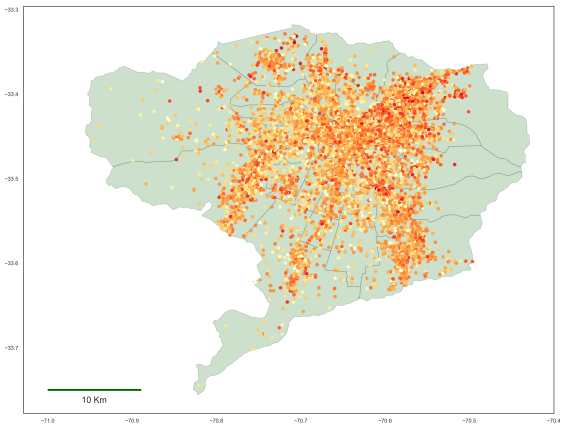
(b) Leisure



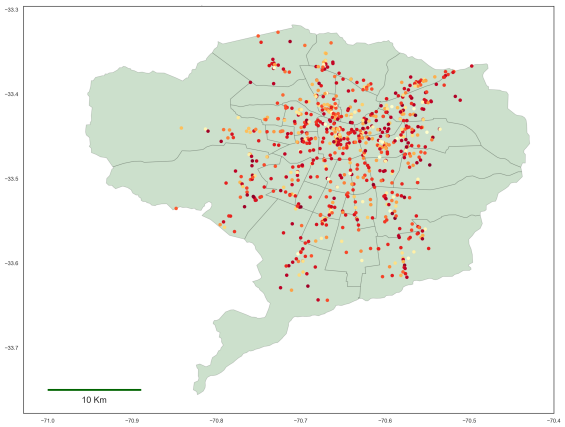
(c) Car Selling



(d) Household Furniture



(e) Restaurants



(f) Gasoline Stations

Figure 3.6: Spatial distribution of credit card transactions *per* POST and economic sector during year 2017 in Santiago City.

3.4. The social media dataset: geo-tagged urban activity records

3.4.1. Overview

The social media dataset contains geo-tagged urban activities. It corresponds to a subset of around 32 million geo-tagged urban activities obtained from multiple social activity platforms such as Twitter, Foursquare, Yelp, Flickr, Gowalla, Brightkite, and Weeplaces. These geo-tagged urban activities were collected during 17 years. Each urban activity is characterized at least by its geolocation (latitude and longitude) and a timestamp. Additionally, and considering that the objective of this study is a city-level analysis, each urban activity was assigned to the closest city as long as it is less than 30 km from the city center. The cities and their respective centers were obtained from [135], and we only consider cities with more than 1MM inhabitants or capitals. In addition, we define a limit of the ten most populated cities in the country in case many cities meet the conditions mentioned above.

Table 3.2 shows the detail of the dataset used in this study. The source of urban activity: tweets, images, and check-ins are detailed, along with the number of events, the number of cities we associate them with, and the period they cover.

Tabla 3.2: Dataset Description

Source	Dataset	Events	#Cities	#Year	Min Year	Max Year
Brightkite checkins	[136]	1,639,399	107	3	2008	2010
Foursquare checkins	[137]	7,515,201	107	6	2010	2015
	[138]	109,9826	3	2	2012	2013
GeoTagged Images	[139]	4,998,865	130	8	2005	2012
GeoTagged Tweets	[140]	2,041,262	10	4	2007	2010
	[141]	187,802	130	2	2020	2021
	[142]	47,337	7	1	2020	2020
	[141] (Exact Location)	184,547	130	1	2020	2020
	[141] (Inferred Location)	2,604,233	136	1	2020	2020
Gowalla checkins	[136]	1,992,082	107	2	2009	2010
Weeplaces checkins	[136]	4,176,673	107	7	2005	2011
Yelp checkins	[143]	5,695,209	2	12	2010	2021

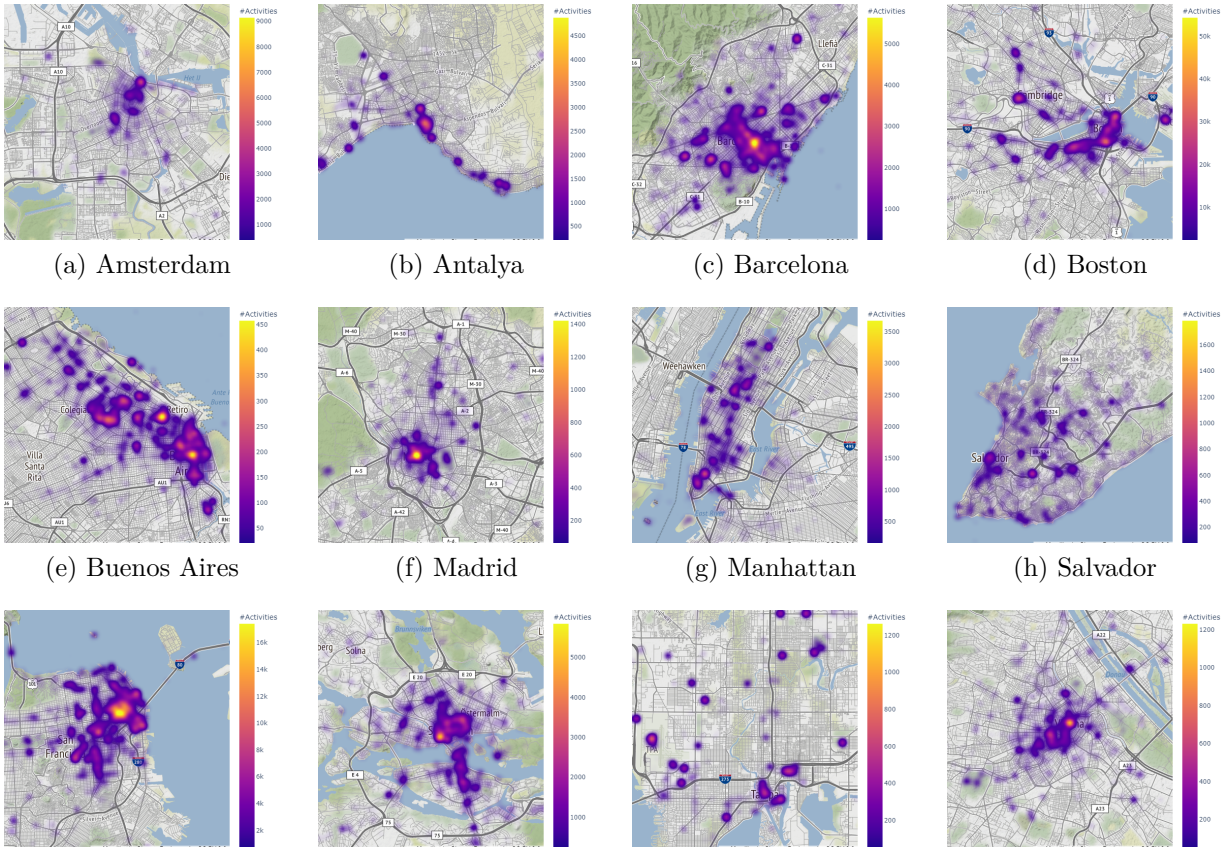
3.4.2. Study Area

The geographical area represented in this dataset corresponds to the world's large cities, for which cities with more than 1MM inhabitants or that are capitals of a country are considered.

Additionally, in order not to represent the behavior of a single country, the number of cities per country is limited to a maximum of ten cities. Figure 3.7 shows the distribution of cities that meet the conditions described above. Additionally, Figure 3.8 shows the detail of geo-tagged urban activities in some cities.

Figure 3.7: Cities considered in the analysis that meet the conditions described

Figure 3.8 shows the density of geotagged urban activities gathered in the abovementioned dataset. For each city shown, only activities within a 30 km radius are considered to describe the city's behavior. In the example, it can be seen that registered activities have a high concentration in the urban centers of each one. For example, in Amsterdam, we observe an area of high activity in the surroundings of the Amsterdam Centraal Railway Station and other sources of high concentration, such as Leidseplein Square, a buzzing nightlife hub surrounded by bars and restaurants. In the case of Manhattan, although they present activity throughout practically the entire island, sectors such as Times Square, the Rockefeller Center, and the One World Trade Center stand out as areas of high activity. Cities that are less touristy than the previous ones, such as Tampa, also show activity in areas of importance to the city, as can be seen in the image: Downtown, The Florida Aquarium, Tampa International Airport.



(a) Amsterdam (b) Antalya (c) Barcelona (d) Boston

(e) Buenos Aires (f) Madrid (g) Manhattan (h) Salvador

(i) San Francisco (j) Stockholm (k) Tampa (l) Vienna

Figure 3.8: Example of cities included in the dataset. The density of geo-tagged digital traces. Yellow indicates a larger activity frequency, while purple indicates a smaller one. Map tiles by Stamen Design under CC BY 3.0, Data by OpenStreetMap contributors under ODbL

Chapter 4

Proposed Methodology

This chapter details the proposed methodology to detect human behavioral patterns from digital traces. As indicated in Figure 1.1. The methodology begins with an allusion to the real world, where individuals interact with the urban infrastructure. Their interactions are recorded by the multiple sensors that people carry with them for much of their day. The real world is analyzed through the lens provided by three digital trace sources, cellphone calls, credit card purchases, and social media urban activities. This kind of information is the most used for analyzing human behavioral patterns. It allows a complete description of the individuals that generate them because the cell phone and the credit card accompany people for a large part of their day.

Regarding the information gathering, Chapter 3 shows the available datasets and the information they contain. These datasets store information about phone calls made with a cell phone, bank transactions made with a credit or debit card, and social media activities made from a smartphone. Each of these activities takes place at the individual level. However, in this study, we will develop the analysis with another level of aggregation. We will use the antenna (BTS) as a record of the call activity, the point of sales (POS) as a record of each cardholder's transactional activity, and the activity in social media will be analyzed at the city level to compare activity worldwide. This way, when we investigate sensor devices, we will also refer to BTS, POS, cities, or a group of them.

In the remainder of this chapter, some definitions are given to understand the generation and storage process of digital traces (See Section 4.1). Then, the data transformation is explained, how to represent the digital traces (See Section 4.2), how they are transformed,

and the concept of spatial aggregation (See Section 4.3). Section 4.4 presents the algorithms used to identify human activity patterns. The proposed model to detect static patterns is presented in 4.4.2, and the model to study spatiotemporal activity patterns is shown in Section 4.4.3. Additionally, the traditional algorithms to detect human activity patterns are presented. They will then operate as a comparison benchmark against the proposed algorithms. Finally, Section 4.6 presents a series of metrics that we propose to validate the patterns obtained and thus reduce the dependence on expert knowledge to perform this task.

4.1. Definitions

A *Base Transceiver Station* (BTS) is a component in mobile telecommunications networks whose primary function is to transmit and emit radio signals between the telecommunications network and mobile cell phones. The function of the BTS is to provide coverage to a specific geographic area, and multiple BTSs are used to cover large geographic areas, therefore building the whole telecommunication network. BTSs play a crucial role in generating call detail records (CDR) that contain information about a call's origin, its recipients, and duration. For this study, we are only interested in knowing the BTS geographical location (latitude and longitude) and the date and time of the calls that are processed by it.

A *Point of Sale Terminal* (POS Terminal) is an electronic device used to process card payments at business locations. Each time a customer pays with a credit or debit card, the transaction is processed by the POS Terminal, and it is geographically tagged with the latitude and longitude of this POS Terminal (business location's latitude and longitude).

For this work, we will refer to a *Sensor Device* as any device that allows the recording and storage of geo-referenced activity.

4.2. Dataset representation

An *Activity Pattern* (AP) is a vector characterizing the activity of each sensor device s over a specified period. In this study, we will consider that each AP describes the weekly activity of the sensor device. We will consider that each AP describes the weekly activity of the sensor device. Formally speaking, we describe a raw Activity Pattern by a vector $\mathbf{XP}^s$, where each component XP_t^s (Activity Block) denotes the number of geo-tagged events or digital traces

carried out on sensor s during period t. Then, the raw activity pattern of the sensor s is given by the vector $\mathbf{XP}^s = (XP_1^s, XP_2^s, \ldots, XP_T^s)(t \in T)$, where T is the set of activity blocks. In this study, we set T into hourly periods during a week; therefore, each raw activity pattern XP^s is a vector with $\mathbf{card}(T) = 168$ components (24 hours, seven days). XP^s

XP^s records the amount of events in the sensor s, this complicates a comparison between different sensors. To facilitate the comparison between sensors, we define the normalized activity pattern $\mathbf{AP}^s$, where $AP_t^s = \frac{XP_t^s}{\sum_{t \in T} XP_t^s}$, i.e., dividing its components by the total number of events. Therefore we can interpret AP_t^s as the percentage of digital traces of the sensor s processed in hourly time window t.

4.3. Data transformation and spatial aggregation

This work aims to understand human behavior through digital traces but links it with the environment where it happens. To analyze the geographical environment, we group the information so that each of these aggregations reflects the behavior of a specific sector or area. For each of the study datasets, the level of aggregation is different. In the case of the telecom dataset, the information is already grouped when it is analyzed at the BTS level because each one records all the calls from the location area; in this case, it only remains to identify the location area. This identification is made through the Voronoi tesselation, as presented in section 3.2.2. In the case of the banking dataset, the situation is different, the number of POS is much greater than the number of BTS, and each POS records the behavior of a point of sale and not of an entire location area. For this dataset, it is assumed that nearby POS have similar behaviors, and therefore the behavior of several POS are grouped. This grouping is done considering the grid presented in Section 3.3.2. Finally, the social media dataset contains geo-tagged social media events from all over the world, and it is for them that the level of aggregation for this dataset is through a grouping at the city level.

Under this spatial aggregation scheme, we define $\mathbf{XP}_s$ as the raw activity pattern for the aggregated area a. The aggregation area a depends on the dataset we are using. For the telecom dataset, $a \in Voronoi(BTS)$ corresponds to a cell of the Voronoi Tesselation whose centroid is a BTS. For the banking dataset, the aggregation area a corresponds to a cell $a = (i, j)$ in the $n \times n$ grid defined in Section 3.3.2 ($i = 1, ..., n$ and $j = 1, ..., n$). For the social media dataset, the aggregation area corresponds to a city $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of

cities that meet the conditions defined in Section 3.4.2. $\mathbf{XP}_a$ is calculated as follow:

$$\mathbf{XP}_a = \frac{1}{\mathbf{card}(H_a)} \sum_{h \in H_a} \mathbf{XP}^h \tag{4.1}$$

where H_a is the set of sensor inside the aggregated area a. Each aggregated area a is characterized by the total number of transactions made in the aggregated area a and the total number of sensor in the aggregated area, denoted $\mathbf{card}(H_a)$. In the same way, $\mathbf{AP}_a$ denotes the normalized activity pattern in the aggregated area a:

$$\mathbf{AP}_a = \frac{1}{\mathbf{card}(H_a)} \sum_{h \in H_a} \mathbf{AP}^h \tag{4.2}$$

From now on, we will use $\mathbf{AP}_a$ and $\mathbf{AP}_s$ interchangeably to refer to the activity pattern of the aggregated area a associated with all the activity recorded by sensor s.

In our study we set every AP related to the aggregated area a as the number of digital traces or geo-tagged events that are managed by the sensor devices in that aggregated area a every hour in a seven-day week. Therefore, each AP is a vector with $N = 168$ components (24 Activity Blocks per day, seven days per week) , where every component reveals the activity of a during one hour. This is achieved by holding the proportion of digital traces during that hour compared with the amount of digital traces gathered in the whole week.

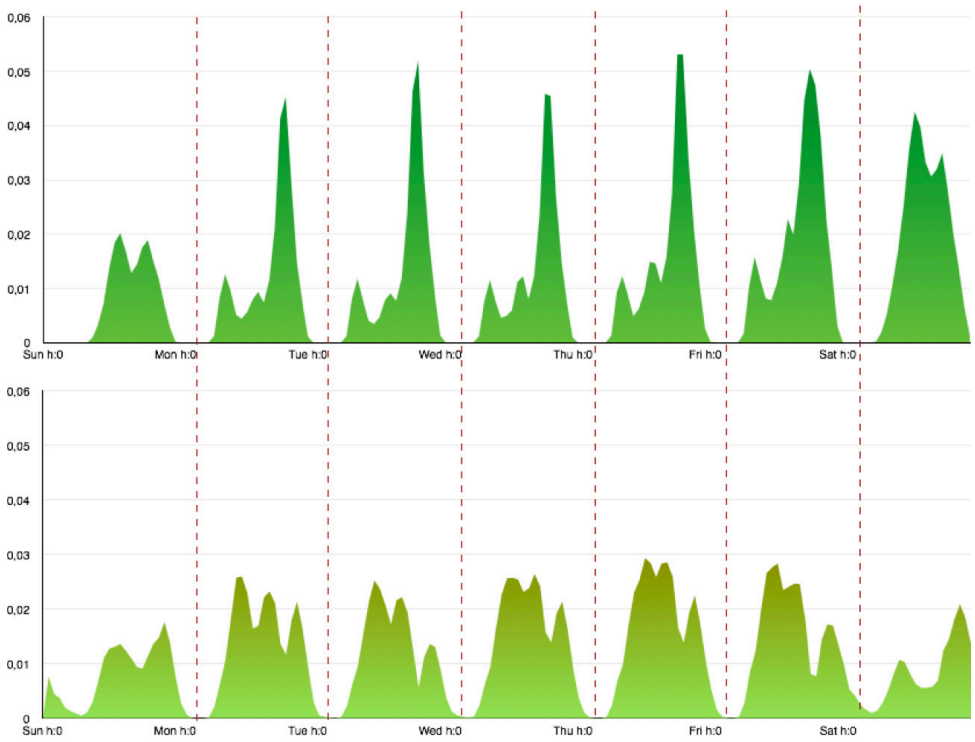


Figure 4.1: Activity Patterns example using the telecom dataset

Figure 4.1 illustrates two AP related to different BTSs in the telecom dataset. Every AP starts on Sunday and finishes on Saturdays.

4.3.1. Determination of activity pattern duration

Before, we declared that each activity pattern is constructed to reflect the behavior of a seven-day week. The rationale behind this assumption is that human behavior patterns recorded through digital traces have a cyclical pattern and that this behavior pattern emerges anew every week. This decision is corroborated below through an analysis of the seasonal component of the activity pattern. Figure 4.2 displays the time series of the number of digital traces (upper plot) gathered in the banking dataset, that is, the number of credit and debit card transactions along with a seasonal decomposition using an additive model from the python library statsmodels [144].

The trend is obtained by applying a convolutional filter that implements a moving average, and the seasonal component corresponds to the average for each period of the de-trended series. The trend component shows an increment at the end of each month, explained because a significant fraction of Chilean citizens is paid precisely at the end of the month. The trend

also shows a moderate increment by the end of the year, associated with Christmas shopping. The seasonal component has a strong weekly frequency. Based on this weekly regularity, we define *AP*s to have a week duration. In other words, one week is our time frame to describe the activity at each aggregated area and to extract representative patterns (topics) of the distribution of digital traces for a typical year.

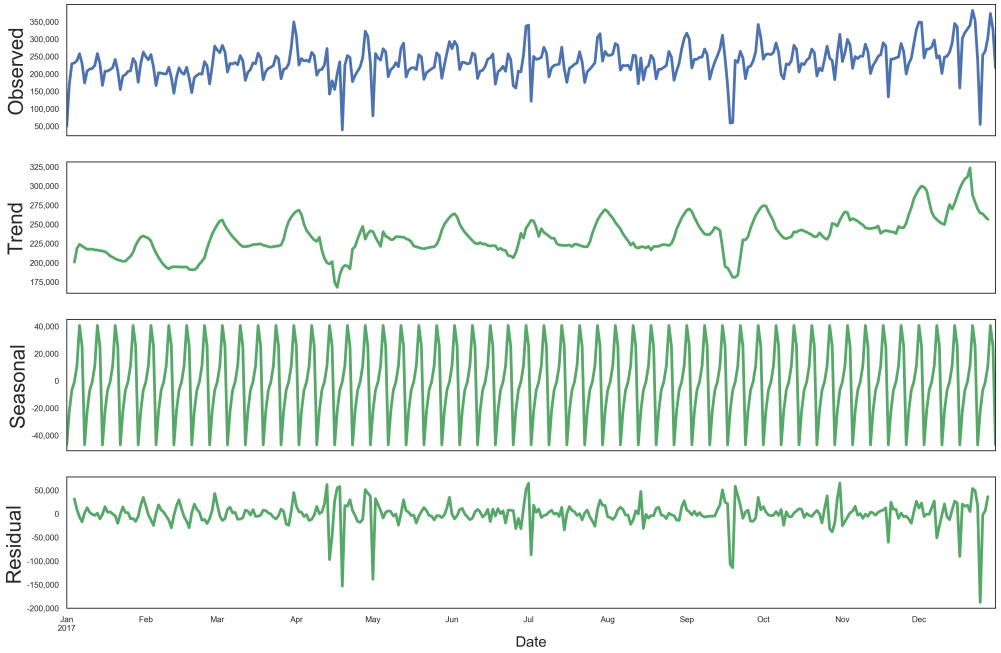


Figure 4.2: Decomposition of the aggregated banking dataset time series into trend, seasonal,and residual components.

4.4. Proposed Models

4.4.1. Topic Modeling overview

Topic modeling is a probabilistic method for uncovering a set of underlying topics in a collection of documents. It originates in natural language processing and has become a popular tool for analyzing and understanding extensive collections of unstructured text. The most widely used technique for topic modeling is Latent Dirichlet Allocation (LDA) [145], which involves identifying a predetermined number of latent topics and assuming each document in the collection can be represented as a combination of these topics. LDA uses a generative process to detect latent topics by considering each one as a distribution of terms.

While LDA has been successful in many applications, it does not consider temporal aspects

that may be present in many document collections. Dynamic Topic Models (DTM) [146] address this limitation by allowing latent topics to evolve over time. DTM extends the idea of LDA by partitioning the collection of documents into defined periods, known as time-slices. The latent topics are modeled within each time-slice, and it is assumed that each latent topic evolves into its corresponding topic in the next time-slice.

4.4.2. Static Topic Modeling using geo-tagged digital traces

A topic model is a probabilistic approach for discovering underlying *topics* that occurs in a collection of documents. The basic premise is that the words that generate the documents are related to latent topics. The most common technique for topic modeling is the Latent Dirichlet Allocation (LDA) [147, 148]. In this work, this concept has been adapted to characterize digital traces and geo-tagged activities in the city using geographically tagged data, cellphone call, credit and debit transaccions and social media activity. Essentially, we will assume that any activity pattern AP_s in a sensor device is drawn from a linear combination of K Human Behavioral Patterns. Each Human Behavioral Pattern $k \in K$ is defined as a distribution $beta_k$ over a set of fixed *Activity Blocks*. Thus, each activity pattern AP_s will have its own mixing proportion of topics θ_s (Human Behavioral Patterns) .

In this context, we represent an *Activity Pattern* as a mixture of topics drawn from a probability distribution $Z_{s,a}$ that can produce the *Activity Blocks* in a *activity pattern* given these topics. The join distribution of a Human Behavioral Pattern mixture θ, a set of Human Behaviors z, and a set of S *activity blocks* $\mathbf{a}$ can be obtained by:

$$p(\theta, z, \mathbf{a}|\alpha, \beta) = p(\theta|\alpha) \prod_{s=1}^{S} p(z_s|\theta) p(a^s|z_s, \beta) \tag{4.3}$$

The main objective of this model is to learn Human Behavioral Patterns from digital traces data distribution by inferring latent topics. The joint posterior probability $p(\theta, z, \mathbf{a}|\alpha, \beta)$ is compose of θ, the distribution of Human Behaviors, one for each Activity Pattern; z Human Behavioral patterns (K) for each sensor device and $\mathbf{a}$ the distribution of Activity Blocks, one for each Human Behavioral Pattern. The parameters β and α are corpus-level hyper-parameters that are assumed to be sampled once. In this context, α is parameter vector for each Activity Pattern (Activity Patterns and Human Behavioral Patterns distribution). With higher α, activity patterns are built from more Human Behavioral patterns, and with

lower α, activity patterns contain less Human Behavioral patterns. β is a parameter vector for each Human Behavioral Pattern (Human Behavioral Patterns - Activity Blocks distribution), with a higher β, Human Behavioral patterns are made up of most activity blocks, and with a low β, they consist of few activity blocks.

A formal description of the data generating process for the LDA model:

1. For each Human Behavioral Pattern $k \in [1, K]$,

   a) Draw a distribution over Activity Blocks $\vec{\beta_k} \sim Dir_K(\eta)$

2. For each sensor device $s \in [1, S]$,

   a) Draw a proportion vector of Human Behavioral Pattern $\vec{\theta_S} \sim Dir_S(\vec{\alpha})$

   b) For each Activity Block $a \in [1, A_p]$ in the sensor device s,

      i. Draw a Human Behavioral Pattern assignment $Z_{s,a} \sim Mult(\vec{\theta_b}), Z_{s,a} \in \{1, \ldots, K\}$

      ii. Draw an Activity Block $W_{s,a} \sim Mult(\vec{\beta_{Z_{s,a}}}), W_{s,a} \in \{1, \ldots, S\}$

Where $Dir_S(\vec{\alpha})$ denote a S-dimensional Dirichlet with vector parameter $\vec{\alpha}$ and $Dir_K(\eta)$ denote a K dimensional symmetric Dirichlet with scalar parameter η.

4.4.3. Dynamic Topic Modeling using geo-tagged digital traces

Topic modeling, LDA [145] and DTM [146] consider that each human behavioral pattern can be represented as a linear combination of k latent topics. We write K_t to denote the set of k latent human behavioral topics that describe the individuals' activities during the time-slice t, and we use $\mathcal{K}$ to denote the set of latent topics throughout the study period, i.e, $\mathcal{K} = \bigcup_{t \in \mathcal{T}} K_t$. Therefore AP_t^s can be expresed as $AP_t^s = \sum_{k=0}^{|K_t|-1} \theta_{t,k}^s AT_{t,k}$ where $AT_{t,k}$ correspond to the k-th activity topic of the sensor device s at time-slice t.

Thus, each human behavioral pattern AP_t^s is described by mixing of activity topics $\theta_{t,k}^s$, also known in topic modeling as per-document topic distribution. In this study, the per-sensor topic distribution $\theta_{t,k}^s$ is modeled using a logit-normal distribution with mean α to represent uncertainty over proportions, $\alpha_t | \alpha_{t-1} \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$ where t and $t-1$ are two adjacent time-slices. Additionally, activity blocks are the equivalent in our problem to the words in the document processing applications, the activity block distribution of activity topics at time-slice t follows a logit-normal distribution $AT_{t,k} | AT_{t-1,k} \sim \mathcal{N}(AT_{t-1,k}, \sigma^2 I)$.

The human behavioral topic distribution η_t^s for sensor device s at time-slice t is modeled as $\eta_t^s \sim N(\alpha_t, a^2 I)$. Finally, human behavioral topics and activity blocks are drawn from multinomial distributions. Similar topic models adaptations for detecting activity patterns can be found in our previous research [49, 50].

4.5. Traditional algorithms to detect human activity patterns

4.5.1. K-means

k-means is an unsupervised algorithm that aims to partition a set of observations into k clusters. k-means assigns each observation to a single cluster in such a way that the observation is assigned to the closest centroid. Mathematically, given a set of n observations $x_1, \ldots, x_n$ in $\mathcal{R}^p$, k-means partitions the n observations into k data sets (where $k <= n$), where $\mathbf{S} = S_1, \ldots, S_k$ is the resulting partition [149]. To obtain $\mathbf{S}$, k-means minimizes the within-cluster sum of squares (also known as inertia or variace criterion).

$$\underset{\mathbf{S}}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{4.4}$$

Where μ_i denotes the centroid of the observations belonging to S_i.

4.5.2. k-Shape

k-Shape [150] is an algorithm for time-series clustering, whose foundation is an iterative improvement process that creates homogeneous and well-separated clusters. k-Shape uses a normalized version of cross-correlation as a distance metric, unlike K-means, which generally uses Euclidean distance. Additionally, it proposes a method based on matrix decomposition to choose the center of the clusters. This method preserves the shape of the time series.

4.5.3. Time series K-means

Time series K-means [151] is an algorithm for clustering time series data. This algorithm proposes a new metric to guide the clustering process. It also generates an iterative process for the cluster search by extracting latent smooth subspaces. The smooth subspaces corre-

spond to weighted time stamps that represent the relative importance of each time stamp in determining the cluster associated with each time series object. TSKmeans solves an optimization problem to assign each time series object to a cluster.

4.6. Human behavioral patterns validation

There are various approaches to characterizing a individuals behavior based on urban activities, and as a result, there are multiple ways to evaluate the quality of the resulting topics. However, common points allow us to establish specific guidelines on the quality of the topics discovered. One common method is to assess the consistency of identified patterns or clusters using metrics such as [152], and other similar metrics. When the discovered topics present a spatial component, geospatial metrics such as pattern distribution and coverage over the study area are used to evaluate the results. Additionally, one of the most used evaluation method relies on the expertise and deep knowledge of the researcher in the area being investigated. This approach allows for a more subjective evaluation of the topics and their relevance to the specific context of the study. However, it can be challenging for researchers to obtain this knowledge when the study areas are extensive or unfamiliar to them. In such cases, alternative evaluation methods may need to be considered, such as consulting with local experts or using additional data sources to gain a better understanding of the area.

To ensure that the human behavioral pattern evaluation method is consistent and reliable, we have identified a set of desired properties for temporal human behavioral patterns and the corresponding metrics that can be used to evaluate the results. These properties and metrics provide a standardized framework for assessing the quality of the patterns discovered and allow us to make meaningful comparisons between different approaches and methods. Some of the desired properties of temporal human behavioral patterns that we propose to consider include:

- Intratemporal Similarity: One of the main expected results is that the topics or patterns describe different activities carried out in a city. In this way, we expect that human behavioral topics be as dissimilar as possible between them. Formally speaking, given a sets of temporal human behavioral patterns $\mathcal{K}$, a time-slice partition $\mathcal{S}$ and K temporal

activity topics the Intratemporal Similarity takes the form:

$$IntraSim(\mathcal{K}) = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}}\frac{2}{K(K-1)}\sum_{k=0}^{K-1}\sum_{l=0}^{K-1}\mathbb{1}_{k\neq l}\cdot sim(AT_{s,k}, AT_{s,l}) \tag{4.5}$$

Where $sim(x, y)$ corresponds to a similarity function between two activity patterns, the objective is to find the set of activity topics that minimizes the Intratemporal Similarity. The minimization problem becomes a maximization problem when the function used to compare activity topic is a distance function.

- Intertemporal Stability: When the human behaviour is analyzed over time, it is expected that the behavior measured from urban activities will not change overnight. We will measure these gradual changes in individuals behavior from the changes between the same human behavioral pattern in adjacent time-slices. Formally speaking, given a sets of temporal human behavioral patterns $\mathcal{K}$, a time-slice partition $\mathcal{S}$ and K temporal human behavioral patterns the Intertemporal Stability takes the form:

$$InterSta(\mathcal{K}) = \frac{1}{K}\sum_{k=0}^{K-1}\frac{1}{2|\mathcal{S}|-2}\sum_{s=0}^{|\mathcal{S}|-1}\sum_{u=0}^{|\mathcal{S}|-1}\mathbb{1}_{|s-u|=1}\cdot sim(AT_{s,k}, AT_{u,k}) \tag{4.6}$$

Where $sim(x, y)$ corresponds to a similarity function between two activity patterns, the objective is to find the set of human behavioral patterns that maximise the Intertemporal Stability. The minimization problem becomes a minimization problem when the function used to compare human behavioral patterns is a distance function.

- Topic Consistency: The human behavioral is based on their daily routines, which is why there is a certain regularity in the activities carried out on working days and during the weekend. This consistency can be observed in the empirical results obtained in multiple studies. In this way, we propose the following metric to study the topic coherence of weekly human behavioral patterns $AT_{s,k}$:

$$TC(AT_{s,k}) = \frac{5}{7}\sum_{i,j\in weekdays}\mathbb{1}_{i\neq j}\cdot sim(AT_{s,k,i}, AT_{s,k,j}) + \frac{2}{7}\sum_{i,j\in weekend}\mathbb{1}_{i\neq j}\cdot sim(AT_{s,k,i}, AT_{s,k,j}) \tag{4.7}$$

Where $AT_{s,k,i}$ correspond to components of day i in the human behavioral pattern $AT_{s,k}$, i.e, the first 24 components of $AT_{s,k}$ represents a Monday ($i = 1$), the next 24 components a Tuesday ($i = 2$) and so on.

- Topic Smoothness: Human behavior while interacting with the urban infrastructure is carried out continuously and without significant restrictions that suddenly limit all activity. For this reason, patterns with smooth changes are preferred to those with high volatility during the day. We propose a simple and effective method to measure the smoothness of an behavioral pattern.

$$TS(AT_{s,k}) = \sqrt{\frac{1}{T-2}\sum_{i=1}^{T-1}(d_{s,k,i} - \bar{d}_{s,k})^2} \tag{4.8}$$

Where $AT_{s,k} \in \mathcal{R}^T$, the difference vector $d_{s,k,i} = AT_{s,k,i+1} - AT_{s,k,i},\ i = 1, \cdots, T-1$ and $\bar{d}_{s,k} = \frac{1}{T-1}\sum_{i=1}^{T-1} d_{s,k,i}$

By considering these desired properties and metrics, we can ensure that our evaluation method is comprehensive and rigorous, allowing us to accurately assess the quality of the temporal human behavioral patterns that we discover.

Chapter 5

Experimental Setup and Results

In this chapter, we detail the proposed experimental setup to answer the research questions presented in Section 1.2. We detail each experiment's objective, expected results, and relationship with the research questions. Subsequently, we present an introduction to the results obtained in this research. Here, we summarize the structure of each study and provide an overview of the following three chapters, which contain the detailed results of each of the experiments carried out in this investigation. In each of these chapters, the approaches used in the corresponding experiment are described, as the results obtained and their interpretation. In addition, the implications of the results are discussed to respond to the objectives of this thesis, and possible areas of future research are proposed.

5.1. Experimental Setup

In order to answer the research questions, three studies are proposed, each based on the previous results, to provide a complete and rigorous assessment of the proposed methodology. The experimental setup for each of the three proposed experiments is presented below in detail.

5.1.1. Experiment 1: Spatial human behavioral patterns

The first study evaluates the proposed methodology's feasibility and effectiveness in detecting human behavioral patterns using new algorithms as an alternative to K-means. To achieve this, we designed an experiment using the proposed methodology in Chapter 4 to detect human behavioral patterns using the telecom dataset. This study proposes an alternative

to K-means to detect static activity patterns. To this end, the results of applying a latent semantics model, Latent Dirichlet Allocation (LDA), to detect human behavioral patterns are analyzed. Additionally, the concept of pattern stability is introduced. For this, the telecom dataset is divided into two subsets, divided in such a way that the first data set contains the first 38 days of the dataset, and the second dataset contains the following 38 days. There is no temporal overlap between the digital traces of these two subsets. The LDA model is applied to obtain between $k = 2$ and $k = 8$ patterns. The topics that best represent the underlying human behavior are selected based on extensive knowledge of the region where the digital traces were gathered. Afterward, the patterns obtained are analyzed and interpreted based on their relationship with the geographic area of data collection. This experiment responds to Aim 2 of this thesis, which seeks to discover if it is possible to detect patterns of human behavioral patterns, the best way to represent them, and their relationship with the geographical area of study.

5.1.2. Experiment 2: Spatiotemporal human behavioral patterns: Multiple static models

In this experiment, the concept of spatiotemporal patterns is introduced. Spatiotemporal patterns not only describe human activity in a limited time window, that is, the description of cyclical human activity based on the weekly routine of people, but also seek to explain how this weekly pattern of activity changes over time when there are data collected at different times or moments in time. This experiment builds on the previous results and proposes a spatiotemporal analysis based on extracting human activity patterns at different moments by training a spatial model with data collected at different moments. This experiment uses the banking dataset, and the study begins by performing a static analysis. The optimal number of human behavioral patterns is determined by training the LDA model between k = 2 and k = 6. The results are compared with additional algorithms for detecting human behavioral patterns, K-mean, Agglomerative Clustering, Gaussian Mixture, and Bayesian Gaussian. Mixture. The set of topics that best represents human activity is selected based on the extensive expert knowledge of the geographical area where the data was collected. Additionally, the patterns obtained with the banking dataset are compared with those obtained using the telecom dataset. Finally, a temporal analysis of the patterns obtained is presented. Multiple

sets of human behavioral patterns are extracted, each with a subset of the digital traces. Each subset contains digital traces obtained in different periods; therefore, it is possible to have a spatiotemporal analysis. This spatiotemporal analysis is complemented to study the effects that the COVID-19 pandemic had, along with all the mobility restrictions imposed, on activity patterns. This experiment validates again what was expected in Aim 2 of this thesis. It gives the first approach to respond to Aim 3, which seeks to study how human behavioral patterns change over time and how this temporal dimension can be incorporated into modeling.

5.1.3. Experiment 3: Spatiotemporal human behavioral patterns: Model-embedded patterns

This experiment aims to study spatiotemporal patterns where the temporal dimension is embedded in the model used to identify human behavioral patterns. This experiment differs from the previous one because the temporal evolution of the patterns is analyzed by running the spatial model multiple times. Additionally, this experiment connects all the knowledge and insights obtained previously to propose a series of metrics that allow us to reduce dependence on extensive expert knowledge of the geographical area of study. These metrics are proposed to answer the question of Aim 1 of this thesis. In this way, this experiment responds to the three objectives set out in this thesis, formalizes the human behavioral pattern validation (Aim 1), collects and transforms digital traces to detect behavior patterns (Aim 2), and finally proposes a method to identify spatiotemporal patterns in such a way that the temporal evolution of the patterns is captured directly by the proposed model (Aim 3).

Unlike the two previous experiments developed with a dataset of digital traces gathered for a single city, this experiment uses the social media dataset (See Section 3.4), whose coverage is worldwide. Consequently, the level of aggregation will no longer be areas of the same city, and this experiment data is aggregated at the city level.

In summary, this study aims to detect the set of activity topics that best represent and describe the activities carried out in various cities worldwide. In order to achieve this objective, a sequence of experiments is defined as follows:

- The first step involves comparing results obtained by applying the methodology proposed in this study (DTM) with existing models for detecting activity patterns. The state-of-

the-art models used for comparison are Latent Dirichlet Allocation (LDA), K-Means, k-Shape (a variation of K-Means specialized in grouping time series), and Time Series K-Means. These models are static; they do not consider temporal dependency. Therefore, we train a separate model for each time slice.

- The static models trained for each time slice assign a different label for similar topics in different time slices. In order to solve this and match activity topics from different time slices. We create a heuristic to assign the same labels to the most similar activity topics obtained independently in different time slices. The results of this analysis are presented in section 8.2.1.

- Additionally, one-year and three-year time slices are compared to determine the best grouping of the temporal dimension. The results of this analysis are presented in section 8.2.2.

- Finally, the model that best describes the city behavior is selected, and the optimal number of activity topics is calculated. The results of this comparison and analysis are presented in section 8.2.3.

5.2. Results

This section introduces the results of applying the methodology described in Chapter 4 in three different scenarios. Chapter 6: **Spatial human behavioral patterns**, shows its application for identifying spatial patterns of human activity using the telecom dataset. These spatial patterns describe the activity assuming a whole week as the temporal basis. Therefore these patterns summarize the activity of the entire study period in this spatial pattern of weekly behavior.

Chapter 7: **Spatiotemporal human behavioral patterns: Multiple static models**, shows a first approach to analyzing the temporal evolution of behavioral patterns. This approach applies the methodology proposed in Chapter 4 multiple times over consecutive time windows. In this way, these results aggregation summarizes all human activity in a set of spatiotemporal human activity patterns. This analysis is applied to the banking dataset. Also, it compares the patterns obtained from purchase behavior with credit and debit cards with the patterns obtained in Chapter 6 that summarize the behavior of telephone calls.

Finally, Chapter 8: **Spatiotemporal human behavioral patterns: Model-embedded patterns**, presents spatiotemporal human activity patterns obtained directly from a dynamic model and not by applying the spatial model multiple times over different time windows. These spatiotemporal patterns characterize social media activity worldwide, showing its evolution for more than a decade. In this analysis, the geographical area of study is at the city level and not at the neighborhood or commune level as in Chapters 6 and 7. This aims to understand structural and behavioral changes in the world population.

Chapter 6

Spatial human behavioral patterns

Today we have the opportunity, without precedents, to analyze human land use or mobility behavior in a city, country, or even the globe. Some studies have analyzed existing data generated daily by mobile networks, primarily using geo-localization in Twitter, Foursquare, or cell phone records. Most of these studies use a small portion of data (a few days or a few million records). In this Chapter, we will apply latent semantic topic models to detect Human Activity Patterns as we explained in Chapter 4. This experiment was designed to address Aim 2 of this thesis (See section 1.2.2.2), determine if it is possible to detect human behavioral patterns from digital traces and propose alternatives to traditional algorithms for this task. Our methodology will be applied using the telecom dataset, a real extensive dataset of 880,000,000 calls made in Santiago City (Chile) over 77 days by about 3 million customers of a major telecommunications company. We proposed to use a latent variables clustering technique which allows us to detect four interesting clusters. We discovered that applying LDA allows us to discover two well-known clusters (residential and office area clusters). We also discover two new clusters: Leisure-Commerce and Rush Hour patterns.

6.1. Topic modeling using the telecom dataset

A topic model [147] can be considered a probabilistic model that relates documents and words through variables representing the main topics inferred from the text itself. In this study, this idea has been adapted to understand human behavioral patterns arising from the activities in the city by applying Latent Dirichlet Allocation (LDA), a topic model, using the telecom dataset. The rationale of using LDA in this problem is to model *human activity patterns*

as arising from multiple latent variables (topics) of behavioral patterns, where a topic is defined to be a distribution over a fixed *Activity Blocks* set. Specifically, we assume that K Human Activity Patterns (topics) are associated with the set of BTS, and each BTS exhibits these topics with different proportions. In this context, a Human Activity Pattern can be considered a mixture of topics, represented by probability distributions that can generate the *Activity Blocks* in a *Human Activity Patterns* given these topics. The inferring process of the latent variables, or topics, is the key component of this model, whose main objective is to learn from cell phone data the distribution of the underlying topics in a given dataset of *Human Activity Patterns*. The methodological detail to apply LDA on the telecom dataset is presented in Chapter 4.

6.2. Experimental setup and results

In our study, we set every activity pattern AP related to the sensor (BTS in the telecom dataset) s as the number of calls managed by that BTS every hour in a seven-day week. Therefore, each AP is a vector with $N = 168$ components (24 Activity Blocks per day, seven days per week), where every component reveals the activity of s during one hour. This analysis is achieved by holding the proportion of calls during that hour compared with the number of calls made during the week. Figure 6.1 illustrates two AP related to different BTSs. Every AP starts on Sunday and finishes on Saturdays.

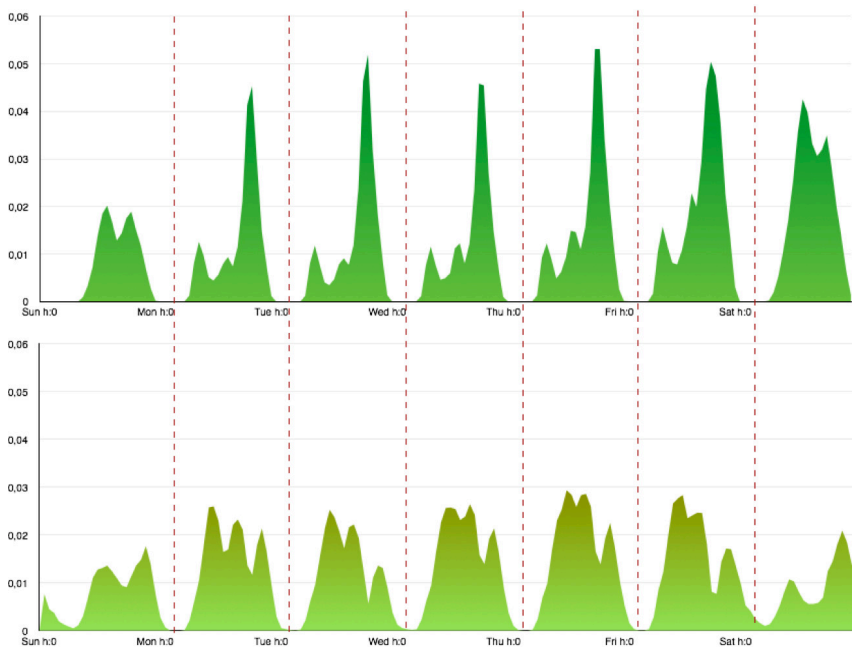


Figure 6.1: Activity Patterns example using the telecom dataset

6.2.1. Spatial human activity patterns identification

We have used Latent Dirichlet Allocation (LDA) to detect latent behavioral patterns arising from the interaction between individuals and the environment. LDA needs as input the number of topics K (human behavioral patterns) representing different activity patterns. In order to validate the optimal number of topics, we executed LDA for each value $K = 2, \ldots, 8$ and selected, using expert knowledge, the value of K, which provides the maximal information and the minimal dimensionality.

Figure 6.2 shows the human behavioral patterns representatives (topics) obtained after applying LDA using $K = 4$. An analysis of these topics allowed us to hypothesize about the behavioral patterns. Figure 6.2(a) describes a behavior characterized by high activity during weekends, especially on Saturdays. During weekdays the behavior is regular throughout the days showing an increasing activity with peaks in afternoons (19:00 hrs). This behavior seems to belong to leisure or commercial areas.

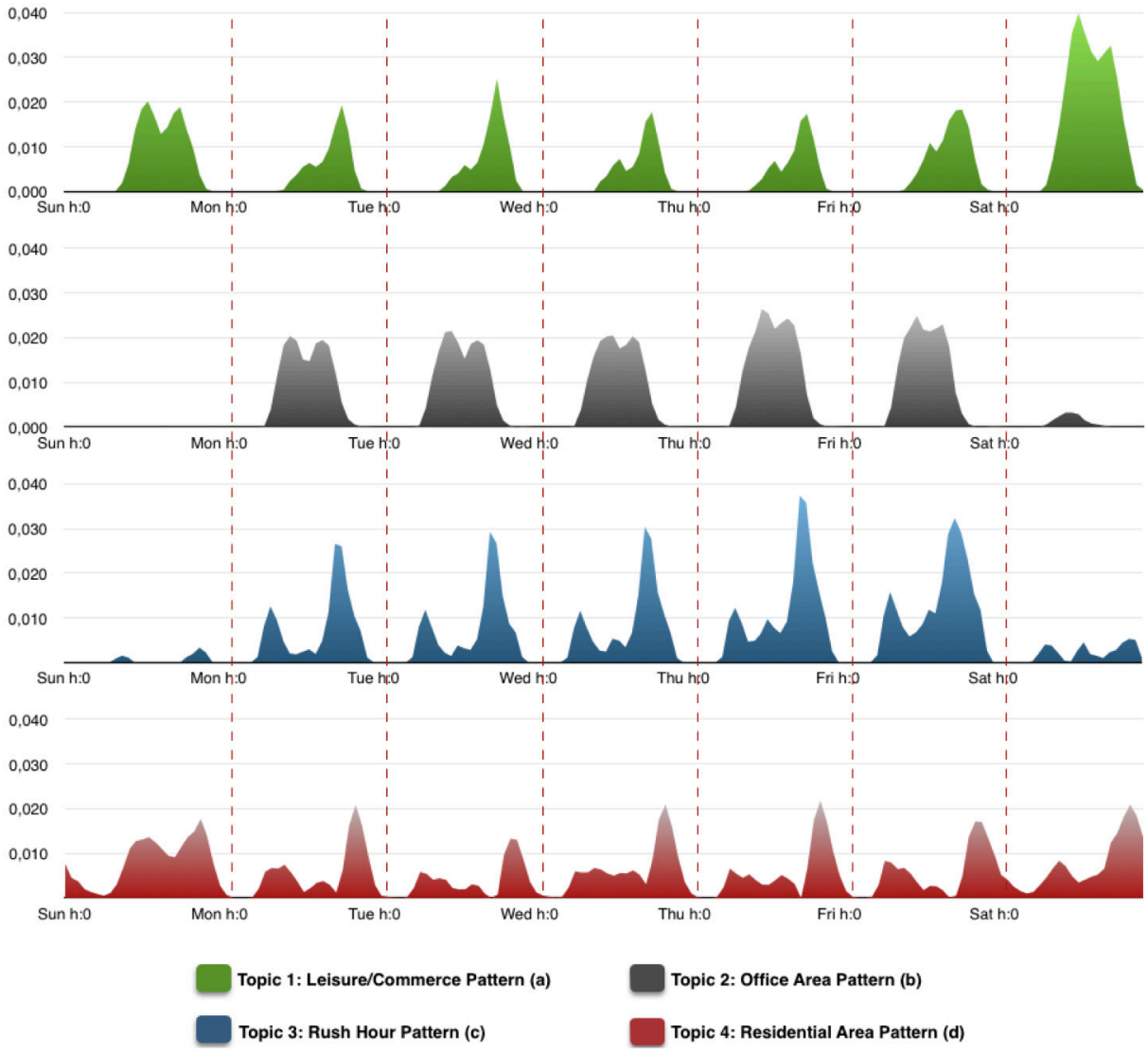


Figure 6.2: Latent behavioral patterns detected after applying LDA

Figure 6.2(b) shows a human behavior characterized by high and regular activity during weekdays and almost non-existing activity during weekends. During the day, there is a decreasing activity at lunchtime (13:00 hrs), indicating probably office areas activity. Similar behavior to 6.2(b) is presented in Figure 6.2(c), but in this case, the activity during the day shows a different pattern. Each day is characterized by three peaks which might be associated with the times when people typically get to work, go for lunch, and leave work. The first one is in the morning at 09:00 hrs., the second one, and considerably less than the others, occurs at lunchtime, and the last one is the highest peak during the day, which occurs in the afternoon at 19:00 hrs. This activity pattern seems to belong to areas with

high human displacement and traffic jams because every peak occurs at rush hour. Moreover, the lunchtime peak almost dissipates on Fridays to contribute to the afternoon peak. This phenomenon could be explained because people leave work early on Fridays.

Finally, Figure 6.2(d) presents human behavior with activity all week, but the behavior during weekdays and weekends differs. During weekdays the activity starts at 06:00 hrs., decreases along the morning, and increases again after 19:00 hrs. with a peak at 21:00 hrs. Moreover, the activity is higher on weekends, especially on Sundays. This behavior is typical of residential areas where individuals come from work in the afternoon on weekdays and weekends and stay at home.

6.2.2. Spatial human activity patterns stability

In order to analyze the stability of the patterns discovered using the methodology presented in this work, we divided the telecom dataset (S) into two equal data subsets. We applied our methodology to discover human activity patterns using each subset. The first dataset ($S1$) contains calls made between 2013-04-18 and 2013-05-26, and the second one ($S2$) between 2013-05-27 and 2013-04-07.

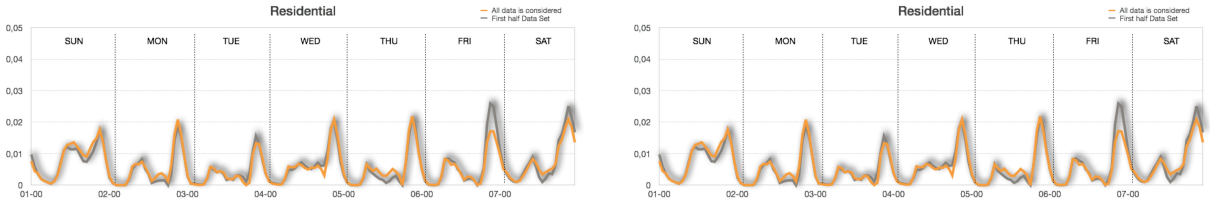
To quantify the pattern stability under different datasets, we used Cosine Similarity. This measure is defined as follows:

$$COS(AP_b, AP_c) = \frac{\sum_{i=1..N} A_i^b \cdot A_i^c}{\sqrt{\sum_{i=1..N}(A_i^b)^2} \cdot \sqrt{\sum_{i=1..N}(A_i^c)^2}}$$

This variable varies in the range $[-1, 1]$ and equals one only when the two AP, AP_b, and AP_c are exactly coincident. We examined how the discovered patterns change as the dataset varies from the first ($S1$) and the second half ($S2$) to the whole dataset (S). The results for the comparison between the patterns discovered using the first half and the whole dataset are shown in figures 6.3(a), 6.4(a), 6.5(a), 6.6(a). In general, results show that most discovered human behavioral patterns are stable when the dataset is reduced. Indeed, cosine similarity between S and $S1$ are 0.978, 0.982, 0.983, and 0.988 for Rush Hour, Residential, Leisure/Commerce, and Office Areas, respectively. Similarly, the comparison between S and $S2$ – presented in Figures 6.3(b), 6.4(b), 6.5(b), 6.6(b) – also exhibit high rates of similarity: 0.885, 0.936, 0.806, 0.9886 for Rush Hour, Residential, Leisure/Commerce and Office Areas respectively. Patterns in this subset are lesser stable than patterns from the first half. The

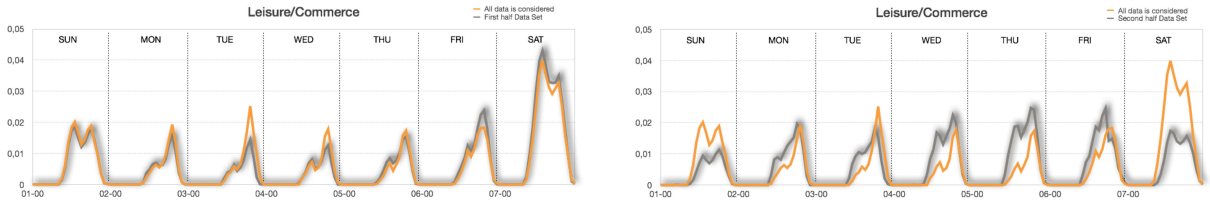
most significant differences occur in Rush Hour and Leisure/Commerce patterns. Some of these differences are explained because the Chilean winter holidays (June to July) are within this period. This reason causes fewer people to circulate through the city in rush hour and displaces some of the recreational/commercial activities to weekdays.

As a final remark, discovered patterns are very stable over time. Also, our methodology persists in finding out the same patterns, although a significant mobility pattern shift was present in $S2$ dataset (winter vacations). Of course, as a subject for future work, it would be interesting to discover the minimum dataset needed to avoid human behavioral pattern shift or vice versa.



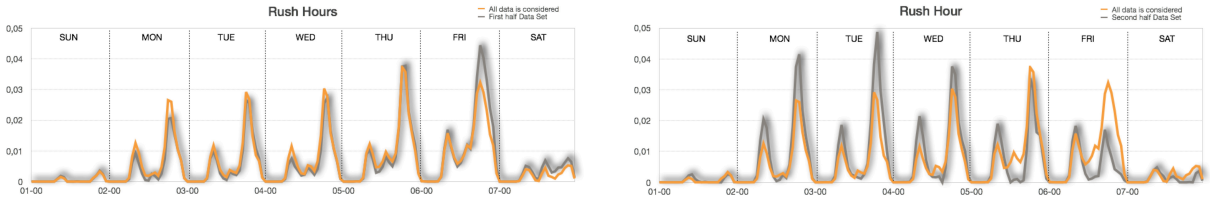
(a) First Half Dataset (b) Second Half Dataset

Figure 6.3: Residential Pattern Stability



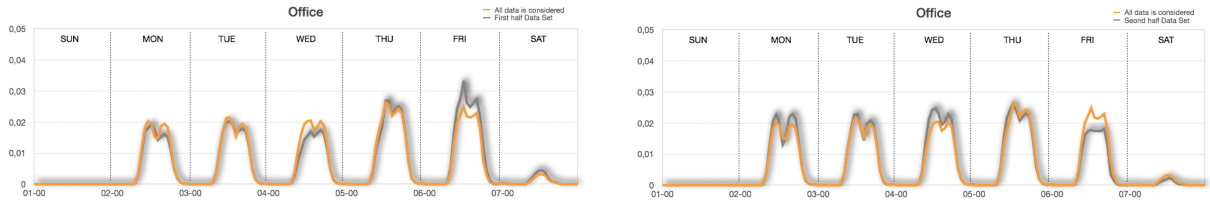
(a) First Half Dataset (b) Second Half Dataset

Figure 6.4: Leisure/Commerce Pattern Stability



(a) First Half Dataset (b) Second Half Dataset

Figure 6.5: Rush Hour Pattern Stability



(a) First Half Dataset

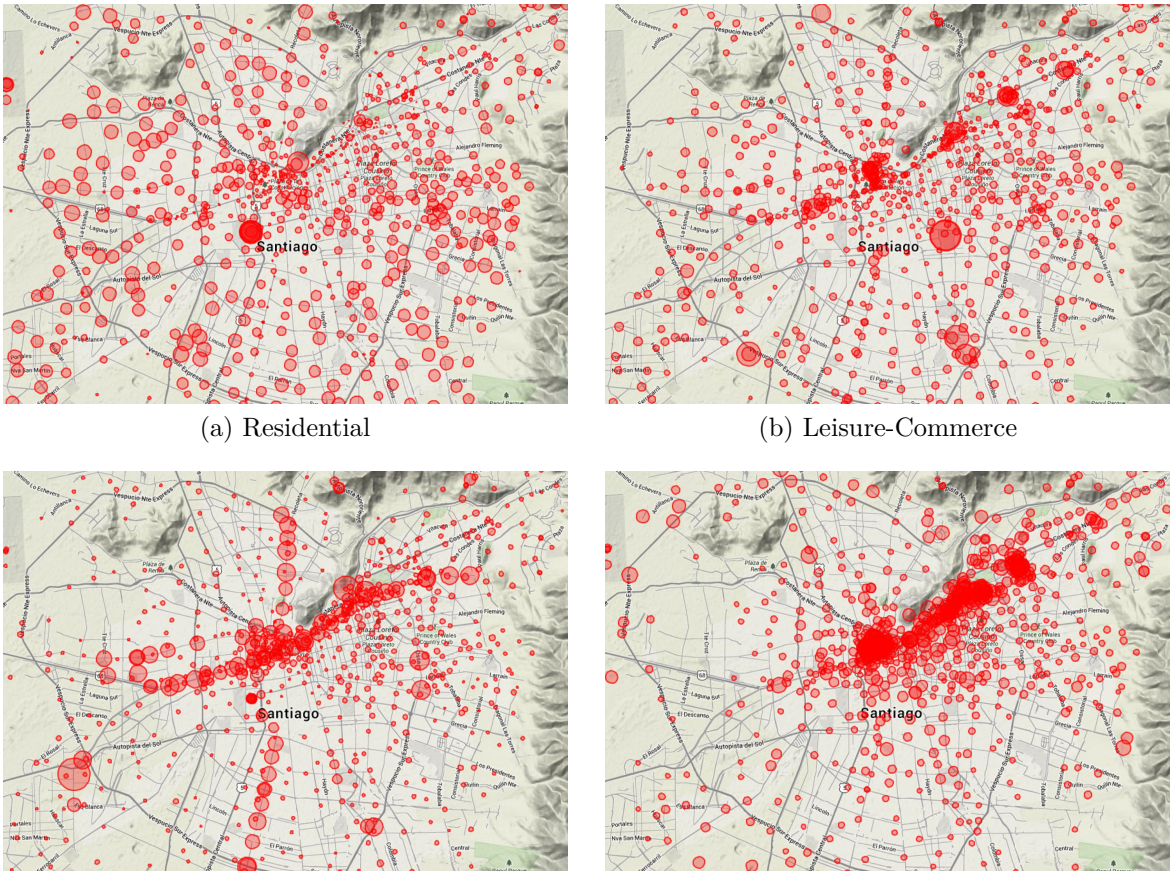
(b) Second Half Dataset

Figure 6.6: Office Areas Pattern Stability

6.2.3. An spatial comparison of static human behavioral patterns

In order to validate our human behavioral hypothesis, we used our results as a layer over the city map to have a geographical representation of the areas where we have real human interaction behavior. The use of LDA allows capturing the degree to which each human activity pattern is present for each BTS s. LDA returns a human activity score g_{ks} for human activity pattern (topic) k and sensor or BTS s, $\sum_k g_{ks} = 1 \; \forall s$. Figure 6.7 illustrates each human behavioral pattern (topic) over the city, where we see how different every pattern distribution is. In order to identify behavioral patterns easily, we discard all BTS towers with a human activity score lower than a given threshold θ.

The validation checks if interpreting the human activity pattern given in Section 6.2.1 correlates with the BTS infrastructure located in a geographical area and its vicinity. Since a database detailing the actual human behavior of the city and the different uses we have identified are unavailable, we use our expert knowledge of the city of Santiago.



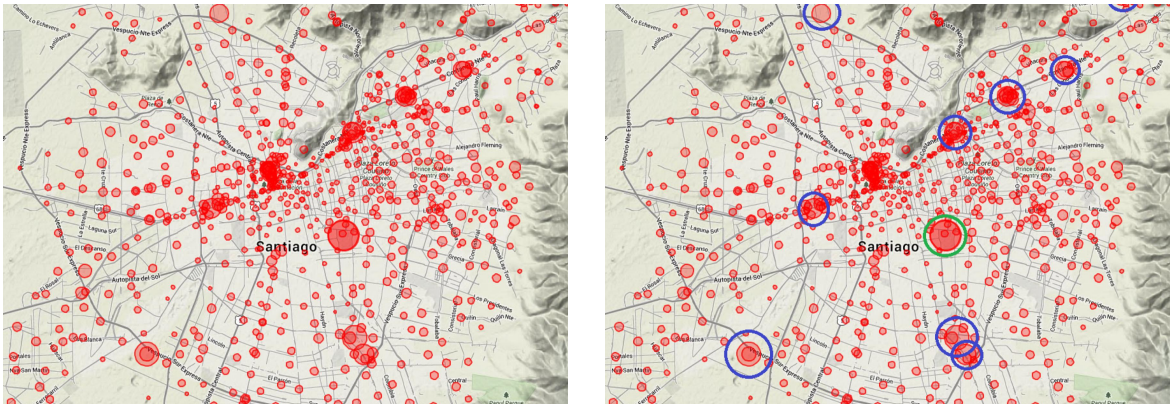
(a) Residential (b) Leisure-Commerce

(c) Rush Hour (d) Offices Areas

Figure 6.7: Geographical representation of human behavioral patterns

The Residential activity pattern represented in Figure 6.7(a) shows a higher human activity score in the periphery of the city center, where the business and commercial center is located. The periphery of the city center contains the most extensive residential zones in Santiago. In this pattern, there is no presence of zones with scores considerably higher than others, but just one area which covers the *Movistar Arena*. It is one of South America's biggest multi-purpose colosseums behind Brazilian arenas like Ginásio Ibirapuera, HSBC Arena (Rio de Janeiro), and Maracana Arena. The high score in this area is due to the events presented in this location being scheduled when the residential pattern presents high activity.

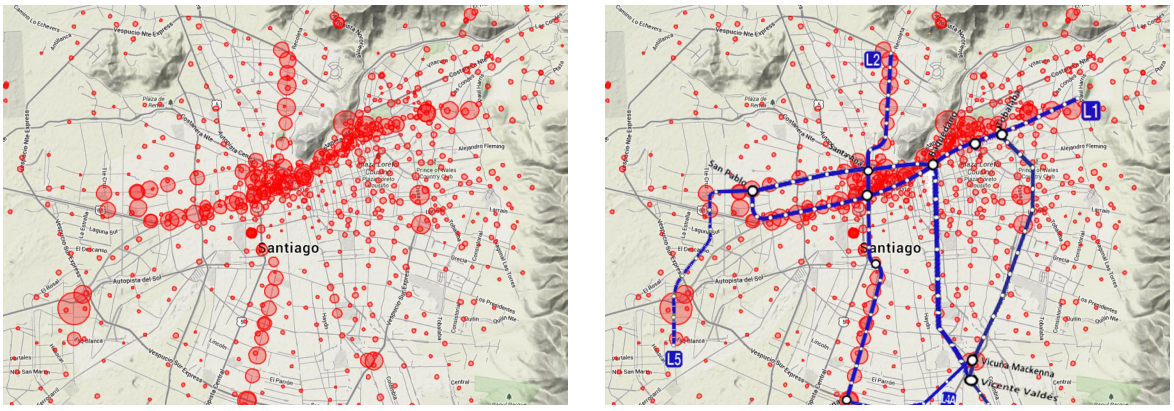
Figure 6.8(a) presents the Leisure-Commerce behavioral pattern. This human activity pattern presents a high intensity in some city points. In order to validate this pattern, we have highlighted these points (See Figure 6.8(b)), where the blue circles contain the main shopping malls in Santiago. The green circle contains Chile's largest stadium, with tennis courts, an aquatic center, a gymnasium, a velodrome, and a BMX circuit.



(a) Leisure-Commerce (b) Shopping Malls of Santiago

Figure 6.8: Leisure-Commerce pattern validation

The Rush Hour activity pattern is shown in Figure 6.9(a). This pattern has a high human activity score in two main areas. The first runs horizontally through Santiago, and the second runs vertically. These patterns are highly correlated to the subway network of Santiago (See Figure 6.9(b)).



(a) Rush Hour (b) Rush Hour & Subway Network

Figure 6.9: Rush Hour pattern validation

6.3. Experiment conclusions

We have applied the proposed methodology to understand the behavior of a city by discovering human activity patterns. The novelty of our approach is the use of latent variables over more than 3 million people's data. Inferring these variables, which are not directly observed in data, has proved to be highly satisfactory. We discovered four human behavioral patterns. Two of these patterns are very well known (human activity associated with office

and residential areas), and two patterns are new information: Leisure-Commerce and rush hour patterns. The leisure-commerce pattern is related to where people can spend their free time, which correlates with shopping centers, cinemas, and parks. Rush hour pattern appears at a specific time and over certain streets, avenues, and highways. This information could be important for urban planning, traffic management, and public transport public policies. As future work will be addressed in the following experiments, we are focusing on finding new sources of information that will allow us to validate our approach with ground truth and explore a dynamic assignment of human behavioral patterns. We also propose to study the evolution of a city over some years, find ways to obtain the optimal number of human behavioral patterns and cross this information with other sources, such as geo-referenced data from Twitter.

Chapter 7

Spatiotemporal human behavioral patterns: Multiple static models

In this Chapter, we use the proposed methodology to analyze the temporal evolution of human behavioral patterns. We apply the spatial model presented in Section 4.4.2 on multiple time-windowed datasets. This way, we compared the patterns obtained from information gathered at different periods. Additionally, we propose and validate a methodology to assess the actual impact of lockdown measures based on the banking dataset, an anonymized and geolocated dataset from credit card transactions,

Additionally, we identify human activity patterns in using credit cards using unsupervised Latent Dirichlet Allocation (LDA) semantic topic discovery. We apply these results to quantitatively assess the changes in people's behavior under the lockdown measures because of the COVID-19 pandemic. An unsupervised latent topic analysis uncovers the main patterns of credit card transaction activity that explain the behavior of the inhabitants of Santiago City. The approach is non-intrusive because it does not require people's collaboration to provide anonymous data. It does not interfere with the actual behavior of the people in the city; hence, it does not introduce any bias.

Splitting the banking dataset into multiple consecutive subsets allows us to analyze the behavior of the activity patterns over time and, in particular, to focus on the effect of non-sanitary measures to control the COVID-19 pandemic. We identify a strong downturn of economic activity as measured by credit card transactions (down to 70%), and thus of the economic activity in city sections (communes) subjected to lockdown versus communes without

lockdown. Independent data from mobile phone connectivity confirm this behavior change. The activity reduction emerged before the lockdowns were enforced, suggesting that the population spontaneously implemented the required measures for slowing virus propagation.

7.1. Topic modeling using the banking dataset

In this experiment, we apply the topic modeling approach presented in Chapter 4 to characterize human activities in the city using geographically tagged credit and debit card transaction data. Essentially, topic modeling assumes that any human behavioral pattern $\mathbf{AP}^s$ of a sensor s (POS) can be expressed as a linear combination of K activity topics $\left\{\mathbf{AT}^0, \ldots, \mathbf{AT}^{K-1}\right\}$, that is, $\mathbf{AP}^s = \sum_{k=0}^{K-1} \theta_k^s \mathbf{AT}^k$. Thus, human behavioral pattern $\mathbf{AP}^s$ is described by a mixing of activity topics $\boldsymbol{\theta}^s$, aka topic distribution of the document. The generative model of LDA assumes that $\boldsymbol{\theta}^s$ follows a Dirichlet distribution of symmetric parameter $\alpha < 1$. Activity blocks are the equivalent in our problem to the words in the document processing applications, that is, the possible values of activity at each hour $\boldsymbol{A} = \{AP_t^s\}_{t,s}$, where t and s extend over the hours in the week and all sensors (point of sales) respectively, without duplicated values. Human activity patterns are composed of activity blocks, with mixing parameters $\boldsymbol{\varphi}^k$ that follow another Dirichlet distribution of symmetric parameter $\beta < 1$. The topic that the activity block AP_t^s belongs to is denoted by z_{ts}, which follows a multinomial distribution of parameters $\boldsymbol{\theta}^s$. Finally, the activity block AP_t^s in each time position t of the activity pattern $\mathbf{AP}^s$ follows a multinomial distribution of parameters $\boldsymbol{\varphi}^{z_{ts}}$. The LDA generative model proceeds by generating the topics in the document (activity pattern), the words (activity blocks) in each document, the precise topic for each word, and the selection of the words in each position of the document.

In this experiment, we used a python implementation provided by Gensim [153, 154] to build up the LDA model, that is, to discover the latent activity topics and the decomposition of the human activity patterns into them.

7.2. Experimental setup and results

In this experiment, we first process the banking dataset well before the pandemic, during the year 2017, in order to obtain reference human activity patterns extracted by the LDA

algorithm. LDA finds a small set of human behavioral patterns extracted from the massive banking dataset. Summarizing vast and sparse data facilitates the decision-making process in policymaking. The use of LDA to detect behavioral patterns provide highly interpretable results. In this methodology, the analyst or policymaker must set up the number of patterns to be detected. For example, if we specify that $k = 2$, we assume that every activity pattern AP of a given point of sales terminal is a combination of two human activity patterns. The parameter K apriori is unknown. Therefore, we need to calibrate it.

To set the parameter k we explored LDA results for each value of $k \in \{2, \ldots, 6\}$. We selected the value of $k = 4$, which maximizes the information content and the interpretability of the human activity topics. We measure the information content of the set of extracted topics by their direction divergence; that is, more divergent vectors provide better representation axes to describe the space of vectors under analysis, in this case, human activity patterns. Therefore we compute the cosine similarities between all possible pairs of activity topics, using the norm of the resulting matrix as the information measure we want to minimize. In all exploration experiments, $k = 4$ provided a minimum value of this information measure relative to other selections of k. As different cells (i, j) contain a different number of terminals, in these evaluations, we weighted the data by the number of terminals $|H_{ij}|$ and the number of transactions T_{ij}.

To train the LDA model, we used two implementations of LDA provided by two python libraries: *gemsim* and *sklearn*, we denote by LDA_g and LDA_s respectively. We compare these LDA implementations against other clustering techniques such as Mini Batch KMeans (MBK), Agglomerative Clustering (AC), Gaussian Mixture (GM) and Bayesian Gaussian Mixture (BGM), each of these techniques was trained using k clusters, where $k \in \{2, 3, 4, 5, 6\}$. Notice that none of these methods directly allow for weighting observations, and, therefore, we oversampled the dataset using random sampling with replacement. The resulting dataset has size os times the size of the original dataset, and we executed experiments for each value of $os \in \{10, 20, 40, 60, 80, 100\}$. Lower values of ov produced unstable land use patterns. Considering all parameter combinations, we ran 4,320 experiments that were trained toward obtaining land-use patterns over the three geographic levels of the Metropolitan Region referenced in section 3.3.

To summarize, we run a series of experiments for each geographical area varying the following parameters:

- Study Area: [Greater Santiago, Inner Santiago, Downtown Santiago]
- Spatial Aggregation: [$100x100$ grid, $400x400$ grid]
- Weights: [C_x_Cell, T_x_Cell, log(C_x_Cell), log(T_x_Cell)]
- oversampling size: [10,20,40,60,80,100]
- Models: [LDA_g, LDA_s, MBK, AC, GM, BGM]
- Number of topics/clusters: [2,3,4,5,6]

Figure 7.1 shows the optimal human behavioral patterns obtained for the area of Santiago City. The vertical partitions correspond to the days of the week starting from Sunday. The x-axis is the time measured in hours. The y-axis is the normalized value of the activity pattern computed by dividing all vector components by the value of the maximum component. Overall, these patterns are consistent with those reported in previous work using other data sources like the telecom dataset from Section 6 (see, for example, [50] and [155]). We interpret these similarities as preliminary evidence that human activity patterns derived from payment data are relatively robust to the pattern extraction methods. However, previous research has shown that LDA models overperform other clustering techniques when detecting behavioral patterns using the telecom dataset ([50]). In our experiments also, LDA models outperform other clustering techniques, especially in interpreting the underlying patterns. For this reason, in the rest of this experiment, we only refer to the results obtained with the LDA techniques.

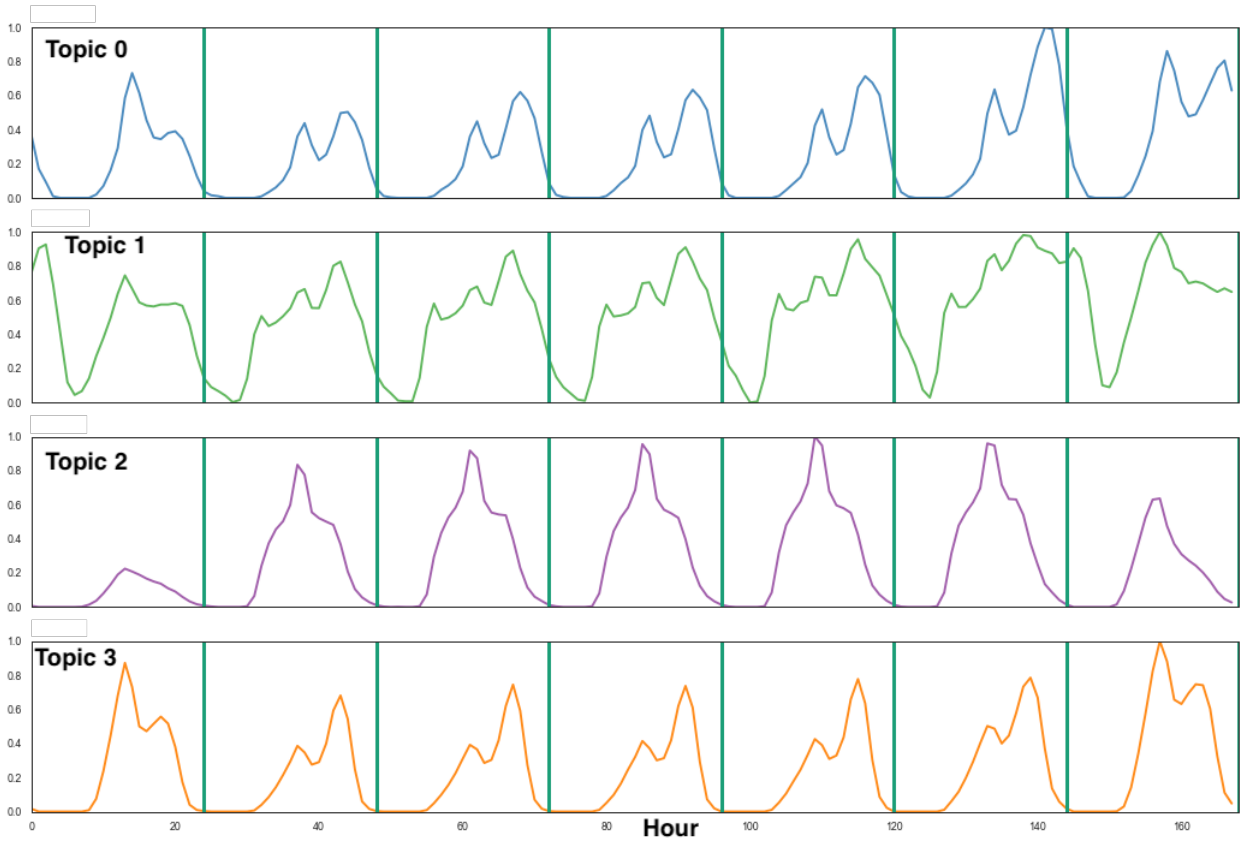


Figure 7.1: LDA detected topics in Santiago city, $k = 4$. The vertical partitions correspond to the days of the week starting from Sunday. The x-axis is the time measured in hours. The y-axis is the normalized value of the activity pattern.

The interpretation of these topics is as follows:

- Topic 0 - **residential** is characterized by two high activity peaks during weekdays, localized around lunch and dinner. Notice that the second peak on Fridays occurs around 11:00 hrs, reflecting that people used to have dinner later this day. Also, the second peak on Sundays is much less pronounced, indicating that citizens were less prone to go out dining on Sunday evenings.

- Topic 1 - **leisure/commerce** presents three peaks during the weekdays at 09:00 hrs, lunchtime, and around 19:00 hrs, roughly corresponding to when people used to commute to or from school or work. During weekends, this pattern is more evenly distributed throughout the day. Notice that there are high activities in the early hours of Saturday and Sunday, corresponding to nightlife habits before the pandemic.

- Topic 2 - **office areas** is characterized by a high and uniform activity during weekdays

and less during weekends. During the day, the main activity is between 09:00 hrs and 18:00 hrs, and there is increased activity at lunchtime (13:00 hrs), corresponding to office areas activity.

- Topic 3 - **rush hour** has peaks that roughly correspond to when people move from or to the working places in office/business areas.

7.2.1. Spatiotemporal human activity patterns validation

The LDA discovery of latent topics is unsupervised. Therefore there is no guarantee regarding the order of discovery or the identity of the topics. In order to establish correspondences among topics in different sets, for example, discovered from data extracted at different times, we compare patterns using the cosine similarity metric. This metric is widely used to compare land use patterns [50, 152, 156, 157] and provides a distance between two human behavioral patterns (AP), and it is defined as follows

$$COS(AP_b, AP_c) = \frac{\sum_{i=1..T} A_i^b \cdot A_i^c}{\sqrt{\sum_{i=1..T}(A_i^b)^2} \cdot \sqrt{\sum_{i=1..T}(A_i^c)^2}} \quad .$$

Cosine similarity value is in the range $[-1, 1]$ and equals one only when the two activity patterns, AP_b and AP_c, exactly coincide. It measures the relative orientation of the high dimensional vectors, thus very insensitive to the absolute magnitude of vector components and equivalent to the correlation measure for zero mean vectors. Cosine similarity has been extensively used in studies about the geographical distribution of land uses [158] using diverse information sources such as Twitter activity [159], Flickr tags [160].

For further confirmation of the above interpretation of the human activity patterns obtained from the banking dataset, we compare them with human activity patterns obtained from the telecom dataset and reported in the previous experiment in Chapter 6

The results for the cosine similarity between the human behavioral patterns discovered using the banking dataset and the telecom dataset are shown in Table 7.1. These results show that discovered activity patterns were relatively stable before the pandemic, independently from the data source and time frame. Indeed, every topic detected from the banking dataset relates to one detected from the telecom dataset with significant cosine similarity magnitude.

The interpretation of the human activity patterns [50] is similar to the one above for the banking dataset. We find cosine similarity above 0.8 for Rush Hour and Residential and above 0.93 for Leisure/Commerce and Office Areas, respectively.

Tabla 7.1: Cosine Similarity between human behavioral patterns discovered using the banking and telecom dataset in the Santiago city area.

		Telecom dataset			
		T0	T1	T2	T3
Banking dataset	T0	0.63	**0.82**	0.72	0.42
	T1	**0.80**	0.76	0.76	0.70
	T2	0.60	0.41	0.59	**0.93**
	T3	0.69	0.63	**0.94**	0.56

When observing the most similar topics among datasets, we have a complete view of human activity patterns and how different data sources can be informative. Figure 7.2 [Topic 1] shows the activity pattern associated with Office Areas obtained from both datasets. For the banking dataset, three patterns (green lines), Office Areas, and human behavioral patterns for Santiago city are shown. Comparing these two patterns indicates they capture very similar patterns with a cosine similarity index of 0.93. However, two differences are worth discussing. The first one occurs from Monday to Friday at lunchtime. While cell phone calls activity decrease during lunchtime, credit card transaction activity increases. This finding suggests that people in office areas tend to make fewer cell phone calls during lunch but must pay for their lunch. Therefore credit card activity increases. The second one occurs during weekends, especially Saturday, because in the areas where this human activity pattern is strong, stores are open during weekends. These distinctions suggest that despite the similar pattern of human behavior recovered by different sensors, an adequate interpretation of the results requires understanding the nature of the interaction between the sensors and the individuals.

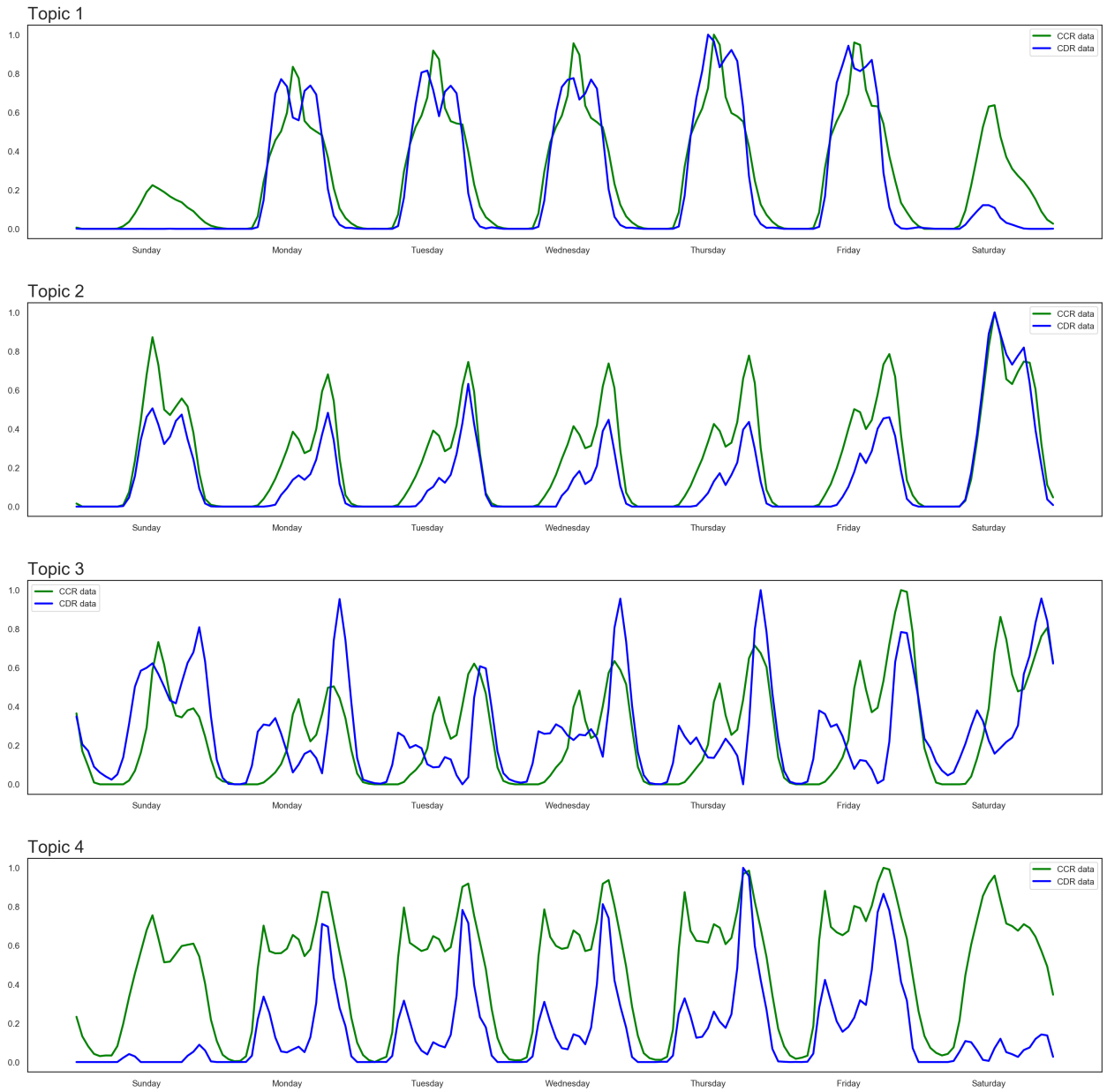


Figure 7.2: Comparison between the human activity patterns obtained by using the telecom dataset (CDR data, blue line) and the banking dataset (CCR Data, green line)

When we compare similar human activity patterns for leisure and commerce behavior, we found they are again remarkably similar, but with a few differences (see Figure 7.2 [Topic 2]). For example, this pattern presents higher intensity when detected by credit card data. In addition, the peak at lunchtime is more clearly identified by the banking dataset. These dissimilarities can be explained because commerce areas are more clearly identified through credit and debit card activity.

The comparison of residential area patterns (see Figure 7.2 [Topic 3]) appear to have

more significant differences. However, they are reasonably consistent when citizens are closer to their residences, such as on weekends and late at night. The observable differences can again be attributed to how users interact with their sensors in the corresponding location. For example, on Sunday evenings, the residence pattern detected using cell phone activity is high, but the credit card activity is not. This result is consistent with what we expect from users when staying at their homes. Similarly, on weekdays, mobile phone activity starts early for this residence pattern, but credit card use does not increase until lunchtime.

We finally compare the patterns associated with urban transportation or rush hours as illustrated in Figure 7.2 [Topic 4]. Compared to previous cases, this pattern presents the most significant differences between what is detected from the banking and telecom datasets. On the one hand, the patterns from payment data are characterized by three peaks associated with the times when people typically commute to work, go for lunch, and then back home. On the other hand, the pattern derived from mobile phone data early morning and late afternoon are very prominent, while the midday peak is more tenuous. These differences are explained because citizens will likely use their mobile phones while traveling during rush hours. However, they are less likely to engage in economic transactions when interacting with transportation infrastructure.

These results suggest that human activity patterns derived from two different data sources at different times present significant commonalities. Even if human activty patterns might exhibit some differences, they are mainly attributed to how users interact with the sensor we consider in the analysis. The patterns for transportation present more considerable differences indicating that for this particular type of activity, the two sensors might not necessarily capture the same usage. These analyses compare how different patterns manifest over time in a typical week. To have a more comprehensive comparison, in the next section we complement this temporal comparison with a special evaluation of how these activity patterns manifest through the city.

Additional validation of our interpretation of the banking activity topics comes from observing the spatial distribution of the topic in Santiago. Remember that LDA's outcome can be interpreted as the degree $g_{k,s}$ that each activity topic k contributes to the overall activity pattern of sensor s, such that $\sum_k g_{k,s} = 1 \ \forall s$. Hence, we calculate the contribution of each activity topic to the aggregated activity of each spatial cell $g_{k,(i,j)}$. Figure 7.3 displays the spatial distribution of Leisure/Commerce activity patterns in the area of Santiago City over-

laid with the place of the main shopping malls in Santiago. In the Figure, we have marketed these shopping landmarks in red, and we observe that their location correlate very well with the Leisure-Commerce pattern discovered through credit card data. This result corroborates the interpretation presented above. Later we will explore more in-depth into this human activity pattern because it is the one that reflects the most significant changes under the lockdown efforts for contention of the COVID-19 pandemic.

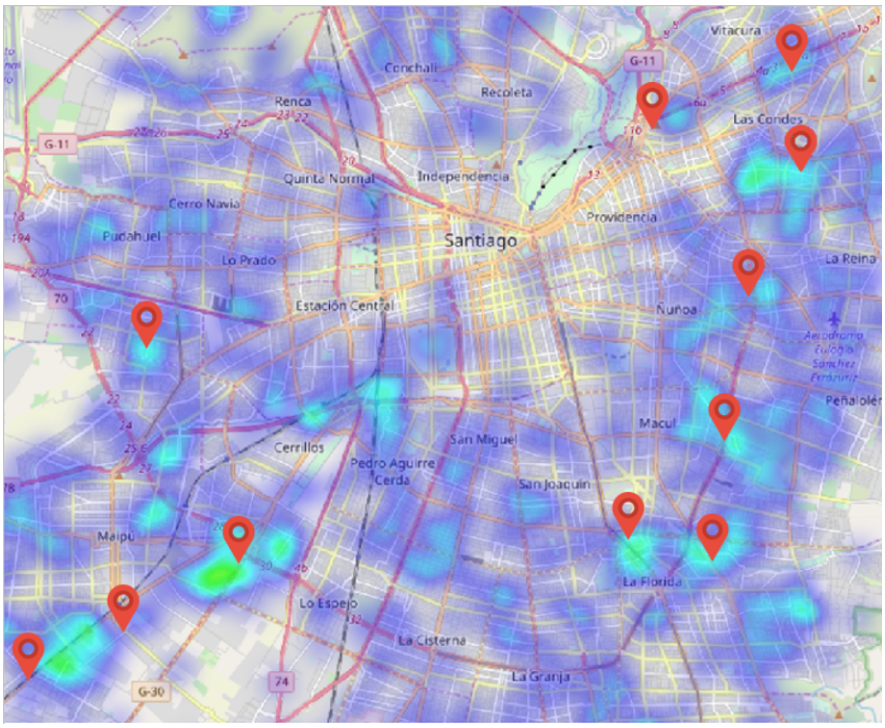
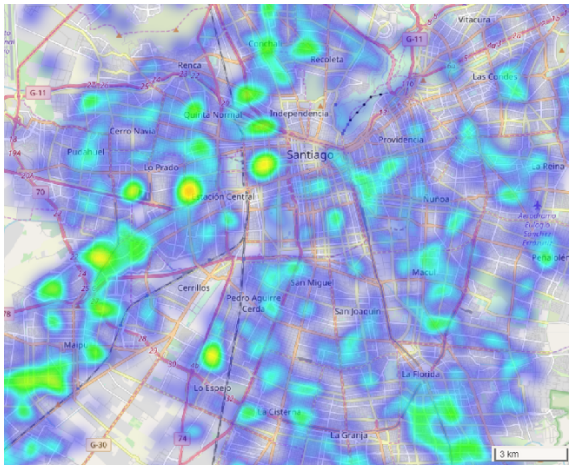


Figure 7.3: Spatial distribution of Leisure-Commerce activity topic obtained using the banking dataset, overlaid by the localization of the main shopping malls (Red markers). Blue color blobs spot the localization of POS with a high contribution of this activity topic in their LDA decomposition.

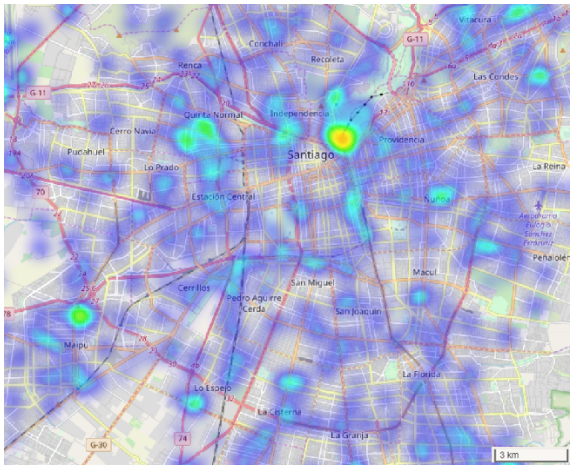
7.2.2. An spatial comparison of dynamic human behavioral patterns

We use a geographical representation of each human activity pattern to complement the previous analysis. Remember that LDA's output can be interpreted as the degree $g_{k,p}$ in which each sensor s belongs to each activity pattern k, such that $\sum_k g_{k,s} = 1 \; \forall s$. Figure 7.4 shows the geographical representation in the city of each human activity pattern detected using the banking dataset. Similarly, Figure 7.5 displays the human activity patterns from the telecom dataset. In both cases, more dense areas (color-coded red) correspond to a higher

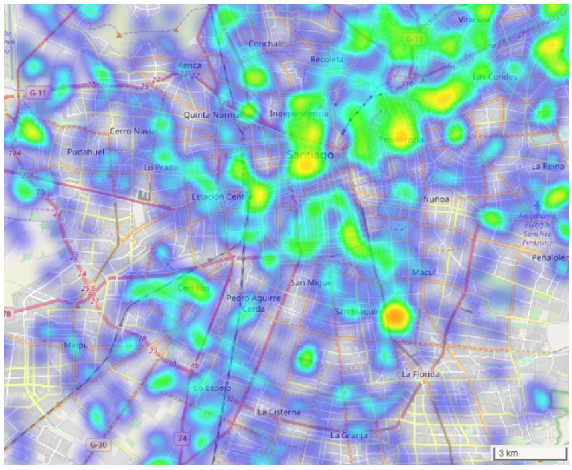
degree of belonging of sensor s to the human activity pattern k. For example, in panels (c) of both Figures displaying activity patterns of Office Areas, we find a more considerable activity in the upper right of the plot (the northeastern part of the city).



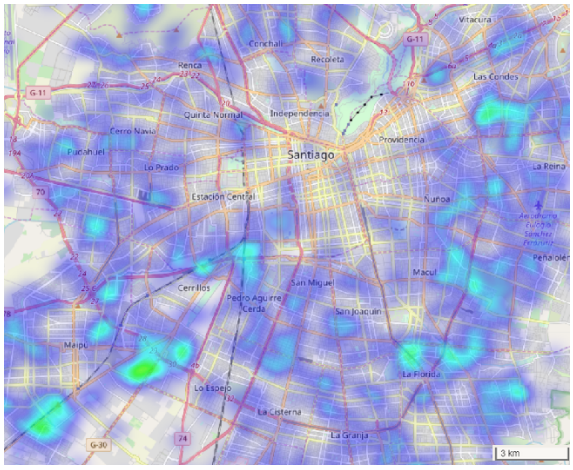
(a) Similar to Residential



(b) Similar to Rush Hour



(c) Similar to Office Areas



(d) Similar to Leisure-Commerce

Figure 7.4: Geographical representation of behavioral patterns - Santiago city

(a) Residential

(b) Rush Hour

(c) Office Areas

(d) Leisure-Commerce

Figure 7.5: Geographical representation of behavioral patterns using the telecom dataset

The spatial distribution of different activity patterns is consistent between the two data sources. This similarity is more clearly seen in the areas more intensively representing each activity. For example, in both Figures, Office Areas appear concentrated in the city's center, while the residential areas are more spread and active in the periphery. Nevertheless, the spatial distribution of cell towers in the telecom dataset is more concentrated in smaller cells. In contrast, the sensors in the banking dataset are more uniformly distributed and describe more complex mixtures of activities. Consistent with the analysis of the previous section, the most notorious differences are between the Rush-Hour Areas Patterns (Figures 7.4 (b), 7.5(b)). This difference can be explained because the patterns found using cell phone data are associated with traveling on a congested infrastructure (bus, car, subway);. At the same time, patterns found using credit card data reflect a different behavior, for example, buying

something before getting the subway home.

7.3. Spatiotemporal assessment of the impact of COVID-19 in human behavioral patterns

7.3.1. Human activity patterns during the pandemic

In order to assess the effect of non-pharmacological interventions (NPI) for COVID-19 on the behavior of the habitants of Santiago, we collected the credit and debit card transactional data in the pre-pandemic (the year 2019) and pandemic (the year 2020) periods. We remind the reader that Chile was in a commune-based lockdown between mid-March 2020 and Sept-2021. Therefore, our 2020 data was gathered during the lockdown period. In order to have a picture of the evolution of the activity topics, we extract the LDA topics ($k = 4$) of activity patterns from windows of 12 consecutive weeks, with an overlap of 10 weeks between consecutive windows. Therefore we have 13 sets of LDA activity topics per year. Figure 7.6 shows the overlaid activity topics of the year 2019 (red) and year 2020 (dark blue). In order to have similar topic assignation, we compute the cosine distance of the LDA detected topics on a 12-week window against the topics extracted from data of the year 2017 described above. We assign 2017 topics to the topics discovered in the time windows of 2019 and 2020, which are more similar according to the cosine distance that the meaning of the topics remains constant. Thus, Topic 1 always corresponds to the Leisure/Commerce activity topic, which has been most strongly affected by lockdowns and curfews. It can be appreciated that the late-night expenditures during the weekend have disappeared (red arrow). For a more quantitative appraisal of the changes between the activity topics assigned to the Leisure/Commerce from 2019 and 2020, we aggregate 3-hour intervals and compute a two-sided non-parametric Wilcoxon test to assess the statistical significance of the differences in activity. We highlighted with a red star those 3-hour periods with strong significant differences ($p < 0.0001$). It can be appreciated the strong impact that the non-pharmaceutical interventions have had on the behavior of the citizens of Santiago City.

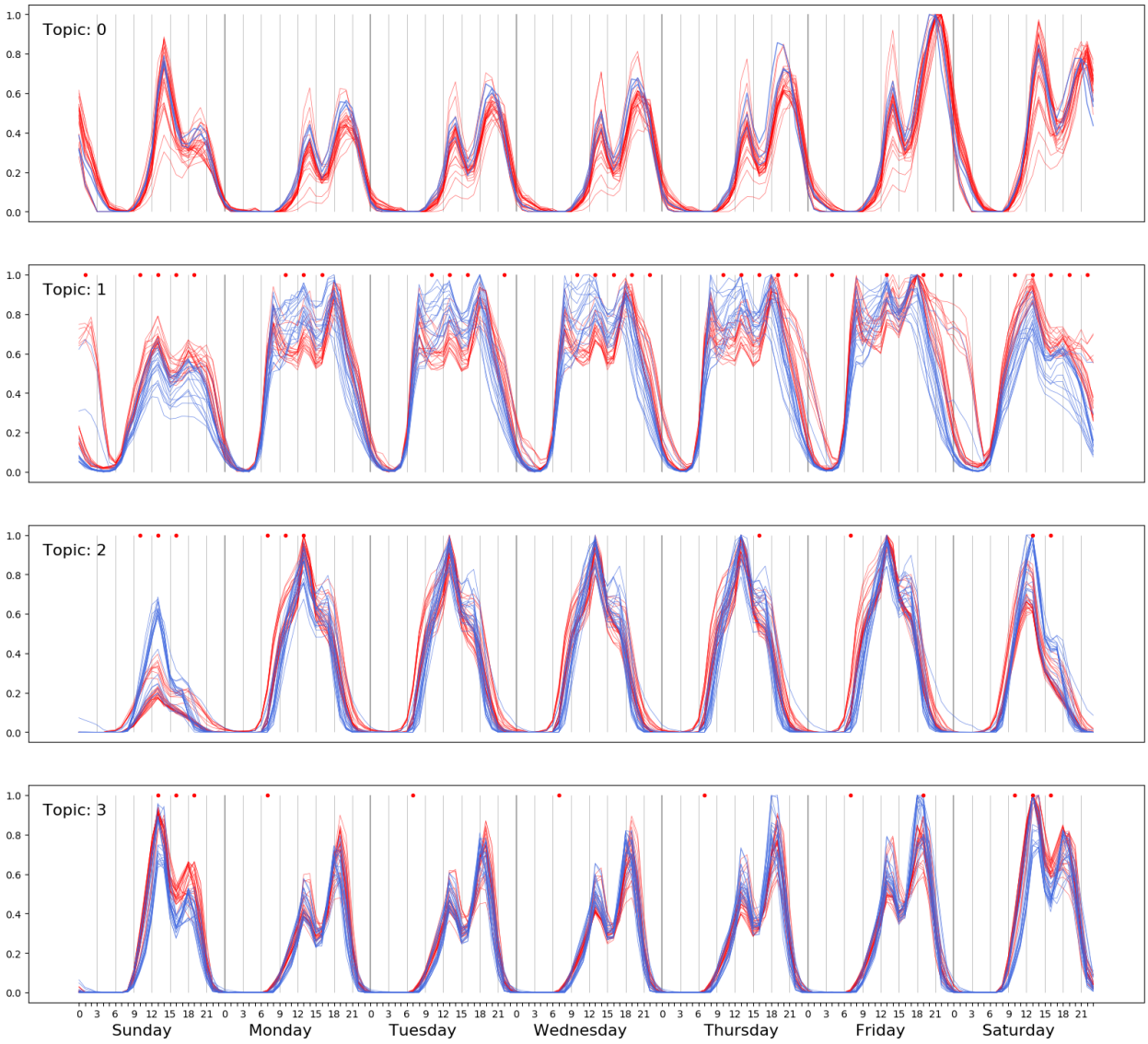


Figure 7.6: Change in activity topics due to the lockdowns and curfews imposed to curb the pandemic. Dark blue and red lines correspond to topics extracted from data from 2020 and 2019, respectively. The red dot denotes statistically significant ($p < 0.0001$) differences among pre-pandemic and pandemic activity topics in aggregations of 3 hours.

7.3.2. Impact of local policies, lockdowns and curfews

Observing the overall economic activity before the declaration of lockdowns and curfews allows us to assess their actual implementation and impact. In figure 7.7, we show the weekly activity pattern inferred from the telecom dataset with and without mobility restrictions. Considering that the contagion dynamics of the pandemic are not strongly related to each terminal's specific activity topic, we display aggregated measures regardless of the underlying

topics, that is, the average hourly activity. Figure 7.7A plots the average activity gathered from credit card records (CCR) in the banking dataset before (green) and after (red) lockdown in communes that did implement lockdown policies, showing a significant decrease in activity. If additionally, we consider the implementation of curfew for these communes, the box-plots in Figure 7.7C show a considerable drop in economic activity after curfew is declared. Communes that did not implement lockdown were less affected, as shown in Figure 7.7B. Nevertheless, the implementation of curfews significantly impacted them, as shown by the box-plots of Figure 7.7D.

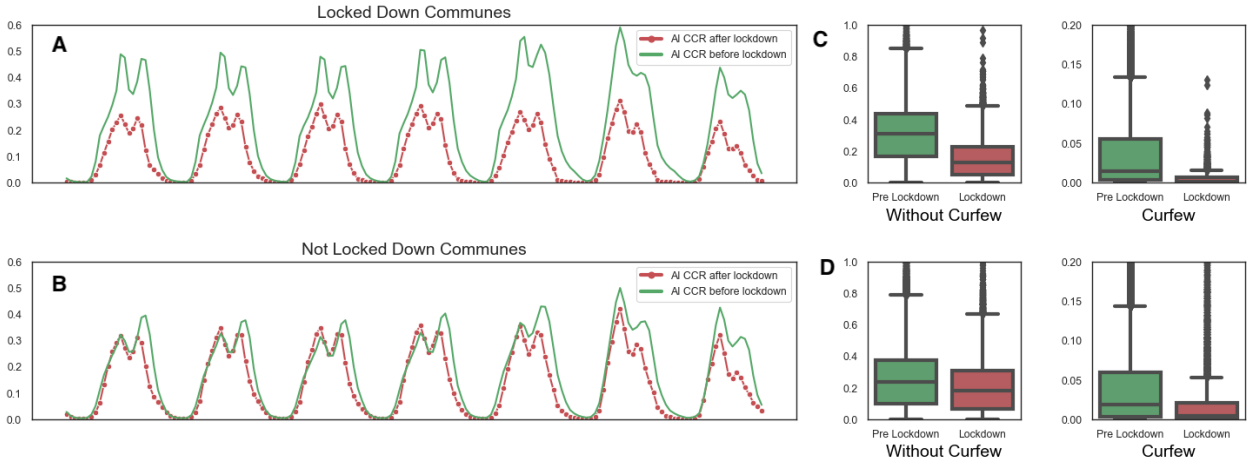


Figure 7.7: The effect of lockdown and curfew policies in Santiago, Chile. (A) Average weekly activity before and after lockdown for communes enforcing lockdown. (B) Same for communes not enforcing lockdown, (C) Additional impact of curfew on communes that enforced lockdown, (D) Same for communes that didn't enforce lockdown.

7.3.3. Aggregate activity measurement of impact

To assess the impact of non-pharmaceutical interventions implemented to curb the COVID-19 pandemic, we are also interested in the overall change in activity inferred from credit card records (CCR) in the banking dataset and how it compares with changes in activity inferred from call detail records (CDR) [161]. We compute the overall daily activity levels from both data sources for communes that have implemented lockdown and those that have not, as illustrated in Figure 7.8 from the beginning of March until mid-April. A red line highlights the critical date of March 26th. We can appreciate in both Figure 7.8B and Figure 7.8D that there is a sharp decrease of activity in both CDR and CCR data for all communes regardless of their implementation of lockdowns. Also, it can be appreciated in both Figure 7.8A and

Figure 7.8C that there was a sharp slowdown of activity in both CDR and CCR data almost ten days before the decision to implement lockdowns. Consistent with the early evidence presented in other countries [162], a relevant reduction in mobility and economic activity was voluntarily adopted for many citizens. However, the lockdown policy generated an additional impact.

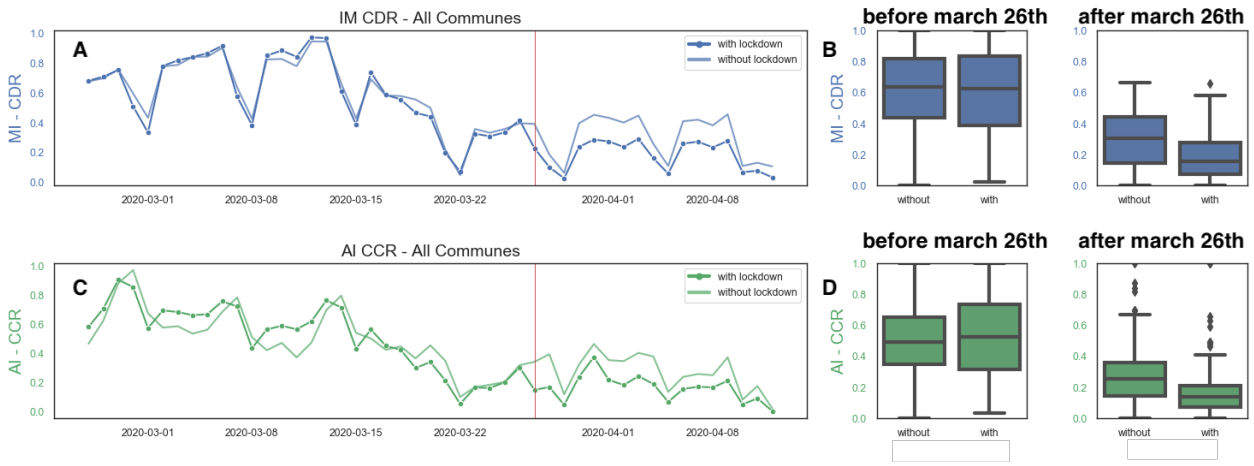


Figure 7.8: The effect of lockdown policies in Santiago, Chile. Aggregated data from the beginning of March 2020 until April 15th. The red line indicates March 26th. (A) CDR activity for communes with and without lockdown. (B) Box-plots of CDR activity in communes with and without lockdown before and after March 26th. (C) CCR activity for communes with and without lockdown. (D) Box-plots of CCR activity in communes with and without lockdown before and after March 26th.

Despite some variations, the overall trend captured by both data sources is consistent. They are associated with similar estimates in reducing activity due to the lockdown policy. Using CCR data, we estimate a reduction of 45.9% for communes not directly affected by the lockdowns and a 50.7% reduction if we estimate it using CDR data. Similarly, for communes implementing a mandatory lockdown, the reduction in activity from both sources is almost identical, with a 69.6% for the CCR estimate and a 69.7% for the CDR data. This agreement between data sources comes as a validation of using CCR data to assess the implementation of non-pharmaceutical interventions.

7.4. Discussion

Regarding assessing the impact of non-pharmaceutical interventions against the COVID-19 pandemic put in place by the Chilean government, the analysis of credit card transactional

Tabla 7.2: The overall effect of lockdown policies in Santiago, Chile, measured from mobile phone connectivity (CDR) and credit card transactions (CCR).

Data Source	Lockdown commune	Before march 26th	After march 26th	Reduction [%]
CDR	No	0.614	0.303	50.7%
CDR	Yes	0.604	0.183	69.7%
CCR	No	0.503	0.272	45.9%
CCR	Yes	0.520	0.158	69.6%

data in the banking dataset can provide information about the mobility of the population and the effect on the effect on the economic sectors of the changes of population mobility. In this regard, we use the banking dataset to evaluate if the impact of the mobility of adopting stay-at-home policies that encourage individuals to reduce non-essential trips has been reflected in changes in economic activities. Furthermore, we have been able to compare the estimates in mobility against those derived from CDR data in the period analyzed [161]. Other approaches to assess the economic impact of the pandemic use electric consumption as a proxy [163].

The analysis of the banking dataset identifies an economic sector that the pandemic-induced economic crisis has deeply hit. The Leisure/Commerce pattern is the most affected of the four activity topics identified in the pre-pandemic data. Our results comparing this activity topic in the year previous to the pandemic and the year of the pandemic show that Santiago inhabitants have changed their behaviors according to the lockdown and curfew policies. Though there is some literature on the disruption of supply chains due to COVID-19 [164], the impact on healthcare [165], forest degradation [166], and the economic impact in African countries, like the impact on cattle exports [167], on residents of some cities [168], there is little regarding the evaluation of the impact of COVID-19 on the people that work in the Leisure/Commerce in developed countries, but for some high-level analysis at corporate level [169, 170]. However, a large percentage of the labor force enrolled in the leisure and hospitality sector, that is, more than 13 million employees in March 2021 in the US according to workforce statistics [171] may fall into poverty with ensuing systemic health critical issues.

One of the facts that we have found is the voluntary reduction of activity that was apparent several days before the implementation of mobility restriction measures taken by the

governments. This result agrees with reported voluntary reductions in mobility estimated from the Google human mobility dataset [172]. Accurate information about the pandemic's evolution helps citizens make appropriate decisions toward curbing the pandemic. How the required mobility reduction impacts economic activity is unclear, as some researchers argue that increasing activity in parks and groceries/pharmacies has much less effect on the reproductive rate than staying at home [173]. Travel patterns appear to significantly impact the propagation of the virus, requiring a combination of sensible public policies and the willing collaboration of the community, as demonstrated by the case of Hong Kong [174]. Big data extensive studies have found that imposed public policies play a small role in the reduction of mobility [175]. The main factor contributing to reductions in mobility appears to be the fear of contagion [176]. Mobility and its relation to economic activity during a pandemic need a further research agenda [177].

7.5. Experiment conclusions

According to the results obtained in this experiment, the proposed methodology not only allows obtaining human behavioral patterns from cell phone digital traces, as we saw in Chapter 4.4.2 but also detects human behavioral patterns using credit and debit card get-tagged transactions. Therefore, human patterns show a certain degree of consistency independent of the type of sensor from which the digital traces were gathered. This experiment also introduced an approach to studying how human behavioral patterns change over time by training multiple spatial models.

Additionally, this experiment shows how the anonymized information about credit card transactions can be used to assess the follow-up and impact of non-pharmaceutical interventions implemented to curb the COVID-19 pandemic. We show how unsupervised latent topic analysis uncovers the main patterns of credit card transaction activity that explain the behavior of the inhabitants of Santiago City. Topics identified in the pre-pandemic year of 2017 are used to identify the topics produced by the analysis in 2019–2020, including the pandemic. Specifically, we can assess the impact on the leisure/commerce sector, which has suffered a substantial activity loss due to the pandemic. Additionally, examining the aggregated activity allows for assessing significant differences between communes that imposed lockdown and those that did not. Lockdown and curfew interventions lead to a reduction of

70% in credit card transaction activity.

Additionally, we found a spontaneous reduction of activity before the implementation of the lockdown of the same magnitude as the reduction achieved with the mandatory restrictions. The need for coercive measures to achieve mobility reduction to stop the virus spread may be reexamined in light of these findings. Future works will be directed to the disaggregate analysis of the information on the points of sale according to their nominal industrial activity to ascertain the pandemic's variable impact on the industry. This analysis will include the distinction between essential and non-essential services. In addition, a detailed analysis of the recovery after lockdown should be carried out independently for each commune.

Chapter 8

Spatiotemporal human behavioral patterns: Model-embedded patterns

Cities are complex and constantly evolving systems. Understanding their dynamics is crucial for transport management, urban planning, disaster response, and policy-making. Many researchers have studied city dynamics based on the information from virtual sensors that record the digital traces gathered from the people's interaction with the city's infrastructure. Telephone calls, purchases with credit card, check-ins to facilities, GPS records, and geo-tagged social media activity are key data sources that can be used as virtual sensors to obtain patterns of mobility and behavior.

This chapter will address the three objectives raised in this thesis. For this, a methodological approach is proposed that combines the results obtained previously in chapters 4.4.2 and 7. For this reason, throughout this chapter, a new algorithm (Aim 2) is proposed to extract spatiotemporal human behavioral patterns (Aim 3). Also, a set of metrics is proposed to reduce the dependence on extensive knowledge in the geographical areas of study (Aim 1). In this way, this study addresses this thesis's main objectives and some gaps detected in the literature review: Most previous studies are restricted to a single city and a single virtual sensor. Also, interpreting the patterns obtained depends on an exhaustive knowledge of the terrain. In addition, algorithms do not manage to incorporate the temporal dependency among patterns to analyze their change over time.

This experiment presents a methodology for detecting spatiotemporal patterns of city-level activity, incorporating the temporal evolution of these patterns. A set of metrics is

proposed to determine the set of patterns that best represent city activity. These metrics reduce subjective interpretations from analyzing city activity patterns of human behavior and their dynamics. We applied this methodology over a multi-sensor dataset of 32 million geo-tagged urban activities collected over 17 years in cities larger than 1 million persons or country capitals. The virtual sensors these data come from diverse public domain information sources detailed in Section 3.4. As a result, we report city-level activity patterns consistent with the known activity profiles carried out in the cities included in the study. Furthermore, our Dynamic Topic Model-based methodology outperforms classical approaches based on K-Means and Latent Dirichlet Allocation for spatiotemporal behavior pattern identification.

8.1. Dynamic Topic Modeling using the social media dataset

8.1.1. Definitions

In order to understand how Dynamic Topic Models (DTM) can be adapted to detect city activity patterns from human behavior, we need to formalize some definitions. First, we define an urban activity as the things people do, where they do, and when they do them. This experiment will represent each urban activity as a combination of its location (latitude and longitude) and a timestamp. These activities will be assigned to the closest city $c \in \mathcal{C}$, where $\mathcal{C}$ represents a set of cities. By aggregating and summing up these individual activities, we can characterize a city's activity pattern, denoted as XP^c. Understanding city activity patterns from human behavior is essential for various applications, including urban planning, transportation management, and resource allocation. By investigating the activity patterns within a city, we can gain insights into how people use the city's resources and infrastructure and identify areas for improvement. Using DTM to detect these patterns allows us to consider temporal aspects and model the evolution of activity patterns over time. This knowledge will provide a more comprehensive view of how a city's activity patterns change and evolve over time.

To investigate the temporal dependence and evolution of human activity patterns at the city level, we divide the period of interest into multiple time-slices, denoted as $\mathcal{S}$. Each time-slice, represented by $s \in \mathcal{S}$, corresponds to a defined period such as a month or a year.

Therefore, XP_s^c represents the activity pattern of city c at time-slice s. The city's activity is studied over multiple consecutive time-frames during a time slice. A time-frame is a period shorter than a time-slice, such as one minute, hour, or day; we will call an activity block the number of events carried out during this time-frame. Figure 8.1 explains the relationship between time-slices and time-frames.

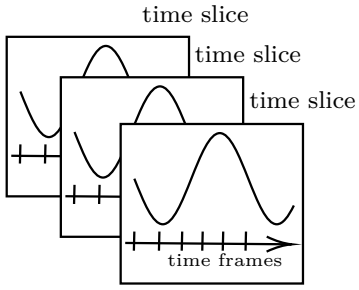


Figure 8.1: Time Slice and Time Frames

Formally speaking, we describe a raw human behavior pattern by a vector XP_s^c, where each component or activity block $XP_{s,t}^c$ denotes the number of events carried out on city c during the time-slice s in period t and therefore $XP_s^c = (XP_{s,1}^c, XP_{s,2}^c, \ldots, XP_{s,T}^c)(t = 1, \ldots, T)$, where T is the number of activity blocks. In this study, we set T into hourly periods during a week; therefore, each raw activity pattern XP_s^c is a vector with $T = 168$ components (24 h, seven days). To facilitate the comparison between cities, we define the normalized city activity pattern AP_s^c, where $AP_s^c = \frac{XP_s^c}{\sum XP_s^c}$, i.e., dividing its components by the total number of events of each city c during the time slice s. Therefore, we can interpret $AP_{s,t}^c$ as the percentage of events of the city c in the hourly time window t.

The complete description of how to adapt and apply DTM to detect human behavior patterns at the city level is presented in section 4.4.3, dynamic topic modeling using geo-tagged digital traces.

8.2. Experimental setup and Results

In this Section, we detail the results obtained in this experiment following the experimental setup proposed in Section 5.1.3.

8.2.1. Temporal matching heuristic

This section analyzes if it is feasible to align the labeling of human behavior patterns so that similar topics across sequential time slices get the same label. This re-assignment is crucial as traditional algorithms ignore the temporal dependence, resulting in activity patterns exhibiting similar behavior being assigned different labels because each time-slice is treated as an independent pattern.

Table 8.1 presents the results of applying the temporal-matching heuristic to the output of the state-of-the-art models. The table displays the Intertemporal Stability (See Section 4.6) percentual increment between the new topic label configuration and the original one. The Intertemporal Stability is measured using cosine similarity. In the table, the first column indicates the time slice aggregation utilized, and the results for each algorithm are presented for each aggregation scenario. The results also present different scenarios where we extract several numbers of topics.

The temporal-matching heuristic aims to increase the similarity of activity topics across consecutive time slices, thereby maximizing Intertemporal Stability. The results show that this objective is broadly achieved in all comparisons. Therefore, based on these results, the activity topic labels reassigned by the time-matching heuristic will be utilized in the following analysis.

Tabla 8.1: The result of temporal matching heuristic on behavior topics. Time matching impact is measured using the Intertemporal Stability with cosine similarity. The percentual increment between the heuristic arrangement and the original topic labels is shown.

Time Aggregation	Model	#Topics			
		2	3	4	5
one-year	K-Means	0.97%	7.57%	5.76%	16.46%
	k-Shape	1.09%	2.45%	0.78%	1.70%
	LDA	0.00%	0.00%	0.29%	0.09%
	TS K-Means	0.98%	0.87%	2.62%	2.41%
three-year	K-Means	3.22%	4.33%	3.53%	4.21%
	k-Shape	1.62%	0.05%	0.48%	0.89%
	LDA	0.00%	0.00%	0.35%	0.00%
	TS K-Means	3.67%	2.70%	4.99%	6.52%

8.2.2. Time-slices Aggregation

This section compares two time-frame choices to train the activity pattern detection models. We train several models using the proposed and traditional models in two different setups to achieve this. Firstly, the models are trained using one-year time-frames, generating 17 pattern subsets, one for each year in the dataset. Secondly, the models are trained using three-year time frames. The final time frame encompasses only two years, 2020 and 2021, which were combined due to the inclusion of information gathered during the COVID-19 pandemic. Both setups include the training of models with k=2 to k=5 topics.

Table 8.2 displays the comparison results between the two selected time frames. The outcomes are presented for the proposed and the traditional models, and for each of these models, results are shown after extracting between k=2 and k=5 topics. Intertemporal Stability, Intratemporal Similarity, Topic Smoothness, and Topic Consistency are presented for each model (See Section 4.6 for metrics details). The values of Intratemporal Similarity and Intertemporal Stability are computed using cosine distance. For each proposed indicator, three columns are displayed:

- The average results over the three-year time-slices
- The average results over the one-year time-slices
- The percentage variation between both results based on the three-year results.

The metrics were calculated using the cosine distance, commonly used to evaluate the similarity between topics [49, 115, 152, 178]. If two topics are identical, their cosine distance equals zero, reflecting a high similarity between them.

When comparing the results obtained with three-year time-slices on the models presented, it was found that the Intertemporal Stability index obtained in these experiments is lower in almost all scenarios. This result indicates that topic sets are more stable over time when using three-year time-slices than shorter time-slices.

In addition, it was noted that the Intratemporal Similarity index of the three-year time-slices experiments is lower than the results obtained using one-year time-slices. This outcome suggests that topics within each three-year time-slice are more similar than topics within one-year time-slices. However, this result cannot be analyzed independently because it could

indicate greater volatility and inconsistency in the topics obtained. In order to complement this matter, it is observed that the Topic Smoothness is less consistent in the topics obtained in the three-year time-slices, as is the Topic Consistency indicator.

In summary, although the results indicate certain disadvantages in using three-year time-slices, they also show better Intertemporal Stability, Topic Smoothness, and Topic Consistency than one-year time-slices. Therefore, it was decided to continue the analysis using three-year time-slice data sets to obtain a more complete and precise view of the topics and their evolution over time.

Tabla 8.2: Inter-temporal Intra-temporal topic validation metrics for different time-slice aggregations. The orange bars correspond to a 1-year time-slice aggregation, and the blue bars correspond to a 3-year time-slice aggregation.

Topics	Model	Intertemporal Stability			Intratemporal Similarity			Topic Smoothness			Topic Consistency		
		3Y	1Y	DIFF	3Y	1Y	DIFF	3Y	1Y	DIFF	3Y	1Y	DIFF
$k = 2$	K-Means	0.31	0.31	-0.58%	0.36	0.42	-13.6%	5.25	15.39	-65.8%	0.82	0.75	8.6%
	k-Shape	0.08	0.14	-40.3%	0.06	0.15	-58.5%	1.78	3.44	-48.1%	0.97	0.92	5.0%
	TS K-Means	0.31	0.29	6.5%	0.36	0.39	-8.4%	5.25	11.15	-52.8%	0.82	0.76	7.1%
	LDA	0.07	0.07	-2.42%	0.03	0.08	-60.3%	1.93	2.92	-33.9%	0.96	0.92	4.9%
	DTM	0.00	0.07	-89.3%	0.32	0.43	-24.4%	1.78	3.78	-52.9%	0.98	0.88	11.0%
$k = 3$	K-Means	0.29	0.41	-27.4%	0.31	0.48	-35.1%	5.37	18.66	-71.1%	0.81	0.65	24.1%
	k-Shape	0.12	0.14	-15.5%	0.05	0.16	-68.5%	1.82	3.62	-49.7%	0.97	0.89	9.0%
	TS K-Means	0.29	0.38	-21.8%	0.31	0.46	-32.0%	5.37	14.04	-61.6%	0.81	0.67	21.0%
	LDA	0.05	0.12	-60.2%	0.12	0.23	-49.0%	2.51	4.68	-46.2%	0.92	0.85	7.8%
	DTM	0.00	0.09	-93.4%	0.40	0.49	-17.4%	2.00	5.91	-66.0%	0.97	0.77	25.2%
$k = 4$	K-Means	0.34	0.43	-20.5%	0.35	0.47	-24.9%	6.47	20.5	-68.4%	0.76	0.65	17.1%
	k-Shape	0.19	0.18	6.3%	0.12	0.15	-21.8%	2.88	3.6	-20.0%	0.91	0.88	3.5%
	TS K-Means	0.34	0.41	-15.5%	0.35	0.46	-22.8%	6.47	15.95	-59.4%	0.76	0.66	15.4%
	LDA	0.04	0.14	-71.8%	0.15	0.26	-42.3%	2.95	5.17	-42.9%	0.89	0.83	7.7%
	DTM	0.00	0.1	-94.7%	0.42	0.52	-18.1%	2.24	6.72	-66.6%	0.96	0.74	30.6%
$k = 5$	K-Means	0.35	0.57	-37.4%	0.37	0.61	-38.8%	6.38	25.03	-74.4%	0.73	0.54	35.5%
	k-Shape	0.15	0.24	-34.8%	0.10	0.21	-49.5%	2.68	5.33	-49.7%	0.92	0.85	8.5%
	TS K-Means	0.31	0.46	-32.6%	0.31	0.46	-32.5%	5.64	17.47	-67.6%	0.78	0.65	19.1%
	LDA	0.02	0.14	-84.3%	0.14	0.26	-45.9%	2.88	5.47	-47.2%	0.89	0.83	7.3%
	DTM	0.00	0.09	-95.1%	0.37	0.63	-39.8%	2.58	9.22	-72.0%	0.95	0.71	33.2%

8.2.3. Model comparison and the optimal number of human behavior patterns

In this section, we will determine the number of patterns/topics that best characterize the activity carried out in the study cities. For this, the previous findings will be considered, and an evaluation will be carried out based on the models trained using the three-year time-slices. In addition, the temporal matching heuristic will be applied to the result of each training.

The first analyzed indicator is the Intertemporal Stability Index. Table 8.3 shows the results obtained for the models trained using the three-year time-slices. Each row shows one

of the trained algorithms, and each column shows the index values as the number of extracted topics varies between $k = 2$ and $k = 5$. From now on, we will use this same table structure to display the other indexes analyzed. For each algorithm, the index values show regularity as the number of topics varies. When comparing different algorithms, considerable differences are noted. The K-Means-based models obtain more unstable topics over time, while the Latent Dirichlet Allocation (LDA) and Dynamic Topic Modeling (DTM) models obtain stable topics over time. In the case of the Intertemporal Stability Index, our methodological proposal using DTM obtains the best results in each of the analyzed scenarios.

Tabla 8.3: Intertemporal Stability Index using cosine distance

Model	#Topics			
	2	3	4	5
K-Means	0.31	0.29	0.34	0.35
k-Shape	0.08	0.12	0.19	0.15
TS K-Means	0.31	0.29	0.34	0.31
LDA	0.07	0.05	0.04	0.02
DTM	0.00	0.00	0.00	0.00

Regarding the Intratemporal Similarity Index, the differences between DTM and the rest of the models are less notorious than in the previous case. This data can be seen in Table 8.4, where DTM obtains similar results to those obtained by K-Means and TS K-Means. Despite this, DTM obtains a better Intratemporal Similarity Index except for when $k = 2$ topics are extracted. Also, the best results for DTM are obtained by extracting $k = 3$ and $k = 4$ Topics.

Tabla 8.4: Intratemporal Similarity Index using cosine distance

Model	#Topics			
	2	3	4	5
K-Means	0.36	0.31	0.35	0.37
k-Shape	0.06	0.05	0.12	0.10
TS K-Means	0.36	0.31	0.35	0.31
LDA	0.03	0.12	0.15	0.14
DTM	0.32	0.40	0.42	0.37

The analysis of Topic Consistency, as presented in Table 8.5, shows that the DTM model has the highest Topic Consistency among all models tested. It is important to note that

Topic Consistency measures the regularity of patterns obtained by comparing days of the week so that the components of an urban activity topic should stay the same from one day to the next. This expected behavior is because the activities carried out in the city reflect the routine of the people who inhabit them. Unlike previous indices, this metric is calculated using cosine similarity instead of cosine distance. In this way, the more consistent the topics obtained, the indicator will be closer to one. However, it is worth mentioning that as the number of behavior patterns increases, topic consistency tends to decrease, but DTM exhibits a minor variation.

Tabla 8.5: Topic Consistency - Cosine Similarity

Model	#Topics			
	2	3	4	5
K-Means	0.82	0.81	0.76	0.73
k-Shape	0.97	0.97	0.91	0.92
TS K-Means	0.82	0.81	0.76	0.78
LDA	0.96	0.92	0.89	0.89
DTM	0.98	0.97	0.96	0.95

Finally, the analysis of Topic Smoothness, as displayed in Table 8.6, indicates that K-Means and Time-Series K-Means reach the highest values for this indicator among all models analyzed. This result indicates that patterns obtained from these methods exhibit significant variations between consecutive hours. This behavior is not expected to be observed in an Human Behavior Pattern where gradual changes are expected rather than drastic fluctuations from one hour to the next. Additionally, the remaining models produce results of similar magnitude, with DTM consistently outperforming in nearly all scenarios. It should be noted that as the number of urban patterns increases, the Topic Smoothness tends to rise.

Tabla 8.6: Topic Smoothness

Model	#Topics			
	2	3	4	5
K-Means	5.25	5.37	6.47	6.38
k-Shape	1.78	1.82	2.88	2.68
TS K-Means	5.25	5.37	6.47	5.64
LDA	1.93	2.51	2.95	2.88
DTM	1.78	2.00	2.24	2.58

In summary, after analyzing the results obtained in this research, we have determined that $k = 3$ is the most appropriate number of human behavior topics to represent the behavior of the cities included in this study. This decision is based on analyzing the Intertemporal Stability and Intratemporal Similarity. Also, it considers the trade-off of increasing the number of human behaviors that were observed when analyzing Topic Consistency and Topic Smoothness.

8.2.4. Final Model: Multi-sensor and multi-temporal city activity patterns from human behavior

Figure 8.2 shows the activity patterns arising from human behavior obtained after applying our proposed methodology to a geo-tagged urban activities dataset. The figure presents three clear columns denoting each extracted topic (human behavior patterns). Each graph row corresponds to the subset of three topics obtained for that particular time-slice. Each time-slice can be identified by the time range shown in the column corresponding to Topic 0. The first row is the topics obtained for the time-slice between 2005 and 2007, while the last row shows the urban activity patterns for 2020-2021. In the diagram, the urban activity patterns start on Monday and end on Sunday.

Next, we give an interpretation of the human behavior patterns

- **Topic 0** is characterized by behavior with certain regularity during the week, behavior that changes during the weekends. During the week, the activity of this pattern increases as the day progresses and presents two clearly defined peaks. The first activity peak is observed at noon and day and then descends to reach a local minimum around 15:00 hrs.; after this, the activity reaches its maximum peak around 09:00 hrs. This behavioral pattern, already observed in our previous research [49, 50], refers to Leisure & Commerce activities. During the weekends, both peaks are preserved. However, the noon peak is much more tenuous to give way to the more significant activity generated during the night peak.

- **Topic 1** shows a relatively low activity during the week, and the most significant activity occurs during the weekend. During the week, the activity increases between 09:00 hrs. and 21:00 hrs., without significant variations in activity during this period. During the weekend, the activity increases from 09:00 hrs., peaking at 15:00 hrs. and then declining.

This fall is more abrupt on Sunday, leaving little activity until dawn the next day.

- In the same way as the previous urban activity patterns. **Topic 2** shows different behaviors during the week and at the weekend. During the week, the activity is concentrated between 09:00 hrs. and 18:00 hrs., and a slight drop in activity around noon. During the weekend, the pattern presents a similar structure. However, after the 09:00 hrs. peak, the activity begins to decline during the rest of the day. This pattern is similar to the office-areas activity pattern detected in our previous investigations [49, 50].

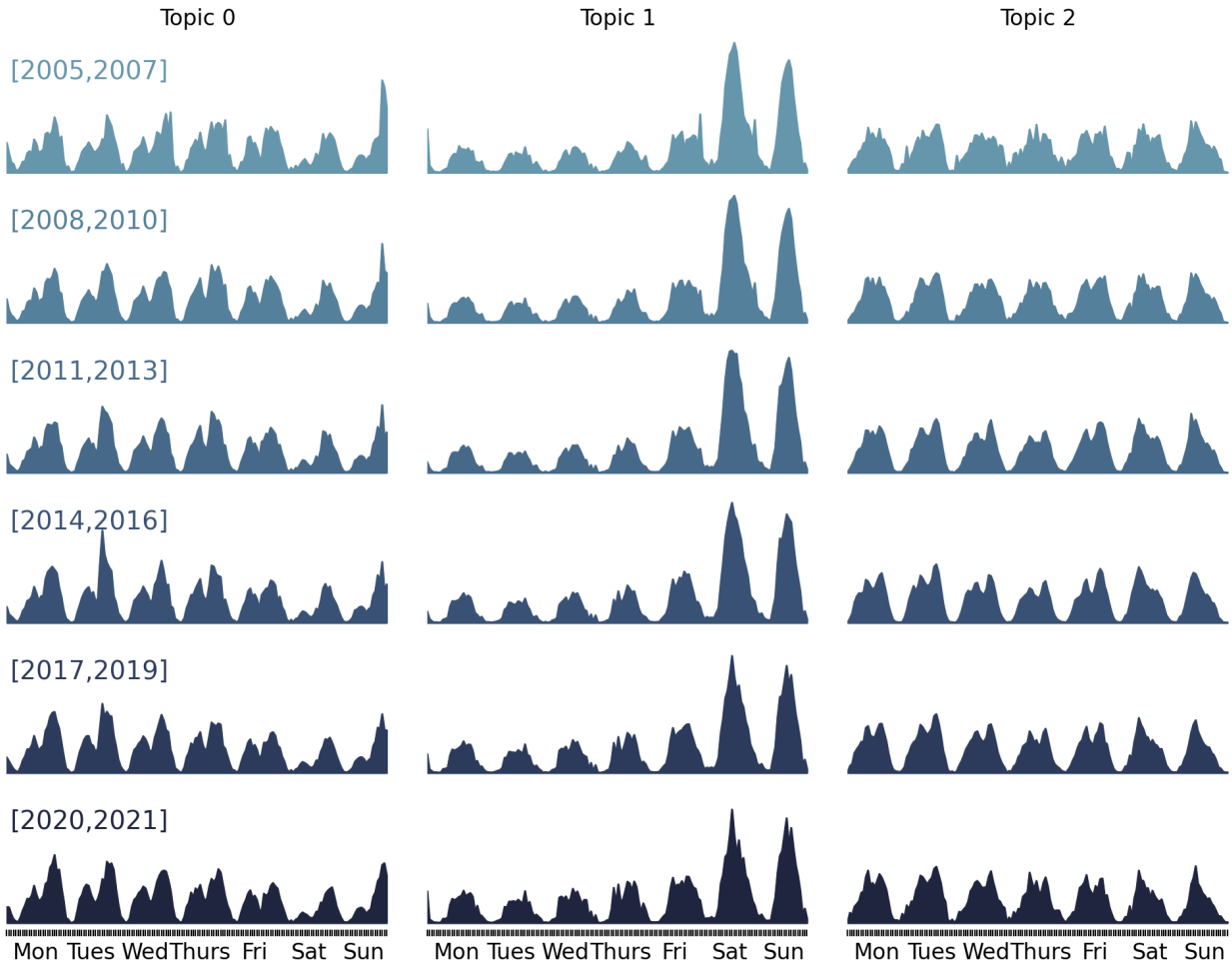


Figure 8.2: Multi-sensor and multi-temporal human behavior patterns obtained using Dynamic Topic Models

One of the advantages of using Dynamic Topic Models to study the temporal behavior of human activity patterns is the possibility of studying changes over time. Figure 8.3 displays the above topics, allowing us to compare their changes over time. The set of patterns obtained with each time-slice is displayed on the same figure, and the intensity of each line

color indicates the periods used for training, the darker the line, the more recent the data used. The first notable point is the stability of the patterns obtained, maintaining their structure over time. This stability is observed, even when the data used for training comes from multiple sensors, and some do not overlap between time-slices. Significant changes to note are the drop in activity on Friday, Saturday, and Sunday nights in Topic 1. Similar behavior has been observed during the last few years in Topic 1, with a noticeable drop in Saturday activity. Finally, a decrease in activity is observed in Topic 2 during the afternoons of Saturdays and Sundays, as well as in the activity observed on Wednesdays and Thursdays from 18:00 hrs. In future work, it is necessary to investigate the root cause of this behavior change. However, the most straightforward hypothesis is due to the changes in mobility produced by the restrictions established to control the covid 19 pandemic.

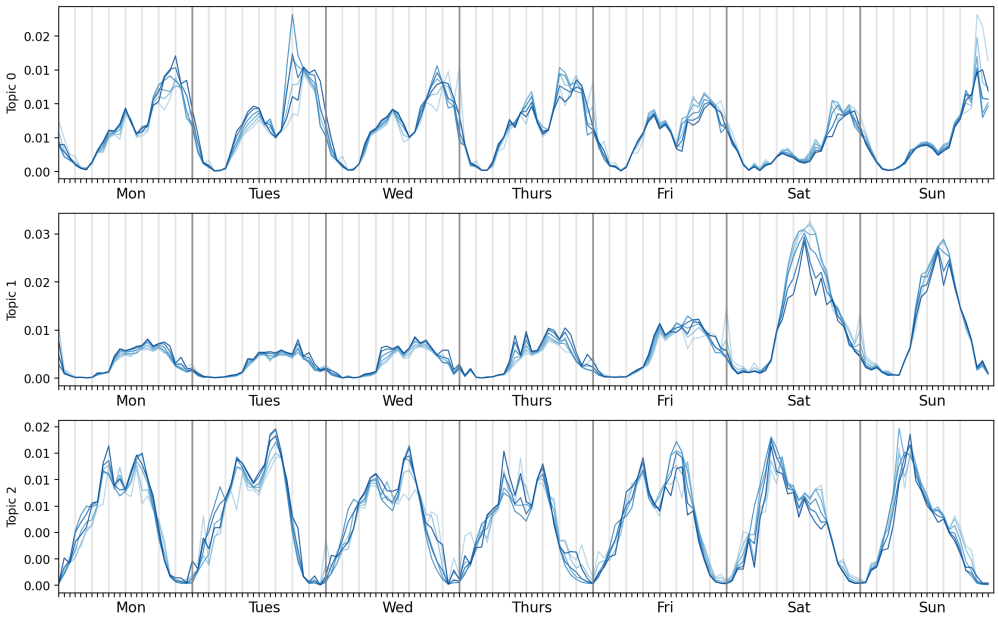


Figure 8.3: Temporal comparison of City activity patterns

8.2.5. Human behavior patterns Characterization

After describing each human activity pattern, it is interesting to understand how this information allows us to characterize the behavior of different cities worldwide. For this, we generated groups of cities based on how Activity Patterns' composition varies over time. To generate these groups, we used the K-Means clustering algorithm. The cities were used as input records for the model. Each city is characterized based on the average activity of

each human behavior pattern and according to the standard deviation of these. In order to determine the optimal number of clusters, the model was trained by varying the number of clusters between 2 and 20 groups. Figure 8.4 shows the average Silhouette Score obtained for each model. Based on these results, it is obtained that three groups correspond to the optimal number of clusters.

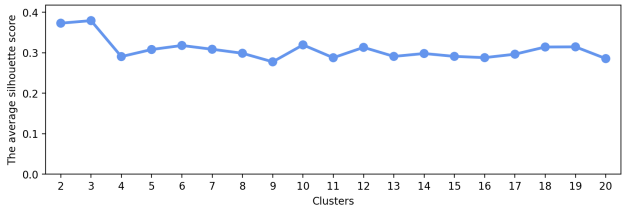


Figure 8.4: Silhouette Score over groups multi-temporal human behaviour patterns at city level

Figure 8.5 shows the result of clustering the cities based on their temporal composition of the human activity patterns. Each color indicates a different cluster. In addition, the name of some cities within each cluster is shown. Thus, it can be seen that within Cluster 0 (C0) are cities such as Campinas (Brazil), Lahore (Pakistan), Jeddah (Saudi Arabia), and Lagos (Nigeria). Cluster 1 (C1) includes cities such as Porto Alegre (Brazil), Athens (Greece), Ecatepec (Mexico), and Atlanta and Boston (USA). In Cluster 2 (C2), we find cities such as Seoul (South Korea), Santo Domingo (Dominica Republic), Nairobi (Kenya), Stockholm (Sweden), Washington (USA), and Birstall (England). As shown in the figure, the limits between each cluster are fuzzy, so a strict interpretation requires more elaboration. However, in general terms, we can understand how to segment and characterize cities based on their activity.

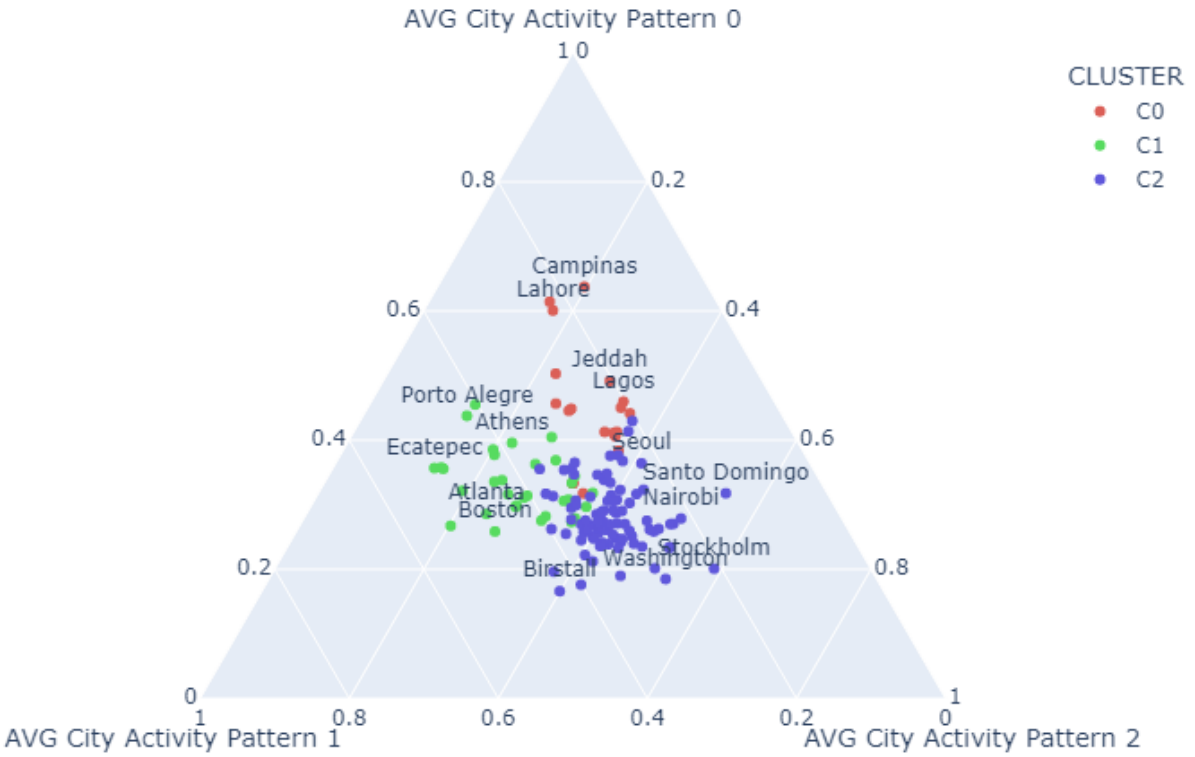


Figure 8.5: Cluster of cities based on their human behavior patterns composition

Figure 8.6 shows a geographical representation of the clusters. This figure shows the spatial distribution of the cities that belong to each cluster. The geographical location of Cluster 0 stands out, whose 20 cities are located mainly in the Middle East, South Asia, and Africa. The cities corresponding to cluster 1 and cluster 2 are distributed in practically the same territories, except that we did not find any of the 32 cities of Cluster 1 in East Asia and Oceania. Finally, Cluster 2 stands out for having several of its 92 cities in central Europe.

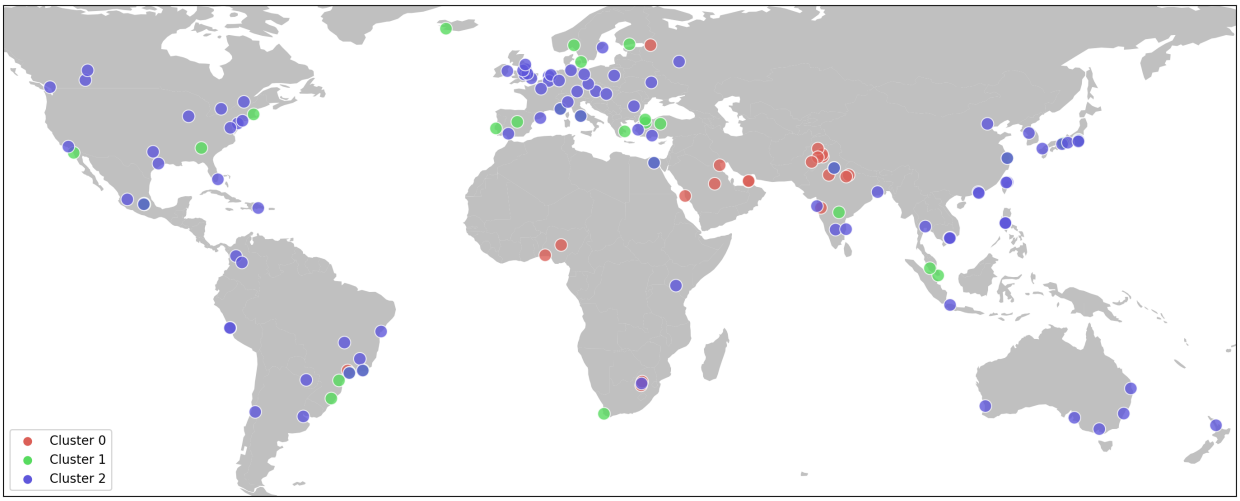


Figure 8.6: Geographical representation of cities clusters

To deepen the segmentation of cities obtained from their composition in the activity patterns, we use additional information that complements the cities analysis adding more details. The Innovation Cities Index [179], an annual quantitative index to rank the most innovative cities worldwide, is used for this. The quantitative index is based on cities' cultural assets, human infrastructure, and networked markets. Cultural assets refer to how culture is experienced within cities and considers arts districts, civic institutions, museums, music events, galleries, political protests, books, media, availability of information, and sports. Human Infrastructure includes the infrastructure deployed in the city for mass transit, finance, universities, hospitals, rail, roads, law, commerce, start-ups, healthcare, and telecommunications. Finally, Networked Markets measure a city's influence and connections in global markets, considering geography, economics, exports and imports, technology, market size, geo-political aspects, and diplomacy.

Figure 8.7 shows the innovation index ranking for 2021. The graph on the left corresponds to a boxplot where each data point corresponds to a city, indicating the position in the innovation ranking. The graph on the right corresponds to the cumulative distribution. When analyzing the figure, the difference between the cities that forms Cluster 0 to the rest of the cities stands out. The cities in Cluster 0 are in the last positions of the innovation ranking, and half of the cities in this cluster are in the last quintile of the innovation ranking. Based on this, we notice a relationship between the human behavior patterns detected in this study and how innovative a city is. Concerning the rest of the clusters, no significant differences are observed in the location of the cities in the ranking. In both groups, half of

the cities are within the first 150 most innovative cities.

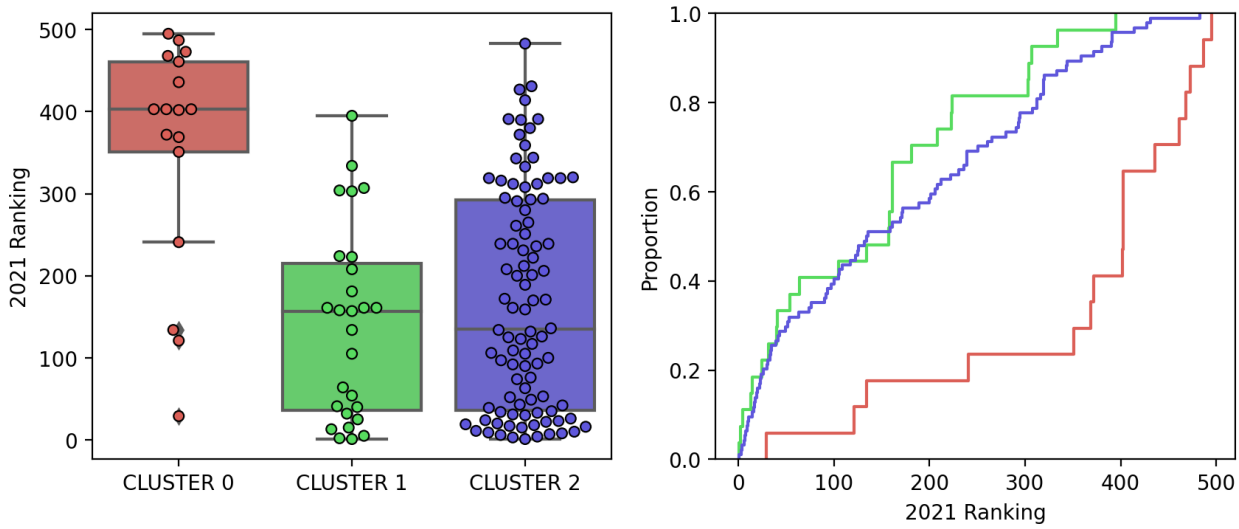
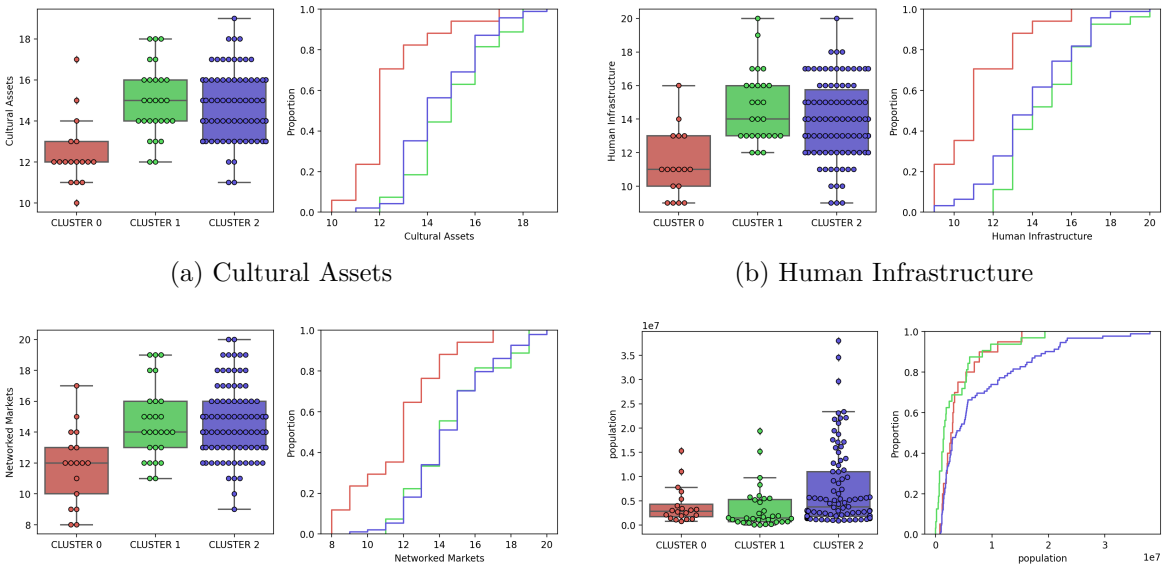


Figure 8.7: Innovation Cities Index 2021 by Cluster

Additionally, we present the factors that make up the Innovation Cities Index for each city under study. Figure 8.8 shows how the Cultural Assets (a), Human Infrastructure (b), and Networked Market (c) indexes are distributed. In addition, the population distribution (d) for each city is shown. The innovation index explanation by the three factors that make up the score does not show variations concerning what we already knew. The cities of Cluster 0 are different from the rest. In this case, these cities have less cultural capital, their infrastructure is also far from world standards, and their markets need to be sufficiently connected and integrated with the rest of the world. On the other hand, Cluster 1 and Cluster 2 show little differences between them when compared based on any of these three indicators. Where it makes present differences is at the population level. In this case, the cities of Cluster 0 and Cluster 1 present a very similar population distribution, while within the cities that belong to Cluster 2, there are some huge ones.



(a) Cultural Assets

(b) Human Infrastructure

(c) Networked Market

(d) City Population

Figure 8.8: Cities population and city score factors by Cluster

8.3. Experiment Conclusions

The analysis of urban activities is a central tool for use in urban planning, traffic management, and even in the design of public policies to prevent damage caused by natural disasters. For this reason, the study of urban activities is an important research topic. In this analysis, we proposed a methodology to address gaps detected in this area. We present a method to include temporal evolution in the problem of detecting activity patterns. Furthermore, we proposed four metrics that aim to reduce dependence on expert knowledge when selecting the set of patterns that best represents the observed activity behavior. The results confirm that the proposed methodology, particularly the Dynamic Topic Models, is an appropriate method to characterize urban mobility through human behavior patterns, obtaining better results when selecting the activity patterns against traditional methods like K-Means and Latent Dirichlet Allocation. The results showed that detecting human behavior patterns from a multi-city dataset where the aggregation level is an entire city is possible. The previous investigations considered single-city analyses, and the study level was conducted in neighborhoods, grids, or Voronoi zones within the same city. On the other hand, our methodology allows us to find a way to directly include in the algorithm the temporal evolution of the activity patterns. We used the proposed metrics to select the best representation of the

city's behavior. These metrics allow us to assess the best representation of activity patterns. The patterns obtained reflect people's behavior in the city, and two of the patterns obtained are similar to patterns already observed in previous single-city activity patterns research. Additionally, this study combines multiple sensors to gather urban activity data, which also were obtained at various instants over 17 years. Our methodology also proved to be a robust model that combines information from multiple sources and different timelines. Finally, the patterns detected are consistent with the proposed metrics and with a validation based on understanding the behavior shown by each pattern. In addition, it provides an alternative to the study of cities because it allows us to distinguish how innovative a city is from the behavior of its inhabitants. In future work, we want to study if there is a hierarchical relationship between the patterns that can be obtained from the single-cities analysis with other types of spatial aggregation, such as cities or countries. On the other hand, we want to delve into the mathematical properties of the proposed metrics, analyze them and test new scenarios for identifying activity patterns.

Chapter 9

Conclusions and Future Work

In this Chapter, we present the conclusions about the work developed in this thesis and also give possible future research lines.

9.1. Conclusions

As technology develops, it becomes increasingly present in people's lives. We interact with different technological devices daily while carrying out our daily activities. Many technological devices store the activities, leaving digital traces of individuals' behavior. A subset of these digital traces georeferences the location where the individual was while performing the activity. This information is very useful because it allows studying urban planning, infrastructure management, public transportation management, and public policies. This thesis addresses the study of multi-sensor and multi-temporal human behavior patterns from digital traces. For them, we set ourselves three objectives covered in three extensive studies presented as experiments in this thesis. In the first experiment, alternatives to traditional algorithms to identify patterns in digital traces were studied. LDA is proposed as an alternative that not only allows us to identify behavioral patterns but also allows us to recognize behaviors that traditional algorithms do not capture. Then, in the second experiment, the identification of spatiotemporal human behavior patterns is addressed by training multiple time-windowed spatial models. This analysis allows us to study the evolution of behavioral patterns over time, but it has some disadvantages because the pattern detection algorithm does not directly include the temporal dimension. Finally, in the third experiment, previous learning is used to formalize the validation of the human behavior pattern through a set of

metrics that aim to reduce dependency on extensive expert knowledge of the geographical area studied. In addition, we proposed a model that incorporates the temporal dimension to detect human behavior patterns. This model, Dynamic Topic Model, overperformed the traditional models and also LDA to detect spatiotemporal patterns. To develop these experiments, we used three digital traces datasets obtained from different sensors, call detail records, credit card purchases, and geo-tagged social media activity. In this way, our study becomes one of the few to study multi-sensor activity patterns and analyze these patterns over time. Finally, the proposed methodology and the data sets used allow us to respond to the objectives specified in this thesis and also allow us to extend the knowledge of the detection of human behavior patterns using multi-sensor and multi-temporal data.

9.2. Future Work

Throughout the study, we aggregated spatial information in various forms, using Voronoi zones in the telecom dataset analysis, grids for the banking dataset, and cities for the social media dataset. Regardless of the aggregation unit, similar behavioral patterns emerged despite being different sensors in different spatial aggregations. One line of future research is to analyze the impact of spatial aggregation on the patterns obtained and investigate whether there is any hierarchical dependency in these aggregations. On the other hand, our last experiment proposes a set of metrics to reduce dependence on extensive expert knowledge and formalize the validation of the patterns obtained. Future work will study the mathematical properties of these metrics and measure the impact of these metrics on the shape that the final patterns will have.

Bibliography

[1] Willis, A., Gjersoe, N., Havard, C., Kerridge, J., y Kukla, R., "Human movement behaviour in urban spaces: Implications for the design and modelling of effective pedestrian environments," Environment and Planning B: Planning and Design, vol. 31, no. 6, pp. 805–828, 2004.

[2] Wang, D., Gong, J., Chen, L., Zhang, L., Song, Y., y Yue, Y., "Spatio-temporal pattern analysis of land use/cover change trajectories in xihe watershed," International Journal of Applied Earth Observation and Geoinformation, vol. 14, no. 1, pp. 12–21, 2012, doi:10.1016/j.jag.2011.08.007.

[3] Valls, F. y Roca, J., "Visualizing digital traces for sustainable urban management: mapping tourism activity on the virtual public space," Sustainability, vol. 13, no. 6, p. 3159, 2021.

[4] Dembski, F., Wössner, U., Letzgus, M., Ruddat, M., y Yamu, C., "Urban digital twins for smart cities and citizens: The case study of herrenberg, germany," Sustainability, vol. 12, no. 6, 2020, doi:10.3390/su12062307.

[5] Wang, Q. y Taylor, J. E., "Quantifying human mobility perturbation and resilience in hurricane sandy," PLOS ONE, vol. 9, pp. 1–5, 2014, doi:10.1371/journal.pone.0112608.

[6] Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A., Canright, G., Engø-Monsen, K., y Bengtsson, L., "Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in bangladesh," Global Environmental Change, vol. 38, pp. 1–7, 2016, doi:10.1016/j.gloenvcha.2016.02.002.

[7] Yu, M., Yang, C., y Li, Y., "Big data in natural disaster management: A review," Geosciences, vol. 8, no. 5, 2018, doi:10.3390/geosciences8050165.

[8] Podesta, C., Coleman, N., Esmalian, A., Yuan, F., y Mostafavi, A., "Quantifying community resilience based on fluctuations in visits to points-of-interest derived from digital trace data," Journal of the Royal Society Interface, vol. 18, no. 177, p. 20210158, 2021.

[9] Fan, C., Zhang, C., Yahja, A., y Mostafavi, A., "Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management," International Journal of Information Management, vol. 56, p. 102049, 2021, doi:10.1016/j.ijinfomgt.2019.102049.

[10] Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., y Ukkusuri, S. V., "Understanding post-disaster population recovery patterns," Journal of The Royal Society Interface, vol. 17, no. 163, p. 20190532, 2020, doi:10.1098/rsif.2019.0532.

[11] Dargin, J. S., Fan, C., y Mostafavi, A., "Vulnerable populations and social media use in disasters: Uncovering the digital divide in three major u.s. hurricanes," International Journal of Disaster Risk Reduction, vol. 54, p. 102043, 2021, doi:10.1016/j.ijdrr.2021.102043.

[12] Farahmand, H., Wang, W., Mostafavi, A., y Maron, M., "Anomalous human activity fluctuations from digital trace data signal flood inundation status," Environment and Planning B: Urban Analytics and City Science, vol. 49, no. 7, pp. 1893–1911, 2022.

[13] Abdar, M., Basiri, M. E., Yin, J., Habibnezhad, M., Chi, G., Nemati, S., y Asadi, S., "Energy choices in alaska: Mining people's perception and attitudes from geotagged tweets," Renewable and Sustainable Energy Reviews, vol. 124, p. 109781, 2020, doi:10.1016/j.rser.2020.109781.

[14] Kubo, T., Uryu, S., Yamano, H., Tsuge, T., Yamakita, T., y Shirayama, Y., "Mobile phone network data reveal nationwide economic value of coastal tourism under climate change," Tourism Management, vol. 77, p. 104010, 2020, doi:10.1016/j.tourman.2019.104010.

[15] Milojevic-Dupont, N. y Creutzig, F., "Machine learning for geographically differentiated climate change mitigation in urban areas," Sustainable Cities and Society, vol. 64, p. 102526, 2021, doi:10.1016/j.scs.2020.102526.

[16] Sottini, V. A., Barbierato, E., Bernetti, I., y Capecchi, I., "Impact of climate change on

wine tourism: An approach through social media data," Sustainability, vol. 13, no. 13, 2021, doi:10.3390/su13137489.

[17] Funada, S. y Tsutsumida, N., "Mapping cherry blossoms from geotagged street-level photos," bioRxiv, 2022, doi:10.1101/2022.01.18.476550.

[18] Garcia, D. y Rimé, B., "Collective emotions and social resilience in the digital traces after a terrorist attack," Psychological Science, vol. 30, no. 4, pp. 617–628, 2019, doi:10.1177/0956797619831964. PMID: 30865565.

[19] Schafer, V., Truc, G., Badouard, R., Castex, L., y Musiani, F., "Paris and nice terrorist attacks: Exploring twitter and web archives," Media, War & Conflict, vol. 12, no. 2, pp. 153–170, 2019, doi:10.1177/1750635219839382.

[20] Bérubé, M., Tang, T.-U., Fortin, F., Ozalp, S., Williams, M. L., y Burnap, P., "Social media forensics applied to assessment of post–critical incident social reaction: The case of the 2017 manchester arena terrorist attack," Forensic Science International, vol. 313, p. 110364, 2020, doi:10.1016/j.forsciint.2020.110364.

[21] Ramadona, A. L., Tozan, Y., Lazuardi, L., y Rocklöv, J., "A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in yogyakarta, indonesia," PLOS Neglected Tropical Diseases, vol. 13, pp. 1–12, 2019, doi:10.1371/journal.pntd.0007298.

[22] Masri, S., Jia, J., Li, C., Zhou, G., Lee, M.-C., Yan, G., y Wu, J., "Use of twitter data to improve zika virus surveillance in the united states during the 2016 epidemic," BMC public health, vol. 19, no. 1, pp. 1–14, 2019.

[23] Tran, T. y Lee, K., "Understanding citizen reactions and ebola-related information propagation on social media," en 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 106–111, 2016, doi:10.1109/ASONAM.2016.7752221.

[24] Brugh, K. N., Lewis, Q., Haddad, C., Kumaresan, J., Essam, T., y Li, M. S., "Characterizing and mapping the spatial variability of hiv risk among adolescent girls and young women: A cross-county analysis of population-based surveys in eswatini, haiti, and mozambique," PLOS ONE, vol. 16, pp. 1–21, 2021, doi:10.1371/journal.pone.0261520.

[25] Abdallah, H. S., Khafagy, M. H., y Omara, F. A., "Case study: Spark gpu-enabled framework to control covid-19 spread using cell-phone spatio-temporal data," Computers, Materials & Continua, vol. 65, no. 2, pp. 1303–1320, 2020.

[26] Bisanzio, D., Kraemer, M. U., Brewer, T., Brownstein, J. S., y Reithinger, R., "Geolocated twitter social media data to describe the geographic spread of sars-cov-2," Journal of Travel Medicine, vol. 27, no. 5, p. taaa120, 2020.

[27] Luca, Massimiliano, Lepri, Bruno, Frias-Martinez, Enrique, y Lutu, Andra, "Modeling international mobility using roaming cell phone traces during covid-19 pandemic," EPJ Data Sci., vol. 11, no. 1, p. 22, 2022, doi:10.1140/epjds/s13688-022-00335-9.

[28] Cope, M., "Commentary: Geographies of digital lives: Trajectories in the production of knowledge with user-generated content," Landscape and Urban Planning, vol. 142, pp. 212–214, 2015, doi:10.1016/j.landurbplan.2015.08.009. Special Issue: Critical Approaches to Landscape Visualization.

[29] Saldana-Perez, M., Torres-Ruiz, M., y Moreno-Ibarra, M., "Geospatial modeling of road traffic using a semi-supervised regression algorithm," IEEE Access, vol. 7, pp. 177376–177386, 2019.

[30] Elleuch, W., Wali, A., y Alimi, A. M., "Towards an efficient traffic congestion prediction method based on neural networks and big gps data," IIUM Engineering Journal, vol. 20, p. 108–118, 2019, doi:10.31436/iiumej.v20i1.997.

[31] P, A. B. y Sumathi, R., "Data sources for urban traffic prediction: A review on classification, comparison and technologies," en 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), pp. 628–635, 2020, doi:10.1109/ICISS49785.2020.9316096.

[32] Montoya-Torres, J. R., Moreno, S., Guerrero, W. J., y Mejía, G., "Big data analytics and intelligent transportation systems," IFAC-PapersOnLine, vol. 54, no. 2, pp. 216–220, 2021.

[33] Salazar-Carrillo, J., Torres-Ruiz, M., Davis, C. A., Quintero, R., Moreno-Ibarra, M., y Guzmán, G., "Traffic congestion analysis based on a web-gis and data mining of traffic events from twitter," Sensors, vol. 21, no. 9, 2021, doi:10.3390/s21092964.

[34] Goh, G., Koh, J. Y., y Zhang, Y., "Twitter-informed crowd flow prediction," en 2018 IEEE International Conference on Data Mining Workshops (ICDMW), pp. 624–631, IEEE, 2018.

[35] Zhao, Y., Li, J., Miao, X., y Ding, X., "Urban crowd flow forecasting based on cellular network," en Proceedings of the ACM Turing Celebration Conference - China, ACM TURC '19, (New York, NY, USA), Association for Computing Machinery, 2019, doi:10.1145/3321408.3321579.

[36] Ebrahimpour, Z., Wan, W., Cervantes, O., Luo, T., y Ullah, H., "Comparison of main approaches for extracting behavior features from crowd flow analysis," ISPRS International Journal of Geo-Information, vol. 8, no. 10, p. 440, 2019.

[37] Terroso-Saenz, F., Flores, R., y Muñoz, A., "Human mobility forecasting with region-based flows and geotagged twitter data," Expert Systems with Applications, vol. 203, p. 117477, 2022, doi:10.1016/j.eswa.2022.117477.

[38] Perola, E., Todorovic, S., Muukkonen, P., y Järv, O., "Exploratory visual methods to aggregate origin-destination geodata," Examples and progress in geodata science, 2020.

[39] Waller, S. T., Chand, S., Zlojutro, A., Nair, D., Niu, C., Wang, J., Zhang, X., y Dixit, V. V., "Rapidex: A novel tool to estimate origin–destination trips using pervasive traffic data," Sustainability, vol. 13, no. 20, 2021, doi:10.3390/su132011171.

[40] Graff, M., Moctezuma, D., Miranda-Jiménez, S., y Tellez, E. S., "A python library for exploratory data analysis on twitter data based on tokens and aggregated origin–destination information," Computers & Geosciences, vol. 159, p. 105012, 2022, doi:10.1016/j.cageo.2021.105012.

[41] Srinon, R., "Smart mobility systems planning: Incorporating customer insights to enhance local experience," en Workshop proceedings, p. 125, 2018.

[42] Chen, P., Shi, W., Zhou, X., Liu, Z., y Fu, X., "Stlp-gsm: a method to predict future locations of individuals based on geotagged social media data," International Journal of Geographical Information Science, vol. 33, no. 12, pp. 2337–2362, 2019, doi:10.1080/13658816.2019.1630630.

[43] Fan, T., Guo, N., y Ren, Y., "Consumer clusters detection with geo-tagged social network data using dbscan algorithm: a case study of the pearl river delta in china," GeoJournal, vol. 86, no. 1, pp. 317–337, 2021.

[44] Miah, S. J., Vu, H. Q., y Gammack, J. G., "A location analytics method for the utilisation of geotagged photos in travel marketing decision-making," Journal of Information & Knowledge Management, vol. 18, no. 01, p. 1950004, 2019, doi:10.1142/S0219649219500047.

[45] Pachni-Tsitiridou, O. y Fouskas, K., "Location-aware technologies: How they affect customer experience," en Strategic Innovative Marketing and Tourism (Kavoura, A., Kefallonitis, E., y Giovanis, A., eds.), (Cham), pp. 1199–1206, Springer International Publishing, 2019.

[46] Jiang, W., Wang, Y., Dou, M., Liu, S., Shao, S., y Liu, H., "Solving competitive location problems with social media data based on customers' local sensitivities," ISPRS International Journal of Geo-Information, vol. 8, no. 5, 2019, doi:10.3390/ijgi8050202.

[47] Ferro-Díez, L. E., Villegas, N. M., y Díaz-cely, J., "Location data analytics in the business value chain: A systematic literature review," IEEE Access, vol. 8, pp. 204639–204659, 2020, doi:10.1109/ACCESS.2020.3036835.

[48] Fayyad, U., Piatetsky-Shapiro, G., y Smyth, P., "The kdd process for extracting useful knowledge from volumes of data," Communications of the ACM, vol. 39, no. 11, pp. 27–34, 1996.

[49] Muñoz-Cancino, R., Rios, S. A., Goic, M., y Graña, M., "Non-intrusive assessment of covid-19 lockdown follow-up and impact using credit card information: Case study in chile," International Journal of Environmental Research and Public Health, vol. 18, no. 11, 2021, doi:10.3390/ijerph18115507.

[50] Ríos, S. A. y Muñoz, R., "Land use detection with cell phone data using topic models: Case santiago, chile," Computers, Environment and Urban Systems, vol. 61, pp. 39–48, 2017.

[51] Muñoz-Cancino, R., Bravo, C., Ríos, S. A., y Graña, M., "On the combination of graph data for assessing thin-file borrowers' creditworthiness," Expert Systems with Applications, vol. 213, p. 118809, 2023, doi:10.1016/j.eswa.2022.118809.

[52] Muñoz-Cancino, R., Bravo, C., Ríos, S. A., y Graña, M., "On the dynamics of credit history and social interaction features, and their impact on creditworthiness assessment performance," Expert Systems with Applications, vol. 218, p. 119599, 2023, doi:10.1016/j.eswa.2023.119599.

[53] Muñoz-Cancino, R., Bravo, C., Ríos, S. A., y Graña, M., "Assessment of creditworthiness models privacy-preserving training with synthetic data," en Hybrid Artificial Intelligent Systems, (Cham), pp. 375–384, Springer International Publishing, 2022.

[54] BankMyCell, "How many smartphones are in the world?," 2022. Retrieved from https://www.bankmycell.com/blog/how-many-phones-are-in-the-world. Accessed October 13, 2022.

[55] Zhao, K., Tarkoma, S., Liu, S., y Vo, H., "Urban human mobility data mining: An overview," en 2016 IEEE International Conference on Big Data (Big Data), pp. 1911–1920, IEEE, 2016.

[56] Hudson-Smith, A., Batty, M., Crooks, A., y Milton, R., "Mapping for the masses: Accessing web 2.0 through crowdsourcing," Social science computer review, vol. 27, no. 4, pp. 524–538, 2009.

[57] Goetz, M. y Zipf, A., "The evolution of geo-crowdsourcing: bringing volunteered geographic information to the third dimension," en Crowdsourcing geographic knowledge, pp. 139–159, Springer, 2013.

[58] Goodchild, M. F., "Citizens as sensors: the world of volunteered geography," GeoJournal, vol. 69, no. 4, pp. 211–221, 2007.

[59] Granell, C. y Ostermann, F. O., "Beyond data collection: Objectives and methods of research using vgi and geo-social media for disaster management," Computers, Environment and Urban Systems, vol. 59, pp. 231–243, 2016, doi:10.1016/j.compenvurbsys.2016.01.006.

[60] Fujisaka, T., Lee, R., y Sumiya, K., "Exploring urban characteristics using movement history of mass mobile microbloggers," en Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, HotMobile '10, (New York, NY, USA), pp. 13–18, ACM, 2010, doi:10.1145/1734583.1734588.

[61] Wakamiya, S., Lee, R., y Sumiya, K., "Urban area characterization based on semantics of crowd activities in twitter," en Proceedings of the 4th international conference on GeoSpatial semantics, GeoS'11, (Berlin, Heidelberg), pp. 108–123, Springer-Verlag, 2011, http://dl.acm.org/citation.cfm?id=2008664.2008674.

[62] Noulas, A., Scellato, S., Mascolo, C., y Pontil, M., "Exploiting semantic annotations for clustering geographic areas and users in location-based social networks.," en The Social Mobile Web, vol. WS-11-02 de AAAI Workshops, AAAI, 2011.

[63] Crandall, D. J., Backstrom, L., Huttenlocher, D., y Kleinberg, J., "Mapping the world's photos," en Proceedings of the 18th international conference on World wide web, WWW '09, (New York, NY, USA), pp. 761–770, ACM, 2009, doi:10.1145/1526709.1526812.

[64] Frias-Martinez, V., Soto, V., Hohwald, H., y Frias-Martinez, E., "Characterizing urban landscapes using geolocated tweets.," en SocialCom/PASSAT, pp. 239–248, IEEE, 2012.

[65] Frias-Martinez, V., Soto, V., Hohwald, H., y Frias-Martinez, E., "Sensing urban land use with twitter activity." Preprint submitted to Elsevier, 2014.

[66] Pushkarev, B. S., Urban Space for Pedestrians: A Quantitative Approach. The MIT Press: Cambridge, MA, USA, 1976.

[67] Whyte, W. H., The social life of small urban spaces. Conservation Foundation Washington, DC, 1980.

[68] Back, A. y Marjavaara, R., "Mapping an invisible population: the uneven geography of second-home tourism," Tourism Geographies, vol. 19, no. 4, pp. 595–611, 2017.

[69] Reades, J., Zhong, C., Manley, E., Milton, R., y Batty, M., "Finding pearls in london's oysters," Built Environment, vol. 42, no. 3, pp. 365–381, 2016.

[70] Reades, J., Calabrese, F., Sevtsuk, A., y Ratti, C., "Cellular census: Explorations in urban data collection," IEEE Pervasive computing, vol. 6, no. 3, pp. 30–38, 2007.

[71] Gonzalez, M. C., Hidalgo, C. A., y Barabasi, A.-L., "Understanding individual human mobility patterns," Nature, vol. 453, pp. 779–782, 2008, doi:10.1038/nature06958.

[72] Mohammadi, A., Karimzadeh, S., Valizadeh Kamran, K., y Matsuoka, M., "Extraction

of land information, future landscape changes and seismic hazard assessment: A case study of tabriz, iran," Sensors, vol. 20, no. 24, p. 7010, 2020.

[73] Li, K., Feng, M., Biswas, A., Su, H., Niu, Y., y Cao, J., "Driving factors and future prediction of land use and cover change based on satellite remote sensing data by the lcm model: A case study from gansu province, china," Sensors, vol. 20, no. 10, p. 2757, 2020.

[74] Lehmann, A. y Gross, A., "Towards vehicle emission estimation from smartphone sensors," en 2017 18th IEEE International Conference on Mobile Data Management (MDM), pp. 154–163, IEEE, 2017.

[75] Alam, M. S., Duffy, P., Hyde, B., y McNabola, A., "Downscaling national road transport emission to street level: A case study in dublin, ireland," Journal of Cleaner Production, vol. 183, pp. 797–809, 2018.

[76] Krause, J., Small, M. J., Haas, A., y Jaeger, C. C., "An expert-based bayesian assessment of 2030 german new vehicle co2 emissions and related costs," Transport Policy, vol. 52, pp. 197–208, 2016.

[77] Kraemer, M. U., Bisanzio, D., Reiner, R., Zakar, R., Hawkins, J. B., Freifeld, C. C., Smith, D. L., Hay, S. I., Brownstein, J. S., y Perkins, T. A., "Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from lahore, pakistan," EPJ Data Science, vol. 7, pp. 1–17, 2018.

[78] Karimifar, M. J., Sikarudi, M. K., Moradi, E., y Bidkhori, M., Competitive Location Problem, pp. 271–294. Heidelberg: Physica-Verlag HD, 2009, doi:10.1007/978-3-7908-2151-2_12.

[79] Cakmakli, C., Demiralp, S., Ozcan, S. K., Yesiltas, S., y Yildirim, M. A., "COVID-19 and emerging markets: The case of turkey," Rep. Tec. 2011, Koc University-TUSIAD Economic Research Forum., 2020.

[80] Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., Iosifidis, C., y Agha, R., "World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19)," International Journal of Surgery, vol. 76, pp. 71–76, 2020, doi:10.1016/j.ijsu.2020.02.034.

[81] Yan, Y., Shin, W. I., Pang, Y. X., Meng, Y., Lai, J., You, C., Zhao, H., Lester, E., Wu, T., y Pang, C. H., "The first 75 days of novel coronavirus (sars-cov-2) outbreak: Recent advances, prevention, and treatment," International Journal of Environmental Research and Public Health, vol. 17, no. 7, 2020, doi:10.3390/ijerph17072323.

[82] Barański, K., Brożek, G., Kowalska, M., Kaleta-Pilarska, A., y Zejda, J. E., "Impact of covid-19 pandemic on total mortality in poland," International Journal of Environmental Research and Public Health, vol. 18, no. 8, 2021, doi:10.3390/ijerph18084388.

[83] Emmerich, F. G., "Comparisons between the neighboring states of amazonas and pará in brazil in the second wave of covid-19 outbreak and a possible role of early ambulatory treatment," International Journal of Environmental Research and Public Health, vol. 18, no. 7, 2021, doi:10.3390/ijerph18073371.

[84] Jindal, C., Kumar, S., Sharma, S., Choi, Y. M., y Efird, J. T., "The prevention and management of covid-19: Seeking a practical and timely solution," International Journal of Environmental Research and Public Health, vol. 17, no. 11, 2020, doi:10.3390/ijerph17113986.

[85] Frediansyah, A., Tiwari, R., Sharun, K., Dhama, K., y Harapan, H., "Antivirals for covid-19: A critical review," Clinical Epidemiology and Global Health, vol. 9, pp. 90–98, 2021, doi:10.1016/j.cegh.2020.07.006.

[86] Lagier, J.-C., Million, M., Gautret, P., Colson, P., Cortaredona, S., Giraud-Gatineau, A., Honoré, S., Gaubert, J.-Y., Fournier, P.-E., Tissot-Dupont, H., *et al.*, "Outcomes of 3,737 covid-19 patients treated with hydroxychloroquine/azithromycin and other regimens in marseille, france: A retrospective analysis," Travel medicine and infectious disease, vol. 36, p. 101791, 2020.

[87] Gentile, I., Maraolo, A. E., Piscitelli, P., y Colao, A., "Covid-19: Time for post-exposure prophylaxis?," International Journal of Environmental Research and Public Health, vol. 17, no. 11, 2020, doi:10.3390/ijerph17113997.

[88] Zhou, Y., Hou, Y., Shen, J., Mehra, R., Kallianpur, A., Culver, D. A., Gack, M. U., Farha, S., Zein, J., Comhair, S., Fiocchi, C., Stappenbeck, T., Chan, T., Eng, C., Jung, J. U., Jehi, L., Erzurum, S., y Cheng, F., "A network medicine approach to investigation and population-based validation of disease manifestations and drug repurposing for

covid-19," PLOS Biology, vol. 18, pp. 1–43, 2020, doi:10.1371/journal.pbio.3000970.

[89] Mishra, S. K. y Tripathi, T., "One year update on the covid-19 pandemic: Where are we now?," Acta Tropica, vol. 214, p. 105778, 2021, doi:10.1016/j.actatropica.2020.105778.

[90] Liu, Y., Morgenstern, C., Kelly, J., Lowe, R., y Jit, M., "The impact of non-pharmaceutical interventions on sars-cov-2 transmission across 130 countries and territories.," BMC Med, vol. 19, p. 40, 2021, doi:10.1186/s12916-020-01872-8.

[91] Perra, N., "Non-pharmaceutical interventions during the covid-19 pandemic: A review.," Phys Rep, 2021, doi:10.1016/j.physrep.2021.02.001.

[92] Regmi, K. y Lwin, C. M., "Factors associated with the implementation of non-pharmaceutical interventions for reducing coronavirus disease 2019 (covid-19): A systematic review," International Journal of Environmental Research and Public Health, vol. 18, no. 8, 2021, doi:10.3390/ijerph18084274.

[93] Trabelsi, K., Ammar, A., Masmoudi, L., Boukhris, O., Chtourou, H., Bouaziz, B., Brach, M., Bentlage, E., How, D., Ahmed, M., Mueller, P., Mueller, N., Hsouna, H., Elghoul, Y., Romdhani, M., Hammouda, O., Paineiras-Domingos, L. L., Braakman-Jansen, A., Wrede, C., Bastoni, S., Pernambuco, C. S., Mataruna-Dos-Santos, L. J., Taheri, M., Irandoust, K., Bragazzi, N. L., Strahler, J., Washif, J. A., Andreeva, A., Bailey, S. J., Acton, J., Mitchell, E., Bott, N. T., Gargouri, F., Chaari, L., Batatia, H., Khoshnami, S. C., Samara, E., Zisi, V., Sankar, P., Ahmed, W. N., Ali, G. M., Abdelkarim, O., Jarraya, M., Abed, K. E., Moalla, W., Souissi, N., Aloui, A., Souissi, N., Gemert-Pijnen, L. V., Riemann, B. L., Riemann, L., Delhey, J., Gómez-Raja, J., Epstein, M., Sanderman, R., Schulz, S., Jerg, A., Al-Horani, R., Mansi, T., Dergaa, I., Jmail, M., Barbosa, F., Ferreira-Santos, F., Šimunič, B., Pišot, R., Pišot, S., Gaggioli, A., Steinacker, J., Zmijewski, P., Apfelbacher, C., Glenn, J. M., Khacharem, A., Clark, C. C., Saad, H. B., Chamari, K., Driss, T., Hoekelmann, A., y on behalf of the ECLB-COVID19 Consortium, "Sleep quality and physical activity as predictors of mental wellbeing variance in older adults during covid-19 lockdown: Eclb covid-19 international online survey," International Journal of Environmental Research and Public Health, vol. 18, no. 8, 2021, doi:10.3390/ijerph18084329.

[94] Aguilar-Farias, N., Toledo-Vargas, M., Miranda-Marquez, S., Cortinez-O'Ryan, A.,

Cristi-Montero, C., Rodriguez-Rodriguez, F., Martino-Fuentealba, P., Okely, A. D., y del Pozo Cruz, B., "Sociodemographic predictors of changes in physical activity, screen time, and sleep among toddlers and preschoolers in chile during the covid-19 pandemic," International Journal of Environmental Research and Public Health, vol. 18, no. 1, 2021, doi:10.3390/ijerph18010176.

[95] Salazar-Fernández, C., Palet, D., Haeger, P. A., y Román Mella, F., "Covid-19 perceived impact and psychological variables as predictors of unhealthy food and alcohol consumption trajectories: The role of gender and living with children as moderators," International Journal of Environmental Research and Public Health, vol. 18, no. 9, 2021, doi:10.3390/ijerph18094542.

[96] Bermejo-Martins, E., Luis, E. O., Sarrionandia, A., Martínez, M., Garcés, M. S., Oliveros, E. Y., Cortés-Rivera, C., Belintxon, M., y Fernández-Berrocal, P., "Different responses to stress, health practices, and self-care during covid-19 lockdown: A stratified analysis," International Journal of Environmental Research and Public Health, vol. 18, no. 5, 2021, doi:10.3390/ijerph18052253.

[97] Mayen Huerta, C. y Cafagna, G., "Snapshot of the use of urban green spaces in mexico city during the covid-19 pandemic: A qualitative study," International Journal of Environmental Research and Public Health, vol. 18, no. 8, 2021, doi:10.3390/ijerph18084304.

[98] Carvalho, V., Garcia, J., Hansen, S., Ortiz, A., Rodrigo, T., y More, J., "Tracking the covid-19 crisis with high-resolution transaction data," rep. tec., Cambridge Working Papers in Economics 2030, Faculty of Economics, University of Cambridge., 2020.

[99] Horvath, A., Kay, B., y Wix, C., "The covid-19 shock and consumer credit: Evidence from credit card data," SSRN, 2021, doi:10.2139/ssrn.3613408.

[100] Dunn, A., Hood, K., y Driessen, A., Measuring the effects of the COVID-19 pandemic on consumer spending using card transaction data. US Department of Commerce, Bureau of Economic Analysis, 2020.

[101] Chen, H., Qian, W., y Wen, Q., "The impact of the covid-19 pandemic on consumption: Learning from high frequency transaction data (july 1, 2020)," rep. tec., Available at SSRN, 2020.

[102] Bounie, D., Camara, Y., Fize, E., Galbraith, J., Landais, C., Lavest, C., Pazem, T., y Savatier, Baptiste, C. D. P., "Consumption dynamics in the covid crisis: Real time insights from french transaction & bank data," Rep. Tec. 15474, CEPR Discussion Papers, 2020.

[103] Andersen, A. L., Hansen, E. T., Johannesen, N., y Sheridan, A., "Pandemic, shutdown and consumer spending: Lessons from scandinavian policy responses to covid-19," 2020.

[104] Chen, H., Felt, M.-H., y Huynh, K. P., "Retail payment innovations and cash usage: accounting for attrition by using refreshment samples," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 180, no. 2, pp. 503–530, 2017.

[105] Arraño, E. y Cova, J. P., "Evolución de los medios de pago en chile y su incidencia en el comportamiento de los componentes m1," Studies in Economic Statistics of the Central Bank of Chile, vol. 125, pp. 1–48, 2018.

[106] Wigginton, C., Curran, M., y Brodeur, C., Global mobile consumer trends: Second edition, 2017. https://www2.deloitte.com/us/en/pages/technology-media-and-telecommunications/articles/global-mobile-consumer-trends.html.

[107] Munizaga, M. A. y Palma, C., "Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile," Transportation Research Part C: Emerging Technologies, vol. 24, pp. 9 – 18, 2012, doi:10.1016/j.trc.2012.01.007.

[108] Madhawa, K., Lokanathan, S., Maldeniya, D., y Samarajiva, R., "Land use classification using call detail records," en NetMob 2015 (Fourth Conference on the Scientific Analysis of Mobile Phone Datasets), MIT Media Lab, Cambridge, USA, 8-10 April 201, 2015, doi:10.13140/RG.2.1.4057.7121.

[109] Soto, V. y Frías-martínez, E., "Robust land use characterization of urban landscapes using cell phone data," 2011.

[110] Lenormand, M., Picornell, M., Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., y Ramasco, J. J., "Cross-checking different sources of mobility information," PLoS ONE, vol. 9, pp. 1–10, 2014, doi:10.1371/journal.pone.0105184.

[111] Lenormand, Picornell, Cantú-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frías-Martínez, E., y Ramasco, J. J., "Influence of sociodemographic characteristics on human mobility," Scientific reports, vol. 5, p. 10075, 2015.

[112] Wang, Y., Wang, T., Tsou, M.-H., Li, H., Jiang, W., y Guo, F., "Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (sina-weibo) and commercial points of interest collections in beijing, china," Sustainability, vol. 8, p. 1202, 2016, doi:10.3390/su8111202.

[113] Du, H. y Mulley, C., "The short-term land value impacts of urban rail transit: Quantitative evidence from sunderland, uk," Land Use Policy, vol. 24, no. 1, pp. 223–233, 2007.

[114] Dong, H., Wu, M., Ding, X., Chu, L., Jia, L., Qin, Y., y Zhou, X., "Traffic zone division based on big data from mobile phone base stations," Transportation Research Part C: Emerging Technologies, vol. 58, pp. 278–291, 2015.

[115] Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., y Monterola, C., "Impacts of land use and amenities on public transport use, urban planning and design," Land Use Policy, vol. 57, pp. 356 – 367, 2016, doi:10.1016/j.landusepol.2016.06.004.

[116] Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.-L., Li, T., y Zhou, C., "A new insight into land use classification based on aggregated mobile phone data," International Journal of Geographical Information Science, vol. 28, no. 9, pp. 1988–2007, 2014.

[117] Bonnel, P., Fekih, M., y Smoreda, Z., "Origin-destination estimation using mobile network probe data," Transportation Research Procedia, vol. 32, pp. 69 – 81, 2018, doi:10.1016/j.trpro.2018.10.013.

[118] Graells-Garrido, E., Caro, D., y Parra, D., "Inferring modes of transportation using mobile phone data," EPJ Data Science, vol. 7, 2018, doi:10.1140/epjds/s13688-018-0177-1.

[119] Hong, L., Lee, M., Mashhadi, A., y Frias-Martinez, V., "Towards understanding communication behavior changes during floods using cell phone data," en Social Informatics (Staab, S., Koltsova, O., y Ignatov, D. I., eds.), (Cham), pp. 97–107, Springer International Publishing, 2018.

[120] Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., y Pradhan, B., "Urban flood risk mapping using the garp and quest models: A comparative study of machine learning techniques," Journal of Hydrology, vol. 569, pp. 142 – 154, 2019, doi:10.1016/j.jhydrol.2018.12.002.

[121] Dasgupta, A., "Floods and poverty traps: Evidence from bangladesh," Economic and Political Weekly, pp. 3166–3171, 2007.

[122] Akbar Ali, A., "Detecting the development of land use patterns for building in urban areas by using high resolution image," Jurnal Tataloka, vol. 15, p. 160, 2013, doi:10.14710/tataloka.15.3.160-174.

[123] Golap, A., Sarma, P., y Akturuzzaman, M., "Land use changing pattern detection and analysis in mymensingh district: A gis analysis," International Journal of Science and Research (IJSR), vol. 7, pp. 1823–1826, 2018.

[124] Yao, Y., Liang, H., Li, X., Zhang, J., y He, J., "Sensing urban land-use patterns by integrating google tensorflow and scene-classification models," CoRR, vol. abs/1708.01580, 2017, http://arxiv.org/abs/1708.01580.

[125] Gao, S., Rao, J., Kang, Y., Liang, Y., y Kruse, J., "Mapping county-level mobility pattern changes in the united states in response to covid-19," Available at SSRN 3570145, 2020.

[126] Lang, W., Long, Y., y Chen, T., "Rediscovering chinese cities through the lens of land-use patterns," Land Use Policy, vol. 79, pp. 362 – 374, 2018, doi:10.1016/j.landusepol.2018.08.031.

[127] Brockmann, D., Hufnagel, L., y Geisel, T., "The scaling laws of human travel," Scientific reports, vol. 1, 2006.

[128] Gonzalez, M. C., Hidalgo, C. A., y Barabasi, A.-L., "Understanding individual human mobility patterns," Nature, vol. 453, pp. 779–782, 2008, doi:10.1038/nature06958.

[129] Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., y Gonzalez, M. C., "Sequences of purchases in credit card data reveal life styles in urban populations," Nature Communications, vol. 9, 2018, doi:10.1038/s41467-018-05690-8.

[130] Lenormand, M., Picornell, M., Cantú-Ros, O. G., Louail, T., Herranz, R., Barthelemy,

M., Frías-Martínez, E., San Miguel, M., y Ramasco, J. J., "Comparing and modelling land use organization in cities," Royal Society Open Science, vol. 2, no. 12, p. 150449, 2015, doi:10.1098/rsos.150449.

[131] Luca, M., Barlacchi, G., Oliver, N., y Lepri, B., "Leveraging mobile phone data for migration flows," CoRR, vol. abs/2105.14956, 2021, https://arxiv.org/abs/2105.14956.

[132] Jacques, D. C., "Mobile phone metadata for development," 2018, doi:10.48550/ARXIV.1806.03086.

[133] Blondel, V. D., Decuyper, A., y Krings, G., "A survey of results on mobile phone datasets analysis," EPJ data science, vol. 4, no. 1, p. 10, 2015.

[134] Aurenhammer, F., Klein, R., y Lee, D.-T., Voronoi diagrams and Delaunay triangulations. World Scientific Publishing Company, 2013.

[135] Maps, S., World Cities Database, 2021. Retrieved from https://simplemaps.com/data/world-cities. Accessed September 3, 2021.

[136] Cho, E., Myers, S. A., y Leskovec, J., "Friendship and mobility: User movement in location-based social networks," en Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, (New York, NY, USA), p. 1082–1090, Association for Computing Machinery, 2011, doi:10.1145/2020408.2020579.

[137] Le Falher, G., Gionis, A., y Mathioudakis, M., "Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities," en 9th AAAI Conference on Web and Social Media - ICWSM 2015, (Oxford, United Kingdom), 2015, https://hal.archives-ouvertes.fr/hal-01134117.

[138] Yang, D., Zhang, D., Zheng, V. W., y Yu, Z., "Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 45, no. 1, pp. 129–142, 2015, doi:10.1109/TSMC.2014.2327053.

[139] Mousselly-Sergieh, H., Watzinger, D., Huber, B., Döller, M., Egyed-Zsigmond, E., y Kosch, H., "World-wide scale geotagged image dataset for automatic image annotation and reverse geotagging," en Proceedings of the 5th ACM Multimedia Systems

Conference, MMSys '14, (New York, NY, USA), p. 47–52, Association for Computing Machinery, 2014, doi:10.1145/2557642.2563673.

[140] Cheng, Z., Caverlee, J., y Lee, K., "You are where you tweet: A content-based approach to geo-locating twitter users," en Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10, (New York, NY, USA), p. 759–768, Association for Computing Machinery, 2010, doi:10.1145/1871437.1871535.

[141] Lamsal, R., "Design and analysis of a large-scale covid-19 tweets dataset," Applied Intelligence, vol. 51, no. 5, pp. 2790–2804, 2021.

[142] Kejriwal, M. y Melotte, S., "A Geo-Tagged COVID-19 Twitter Dataset for 10 North American Metropolitan Areas," 2021, doi:10.5281/zenodo.4434972.

[143] Inc., Y., Yelp Open Dataset, 2021. Retrieved from https://www.yelp.com/dataset. Accessed October 26, 2021.

[144] Statsmodels, "Statsmodels seasonal decompose," 2021. Retrieved from https://www.statsmodels.org/stable/generated/statsmodels.tsa.seasonal.seasonal_decompose.html. Accessed April 27, 2021.

[145] Blei, D. M., Ng, A. Y., y Jordan, M. I., "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003, doi:10.1162/jmlr.2003.3.4-5.993.

[146] Blei, D. M. y Lafferty, J. D., "Dynamic topic models," en Proceedings of the 23rd International Conference on Machine Learning, ICML '06, (New York, NY, USA), p. 113–120, Association for Computing Machinery, 2006, doi:10.1145/1143844.1143859.

[147] Blei, D. M., Ng, A. Y., y Jordan, M. I., "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, 2003, doi:10.1162/jmlr.2003.3.4-5.993.

[148] Li, Z., White, J. C., Wulder, M. A., Hermosilla, T., Davidson, A. M., y Comber, A. J., "Land cover harmonization using latent dirichlet allocation," International Journal of Geographical Information Science, vol. 35, no. 2, pp. 348–374, 2021, doi:10.1080/13658816.2020.1796131.

[149] Bock, H.-H., Clustering Methods: A History of k-Means Algorithms, pp. 161–172. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, doi:10.1007/978-3-540-73560-1_15.

[150] Paparrizos, J. y Gravano, L., "K-shape: Efficient and accurate clustering of time series," SIGMOD Rec., vol. 45, p. 69–76, 2016, doi:10.1145/2949741.2949758.

[151] Huang, X., Ye, Y., Xiong, L., Lau, R. Y., Jiang, N., y Wang, S., "Time series k-means: A new k-means type smooth subspace clustering for time series data," Information Sciences, vol. 367-368, pp. 1–13, 2016, doi:10.1016/j.ins.2016.05.040.

[152] Frias-Martinez, V. y Frias-Martinez, E., "Spectral clustering for sensing urban land use using twitter activity," International Scientific Journal Engineering Applications of Artificial Intelligence, vol. 35, 2014.

[153] Řehůřek, R. y Sojka, P., "Software Framework for Topic Modelling with Large Corpora," en Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, (Valletta, Malta), pp. 45–50, ELRA, 2010.

[154] Řehůřek, R. y Sojka, P., "Gensim: Topic modeling for humans," 2021. Retrieved from https://radimrehurek.com/gensim/. Accessed October 13, 2021.

[155] Yuan, G., Chen, Y., Sun, L., Lai, J., Li, T., y Liu, Z., "Recognition of functional areas based on call detail records and point of interest data," Journal of Advanced Transportation, vol. 2020, 2020.

[156] Sohn, Y., Moran, E., y Gurri, F., "Deforestation in north-central yucatan (1985-1995): Mapping secondary succession of forest and agricultural land use in sotuta using the cosine of the angle concept," Photogrammetric Engineering & Remote Sensing, vol. 65, 1999.

[157] Caceres, N., Benitez, F. G., y Cantarella, G. E., "Supervised land use inference from mobility patterns," Journal of Advanced Transportation, vol. 35, 2018.

[158] Hu, N., Legara, E. F., Lee, K. K., Hung, G. G., y Monterola, C., "Impacts of land use and amenities on public transport use, urban planning and design," Land Use Policy, vol. 57, pp. 356–367, 2016, doi:10.1016/j.landusepol.2016.06.004.

[159] Frias-Martinez, V. y Frias-Martinez, E., "Spectral clustering for sensing urban land use using twitter activity," Engineering Applications of Artificial Intelligence, vol. 35, pp. 237–245, 2014, doi:10.1016/j.engappai.2014.06.019.

[160] Yan, Y., Schultz, M., y Zipf, A., "An exploratory analysis of usability of flickr tags

for land use/land cover attribution," Geo-spatial Information Science, vol. 22, no. 1, pp. 12–22, 2019, doi:10.1080/10095020.2018.1560044.

[161] Ferres, L., Schifanella, R., Perra, N., Vilella, S., Bravo, L., Paolotti, D., Ruffo, G., y Sacasa, M., "Measuring levels of activity in a changing city. a study using cellphone data streams," A Study Using Cellphone Data Streams, 2020. http://datascience.udd.cl/covid_ids_tef_01.pdf.

[162] Sears, J., Villas-Boas, J. M., Villas-Boas, S. B., y Villas-Boas, V., "Are we# stayinghome to flatten the curve?," American Journal of Health Economics, vol. 9, no. 1, pp. 000–000, 2023.

[163] Beyer, R. C., Franco-Bedoya, S., y Galdo, V., "Examining the economic impact of covid-19 in india through daily electricity consumption and nighttime light intensity," World Development, vol. 140, p. 105287, 2021, doi:10.1016/j.worlddev.2020.105287.

[164] Goel, R. K., Saunoris, J. W., y Goel, S. S., "Supply chain performance and economic growth: The impact of covid-19 disruptions," Journal of Policy Modeling, vol. 43, no. 2, pp. 298–316, 2021, doi:10.1016/j.jpolmod.2021.01.003.

[165] Kaye, A. D., Okeagu, C. N., Pham, A. D., Silva, R. A., Hurley, J. J., Arron, B. L., Sarfraz, N., Lee, H. N., Ghali, G., Gamble, J. W., Liu, H., Urman, R. D., y Cornett, E. M., "Economic impact of covid-19 pandemic on healthcare facilities and systems: International perspectives," Best Practice & Research Clinical Anaesthesiology, 2020, doi:10.1016/j.bpa.2020.11.009.

[166] Golar, G., Malik, A., Muis, H., Herman, A., Nurudin, N., y Lukman, L., "The socialeconomic impact of covid-19 pandemic: implications for potential forest degradation," Heliyon, vol. 6, no. 10, p. e05354, 2020, doi:10.1016/j.heliyon.2020.e05354.

[167] Mtimet, N., Wanyoike, F., Rich, K. M., y Baltenweck, I., "Zoonotic diseases and the covid-19 pandemic: Economic impacts on somaliland's livestock exports to saudi arabia," Global Food Security, vol. 28, p. 100512, 2021, doi:10.1016/j.gfs.2021.100512.

[168] Kithiia, J., Wanyonyi, I., Maina, J., Jefwa, T., y Gamoyo, M., "The socio-economic impacts of covid-19 restrictions: Data from the coastal city of mombasa, kenya," Data in Brief, vol. 33, p. 106317, 2020, doi:10.1016/j.dib.2020.106317.

[169] Kaczmarek, T., Perez, K., Demir, E., y Zaremba, A., "How to survive a pandemic: The corporate resiliency of travel and leisure companies to the covid-19 outbreak," Tourism Management, vol. 84, p. 104281, 2021, doi:10.1016/j.tourman.2020.104281.

[170] Chen, M.-H., Demir, E., García-Gómez, C. D., y Zaremba, A., "The impact of policy responses to covid-19 on u.s. travel and leisure companies," Annals of Tourism Research Empirical Insights, vol. 1, no. 1, p. 100003, 2020, doi:10.1016/j.annale.2020.100003.

[171] U.S. Bureau of labor statistics, "Workforce statistics," 2021. Retrieved from https://www.bls.gov/iag/tgs/iag70.htm#iag70emp1.f.p. Accessed April 27, 2021.

[172] Mendolia, S., Stavrunova, O., y Yerokhin, O., "Determinants of the community mobility during the covid-19 epidemic: The role of government regulations and information," Journal of Economic Behavior & Organization, vol. 184, pp. 199–231, 2021, doi:10.1016/j.jebo.2021.01.023.

[173] Noland, R. B., "Mobility and the effective reproduction rate of covid-19," Journal of Transport & Health, vol. 20, p. 101016, 2021, doi:10.1016/j.jth.2021.101016.

[174] Chan, H.-Y., Chen, A., Ma, W., Sze, N.-N., y Liu, X., "Covid-19, community response, public policy, and travel patterns: A tale of hong kong," Transport Policy, vol. 106, pp. 173–184, 2021, doi:10.1016/j.tranpol.2021.04.002.

[175] Hu, S., Xiong, C., Yang, M., Younes, H., Luo, W., y Zhang, L., "A big-data driven approach to analyzing and modeling human mobility trend under non-pharmaceutical interventions during covid-19 pandemic," Transportation Research Part C: Emerging Technologies, vol. 124, p. 102955, 2021, doi:10.1016/j.trc.2020.102955.

[176] Borkowski, P., Jażdżewska-Gutta, M., y Szmelter-Jarosz, A., "Lockdowned: Everyday mobility changes in response to covid-19," Journal of Transport Geography, vol. 90, p. 102906, 2021, doi:10.1016/j.jtrangeo.2020.102906.

[177] Benita, F., "Human mobility behavior in covid-19: A systematic literature review and bibliometric analysis," Sustainable Cities and Society, vol. 70, p. 102916, 2021, doi:10.1016/j.scs.2021.102916.

[178] Yan, Y., Schultz, M., y Zipf, A., "An exploratory analysis of usability of flickr tags for land use/land cover attribution," Geo-Spatial Information Science, vol. 22, no. 1,

pp. 12–22, 2019.

[179] 2thinknow Innovation Cities Index 2021, Innovation Cities Index 2021: Top 100 World's Most Innovative Cities, 2021. Retrieved from https://innovation-cities.com/worlds-most-innovative-cities-2021-top-100/25477/. Accessed October 23, 2021.