Mobility Analysis & Tourism in Madeira Island

Pedro Miguel Alves Barbosa

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisor: Prof. Duarte Nuno Jardim Nunes

Examination Committee

Chairperson: Prof. Manuel Fernando Cabido Peres Lopes

Supervisor: Prof. Duarte Nuno Jardim Nunes

Members of Committee: Prof. Leonardo Azevedo Guerra Raposo Pereira

October 2021

I declare that this document is an original work of my own authorship and that it fulfils all the requirements of the Code of Conduct and Good Practices of the *Universidade de Lisboa*.

Acknowledgements

I would like to thank my supervisor Professor Nuno Nunes for helping me on the development of this thesis, a big thanks to Miguel Ribeiro for all the ideas and explanations he gave me from start to finish whenever I needed. And to all the Professors who spared time to talk to me, I would like to thank for all the support and help they gave me during the development of this thesis.

A big thanks to my friends, who when I needed a break from the work, took me out and helped me relax, without them I would never be able to finish this journey.

And finally, but not least, a special thanks to my family, who provided me all the necessary tools to have great success during this journey.

Abstract

The main goal of this thesis is to analyse people mobility and how different establishments with different services are distributed around the Island. Furthermore, with these analyses we aim to understand not only the motivation behind people movements but also the relationship between the different kind of services, so that in the future, governments, and city planners can use these data to improve infrastructures such as roads, public transports, and community places in order to reduce congestions and agglomerates of people, which is particularly relevant now that the Covid-19 pandemic is causing movement limitations. Besides helping with the restrictions and prevent congestions, we also analysed the impact that the pandemic had on the Island mobility by examine the services that suffered more with these restrictions. By doing all this, we were able to create a baseline analysis so that everyone who wants to go further or even replicate in a different location can do so and extract important information about mobility and more. These was possible by using 2 years' worth of mobility data that were gathered from a passive Wi-Fi tracking system and the geographic locations of all these establishments. We also used several data mining techniques that will be explained during this thesis as well as all the models we used to interpret our data.

Keywords

Mobility, Points of Interest/amenities, Wi-Fi, PCA, Impacts on Mobility, Correlation

Resumo

O principal objetivo com esta tese é analisar os movimentos e o comportamento das pessoas, de modo a interpretar os motivos das desolações das mesma, para além disso também vamos perceber como diferentes tipo de estabelecimentos, os quais oferecem vários tipos de serviços, estão distribuídos pela ilha e as correlações que existem entre eles. Com estes dados, nós apontamos ajudar governos e planeadores de território na construção e melhoramento de novas infraestruturas, como estradas, transportes públicos e espaços públicos, de modo a reduzir congestionamento e aglomerados de pessoas, o que é particularmente relevante agora com a pandemia, que está a causar limitações de movimento. Em adição á ajuda que estes dados podem dar para prevenir congestionamento, também vai permitir perceber o impacto que a pandemia teve na mobilidade na Ilha, isto é, vamos analisar quais os tipos de estabelecimentos sofreram mais com a pandemia e as restrições. Com todos estes dados que vamos angariar com as nossas análises, também pretendemos criar uma análise que possa ser usado como base para outras análises de modo a que seja possível replicar ou até adicionar informação extra. Tudo isto vai ser possível, usando dados relativos a movimentos durante 2 anos que foram juntos e compilados por uma infraestrutura chamada passive Wi-Fi tracking system e também os dados relativos às localizações dos diferentes estabelecimentos. Para estas análises foi usado várias data mining tecnhiques assim como diferentes modelos que permitem processar e interpretar os dados.

Palavras-chave

Mobilidade, Pontos de Interesse, Wi-Fi, PCA, Impatos na Mobilidade, Correlação

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Understanding human mobility is a major contributor to the development of knowledge on important issues such as the form and function of urban areas, the location of facilities, and the demand for transportation services. [14.] And in an era as the one we live today where there is a high population growth in urban areas, there are numerous challenges up ahead for city planners and policymakers. Being congestion levels due to increasing traffic, toxic air levels, and integration of sustainable transport, developing towards the future integrating with modern technologies. [19] For these reasons there have been an increased number of research on this topic, such as [5 , 16], and all of them adopted different technologies and methods for a similar purpose, understanding human mobility. However, we need to keep in mind that due to the complex nature of mobility these same models that try to explain mobility patterns, are only simplified views of the real world, they do not duplicate reality precisely [10]. Nonetheless these models give some highly valuable information to communities and governments so that they can utilize it in their favour.

When it comes to human mobility, there are many ways to interpreter/analyse these types of data, however some researchers suggests that one important factor that motivates people to move to one location is the available services/activities on that exact location [20]. In other words, citizens tend to move to certain locations where they can be provided some kind of service they are looking for, for example, pharmacies, schools, supermarkets etc. And by understanding this relationship between people mobility and services we can extract highly valuable information to governments and city planners. For example, they can improve infrastructures like public transports, providing traffic reports and detecting commuting patterns for planning of transport systems [8], accesses to points-of-interest and also prevent congestion and excess traffic, which in times of COVID-19 is crucial [3].

The fast evolution of mobile technology also comes with an increased usage of these devices which implement GPS, Wi-Fi, Bluetooth, etc., and with these different technologies, researchers try to use them to create new innovative models in order to understand people movement within urban areas [14]. Simulating people mobility can have many applications from improving public transport routes, help local authorities and much more. Nevertheless, with the appearance of the pandemic, also came new usages to this data because even with the pandemic, governments can still use this data to prevent any unnecessary spread of the virus and furthermore understand possible transmission routes [3]. This information can also locate the main threads in urban areas where there might be an agglomeration of people, and with this try to mitigate these impacts with governmental laws and restrictions.

Among the technologies available to do this sort of analysis the one that will be used in this thesis is Wi-Fi through routers distributed in different locations, a router is a networking device that forwards data packets between computer networks [41]. This already deployed routers infrastructure [12] is a community-based passive wireless tracking system that uses passive Wi-Fi tracking to understand mobility at scale [13].

The work described in this thesis was applied to Madeira Islands, where the population is about 250 thousand, has on average more than 1 million tourists per year (in a typical year before the COVID-19

pandemic) and have more than 100 routers distributed through the island to extract mobility data. With this infrastructure, not only we will try to understand the impacts of the pandemic on people movements, but we also aim to help the community itself by providing important info regarding human mobility in order for them to improve public infrastructures and to make urban planning decisions easier.

For these purposes, we will analyse the Wi-Fi data (Data from the Passive Wi-Fi Monitoring System composed of routers) from April to September from 2019 and 2020, we chose these dates to have a comparison between before and during the pandemic. And by extracting the locations of different kind of establishments, we can utilize these two datasets are understand the relationship between both.

1.2 Objectives and Research Questions

On this thesis what we try to accomplish is a way to extract information about people mobility and at the same time understand what type of relation the establishments (as places that provide services such as, schools, banks, etc.) in one location have with the mobility on the same location. We will also show other types of analyses, such as, analyses regarding population and distribution of services throughout the island, that will help not only on urban planning decision but also give a more insight examination on all these factors so that everyone that benefits with this information such as businesses, environmental groups, healthcare, and tourism, can use these analyses to optimize their methods of operation. All this using, an already implemented infrastructure that uses Passive Wi-Fi tracking technologies with routers distributed through several locations.

With these different datasets we aim to extract relations and information via data mining techniques and different models, while answering the following questions:

- RQ1: Can the Wi-Fi infrastructure (with the routers) translate the real mobility within a large area (an island in this case)?
- RQ2: Does the population affect the type of establishments available on a certain location?
- RQ3: Does the types of establishments available influence the other establishments around them?
- RQ4: Does the mobility have any relationship with the establishments?

1.3 Thesis Organization

This thesis will be organized in six sections. First, we review the Related Work in Section 2, where we examine some the contributions to the literature regarding this topic of people flow and the different technologies used for their implementations. In Section 3, we take a closer look on the Background for our thesis, by analysing the Passive Wi-Fi infrastructure in more dept. Then, in Section 4 the analyses are explained from the filtering to the processing, and in Section 5 we show the implementation of our analyses. In Section 6, we finally talk about the conclusions and future work.

Chapter 2

Related Work

This section presents literature related to movement patterns and distinct ways to gather data about mobility of people using several technologies. Most of the literature here is related to tourists mobility, but we need to keep In mind that for the purpose of this thesis we aim to understand people movements not only tourists.

There are many models to analyse human mobility [4, 5, 7, 8] as well as many ways to count people that enter/exit a certain location. From more rudimentary methods, such as counting by hand, to more sophisticated methods using Call Detail Record (CDR), and wireless signals such as Bluetooth, GPS and Wi-Fi. Throughout this section we will analyse previous works regarding these subjects and try to understand the differences between them in subsections.

2.1 Manual Counting

Nowadays counting people manually is no longer a reliable way to get information from tourists tally since it takes a lot of time and requires someone to be counting 24/7 on a certain location, to get the most accurate results possible.

In models based on Manual Counting, where people want a more insightful view about the routes that tourists take, these models can be a rapid and slightly reliable alternative. However, it can differ from the reality of the touristic routes. Furthermore, these methods only work in small spaces and there is no reliable way to expand this to an entire city. Although presented these disadvantages, with the current situation that we live in with COVID-19, many establishments have adopted this method to create their own models, so that they can get a distinct view of which part of the day they have more congestion and try to minimize it.

Continuing this topic, we can go one step further, by considering some evolved methods of counting not totally manually but through the usage of tickets for example. This is where data starts to get more accurate, however there might be a scenario where someone buys a ticket and does not go. Nonetheless the models generated from this data, by applying data mining techniques, can get closer to the truth.

2.2 Direct Observation

Still without any help of technology, there are models to help interpret tourists' behaviour and patterns in certain locations, one of those being direct observation. Along the years there were many projects around this topic [7, 9], with extremely good results. In these methods the data is gathered by interviewing and observing the tourists along their visit.

This was adopted in many cities, with different subject cases, there was a study [7] that analysed European trips of young American and Canadian tourists, other [9] where besides "tracking" the non-participants they also interviewed them after. And a more recent case, [6] where they observed and interviewed the visitors along their visit at the Old Town of Girona in Spain, these visitors were selected according to a random criterion. Once a visitor was chosen, they would record three types of data regarding the visitor pattern and at the end of the visit the tourist would get a questionnaire based on three types of information: conventional sociodemographic data, characteristics of the visit, and general perception of the city. With the results they could take some interesting conclusions, one of those being the most historical centre on the Old Town is actually not notable to the visitors, which is rather odd but interesting since we are talking about a historical location. They also did a tourist profiling, this is, the purpose of the visit, according to their itineraries, which they divided into four categories, basic, shopping, complex and wall itinerary.

The models created by these methods can be extremely accurate if done correctly, however, although being extremely precise they require a lot of labour and time, and it is difficult to expand to a big city since it is needed to observe the behaviour of a certain group of people around a city. During these studies, researchers had to be extremely careful when observing the non-participants, always maintaining a reasonable distance. With this method “ethical issues may quickly arise for the participant observer when tourists intentionally seek a greater degree of privacy, for instance, in the urban night environment or in the anonymity of the urban scene in general” [7], which also makes the work more difficult when trying to understand the motive of the visit.

2.3 Call Detail Record (CDR)

We will look over a model based on Call Detail Record (CDR) to, again, understand the travel patterns of visitors and potentially predict movements for future tourists. “CDRs are digital footprints of telephone calls, including information of the time a call was made, and the corresponding cell tower used to process the call” [2].

In one of the articles [5], the authors explain how these footprints of telephone calls can be used to create a predictive model of mobility patterns of tourists. With the data gathered they start by identifying if a person is a tourist or not, through a code that the phone carries. This code on the device is linked to the country where the phone came from. Next, they locate the origin of calls made by a phone with the help of the cell phone tower on different locations, and with this, they trace a path based on the different cell towers where the phone was detected.

Then, using OD (Origin-Destination) interactive maps, as shown in Fig. 1, they populate it with all their data, creating a map with different densities according to the utilization of that path. With this information the authors can identify functional traffic problems throughout the country that vary during the day, allowing them to suggest to the government ways to implement a new transportation system in specific areas of the country where improvement is needed the most.

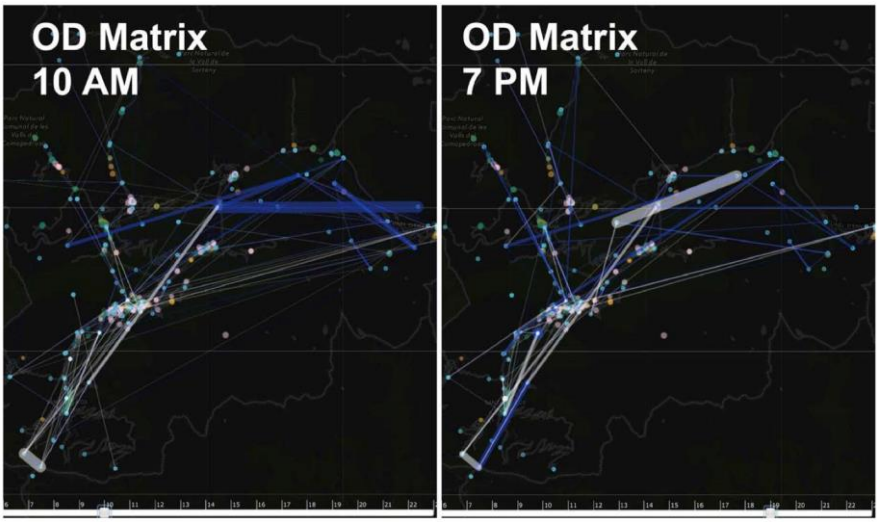

Fig. 1 Origin-Destination Matrix showing changes in mobility pattern between morning and night-time, with white lines representing Spanish tourist movements and blue lines representing French tourist movements. [5]

2.4 Bluetooth

After mentioning the most popular methods for people tracking using only manual labour and CDR, we will analyse more diverse methods using wireless connections available to us nowadays. With the continuous rise of technology and smartphone usage on an everyday basis, people keep getting bombarded with new technologies and information all the time, but with this, also comes new opportunities of developing new studies that were not possible before, in particular for this thesis proposal topic, understanding people flows and behaviour.

Some research uses Bluetooth technology to accomplish a “contribution to the field of spatiotemporal tourism behaviour research by demonstrating the potential of ad-hoc sensing networks in the non-participatory measurement of small-scale movements” [16]. For their study, they deployed 29 Bluetooth sensors on the historical centre and arts quarter of Ghent in Belgium, where they gather data for 15 days. These locations consisted of hotels and eleven of the most visited tourists’ attractions.

Whenever the sensor detected devices within their range, they would save the MAC addresses (media access control address is a unique identifier assigned to a network interface controller (NIC) for use as a network address in communications within a network segment) and COD (class-of-device) as well as the detection timestamp, in this case the MAC address serves as a unique identifier for each device.

One of the data mining methods tested in this study was association rule learning, which helps to discover interesting relationships between variables in a large dataset, but also creates many rules which makes it difficult to visualize, so to help the visualization they used a “visit pattern map”, as the authors say, this map is a “geographical depiction combining two types of information: the spatial

distribution of visits over the study area and the association (combination) of visits to different attractions. The spatial distribution of visits is visualized by proportionally sized circles showing the share of tracked individuals that visited each attraction. The association between the different attractions is visualized by means of lines connecting different attractions" [16], there are three measures to compare these rules:

- Support - measure of the share of tracked individuals;
- Confidence - measure of the probability of its consequent given its antecedent;
- Lift - measure of its support compared with the support that can be expected if and were independent;

Another stage of this Bluetooth research was the filtering of the dataset. Before starting the analysis on the information gathered, the authors had to distinguish the devices of actual tourists from people who lived/worked on those places. To accomplish this objective, they applied a progressive filtering, based on three parameters:

- Type of device;
- Duration of visit to a location;
- Duration of presence at a location;

This step is crucial to any data analysis as it allows the researchers to narrow down the noisy data from the dataset so that the results from the models created are the most accurate possible. Without this filtering any assumptions taken from the dataset could be wrong or deviated from the truth.

The next step is data exploration, more specifically, visitor segment exploration, and this is where the analysis starts to get interesting, by analysing the rules created from the authors models (association rule learning). Taking a look at one example of a rule, from the authors: {Louvre}=>{Arc de Triomphe, Notre Dame}from this we can say that there is a correlation between the attractions, which means the tourists that visited the Louvre also visited the other two attractions, we can read in more detailed about these rules in the authors paper [16] or at [1], but for the purpose of this thesis proposal, it's enough to understand the logic behind it.

With this previous research we can already see a data mining technique that gives good results and that can be implemented in order to understand people patterns and behaviour.

2.5 Wi-Fi

Besides of the previous technologies used to capture people flow and patterns there are also studies that use Wi-Fi technology to gather and analyse movements within a location.

One study on the subjects of estimating movements of people using Wi-Fi [4], focused more specifically on mass events (For example football games, universities, campuses, and hospitals) where the authors applied this method to two distinct events one being a music festival and the other a University campus. They can track the movements of people with help of access points distributed through the locations, where they stored MAC addresses of each device, timestamp from when the device was located to when it left as well as some other information.

And with the help of this information, the authors can gather information about their routes. Systems like the one described in [4] do not require user consent and are therefore capable of tracking a much larger sample set of the population [4], which means that with this technique it is possible to gather more raw data to analyse and create better predictions.

The authors also present some useful and interesting applications for the visitor's patterns/flows, which is achieved through the tracking system, such as:

- Real-time crowd management;
- Mobility models for simulations;
- Ubiquitous computing;

2.6 GPS

Now we present two studies around GPS technology that try to understand human mobility using a combination of volunteered GPS trajectories and contextual spatial information.

In [14], the authors present a new framework for the identification of dynamic (travel modes) and static (significant places) behaviour using trajectory segmentation, data mining, and spatio-temporal analysis. What the authors try to achieve with their study is, if and how the travel modes depend on the residential location, age, or gender of the tracked individuals. Although the purpose on this thesis is not exactly to extract patterns of movement, it is interesting to understand how other parameters, also from a mobility dataset, can be used to take important information about how people move, which can be utilized for further analysis.

To achieve this the authors divided the framework into three phases:

- Phase one: separating dynamic and static behaviour;
- Phase two: analysis of places;
- Phase three: analysis of movement;

In the first phase the authors used a learning method to classify the movement of 250 segments from the training set in three categories, vehicle movement, walk/run and stop according to certain parameters. The next phase is where they identify the locations of significant places and categorize them according to an external special data, the way they identify these locations is by inspecting the stop segments from the first phase and calculate the frequency of reoccurrences and the amount of time spent in there. In the final phase they categorize the movement of the test set.

On the second study [17] the authors goal is to mine interesting locations and classical travel sequences in a given geospatial region, the way they do this is by, first model multiple individuals' location histories with a tree-based hierarchical graph (TBHG) and then propose a HITS (Hypertext Induced Topic Search)-based inference model. To perform this study the authors used a large dataset collected by 107 users over one year.

The authors in this study in order to implement the graph chose several variables [17]:

- GPS log - collection of GPS points;
- GPS trajectory;
- Stay point - geographic region where a user stayed over a certain time interval;
- Location history - record of locations that an entity visited in geographical spaces over a period of time;
- Tree-Based Hierarchy H - where H is a collection of stay points;
- Location Interest;
- User Travel Experience;

With all these variables, especially the location interest and user travel experiences, the authors implemented a graph to mine classical travel sequences. The authors based their graph on scores, where the classical score of a sequence is the integration of the following aspects [17], 1) The sum of hub scores of the users who have taken this sequence. 2) The authority scores of the locations contained in this sequence. 3) These authority scores are weighted based on the probability that people would take a specific sequence. Using a TBHG graph, Fig. 2 demonstrates the calculation of the classical score for a 2-length sequence, A→C. The graph nodes (A, B, C, D and E) stand for locations, and the graph edges denote people's transition sequences among them. The number shown on each edge represents the times users have taken the sequence. From this graph the authors went even further and took a more insightful examination in terms of mobility, explained in [17].

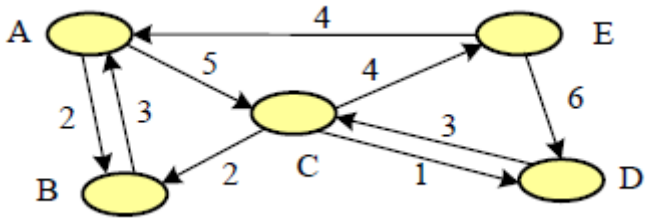

Fig. 2 Demonstrating classical sequence mining with a graph. [17]

2.7 Points of Interest to Estimate Mobility Patterns

We have presented in the previous subsections several ways to analyse people pattern directly with the of different devices capable to track and count people movements. Now we will explore a paper [19.] that uses Points of Interest to estimate mobility patterns, which requires a more in dept analysis of the location in which the study is conducted (London).

For their analysis they need several types of data, such as:

- Shape files of London, divided by warden;
- Population Data;
- OpenStreetMap Data with the Points of Interest;
- Travel Data, with Origin-Destination data from the Transports of London;
- Station Coordinates;

And by using the OSM Data (OpenStreetMap Data), the authors created several maps of London with the distributions of Points of Interest, which were divided into different categories, for example, Educational, Financial, Healthcare, etc… so that they could have a better understanding of the area they were analysing. From here, with the locations of the station and the Origin-Destination data, they could create a model based on the POIs to estimate mobility patterns.

Taking this initial idea of analysing the city itself and the distribution of the POIs, we will try to extend this to our case study in Madeira Island and by adding our routers dataset, this combination will help us to extract important information about human mobility and examen the impact of COVID-19 on the island.

Chapter 3

Case Study – Madeira Islands

Now that after mentioning several techniques for analysing and gather information about people flows, let us get into the background of this thesis proposal and from where it comes, what was done and how it was done.

3.1 Overview

This thesis was conducted in Madeira Island, where the population is about 250 thousand people and has on average more than 1 million tourists per year (in a typical year before the pandemic). And over the last 60 years there has been a significant change on the population and their mobility as well as on job distribution and economic activities on the island, all these due to the big transport networks constructed to try to minimize congestion between different locations [21].

The baseline is a Passive Wi-Fi monitoring system located on Madeira Island, with which is possible to account the number of people(devices) on several locations across the island [12], [13].

The way this is possible is, instead of using methods like Bluetooth, direct observation, or even manually counting, it uses Wi-Fi, where the information is gathered with the help of more than 100 routers distributed throughout the island in numerous locations, where each one stores information about the devices nearby that send probe requests (A message sent by a client requesting information from a specific access point (router)). Each router can save anonymous information of users, like MAC Address, time that the user entered and left a certain location, intensity of the signal, SSID, among other information.

In Fig. 3 [12] is possible to get a clearer view in how the data is store in the data warehouse and the different tables and variables that exists. It was from these tables that all the raw data was extracted, but the main tables that will be used for our proposal are:

- probe_request;
- stay;
- router;
- mac_address;
- vendor.

The probe_request table, gathers all the “messages” (id) that are sent by a device (mac_address_id) to a certain router (router_id) on a certain time (capture_time). The stay table is populated through the probe_request table, which then give us the information about how long a device stayed near a router. As the name suggests the router table stores the information about the routers such as location and typology (description). The same thing goes for the mac_address table where all the new mac addresses are stores. The vendor table is where the brand of the devices is stored.

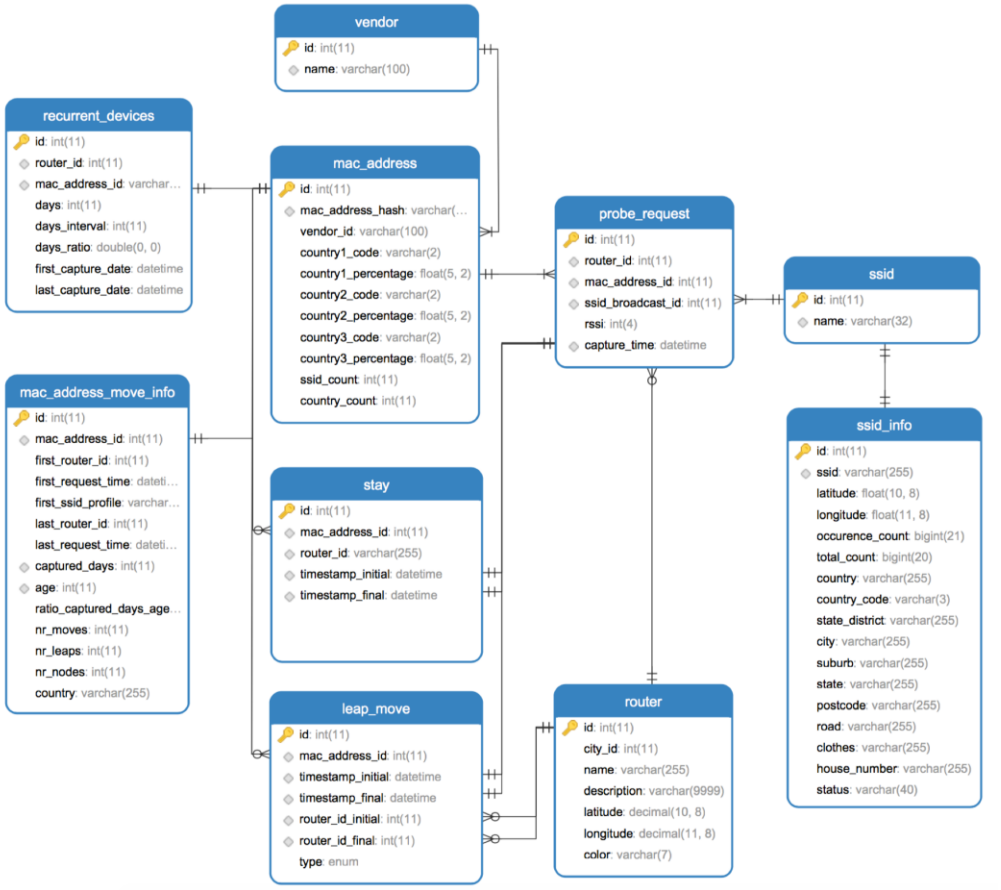

Fig. 3 Capture and process database schema. [12]

All this infrastructure was deployed and maintained by community itself [13], providing a low-cost infrastructure to help the community understand the patterns of their visitors. Each router also has a location typology to designate itself:

- Plazas;
- POIs;
- Events pavilion;
- Football stadium;
- Airport/port;
- Nightclubs.

This way it is easier to understand the count of devices on certain places on a certain time (i.e., on a Nightclub there might be more devices during the night-time).

Fig. 4 Activity heat map of the island with the most crowded locations (red) against the less popular (green).[13]

In Fig. 4 we can examine the distribution of these routers throughout the island and where there is more activity. To check the reliability of the count of the devices on some locations, it was used a source of ground truth to compare the results taken from the routers. Some of these ground truth sources were, for example, arrival counts from the airport itself, for the Port routers was the number of ships at the boat docks and finally in the stadium the count was compared to the tickets sold. However, when comparing the routers count with the real results, it is not taken into account the fact that one person can carry more than one device, and this might affect the end results. Another factor that makes this count method not 100% effective to all typologies, as mentioned by the authors [13], is the fact that depending on the location we are examining the results might be affected by the type of traffic and if people are stationary or passing nearby the location.

One analysis done by the authors was the differences between normal days and event days, and for these, they created three types of detections:

- Daily events - Which compares each day with all the other days within the selected dates;
- Weekday events – Which compares each day with only the same weekdays days from other weeks;
- Hourly events - Which displays the number of events within 15-minute intervals for each day.

They used these designations to create a detection algorithm to identify these events in the data.

To support this infrastructure, the authors also developed an interactive map (https://mare.iti.arditi.pt/) for the public to analyse the data live and verify the counts of the routers along the island. One interesting feature added recently was, the comparison between the number of people in a location to the maximum number of people recommended by the government with COVID-19.

With the help of all the data gather from the Passive Wi-Fi monitoring system, in this thesis proposal we will suggest a model that allows any client an understanding on how the tourists move on the island and predictions of the future visitors' patterns.

3.2 Architecture

In this section we will go through the different components of the infrastructure we are using for this thesis proposal and inspect in more dept how the setup was done [11], in Fig. 5 we can visualize all the connections between the different components. For the routers it was used a commercial TPLink MR3240v2 home router to capture all the data. Since the authors were targeting multiple location and this solution was capable of running Linux system, they opted for installing OpenWRT – Barrier Breaker 14.07. Witch had all the wireless interfaces needed for the implementation as well as the minimum storage necessary to do all the tasks [11].

The wireless interface was assigned to a single network, and configured to monitoring mode, for security reasons. In order to set the wireless interface on monitoring mode, it was added a Python script to run at the start scripts of the router. This basic script uses the package Scapy to capture packets and filter them for the type 0 (management) and the subtype 4 (probe requests). The probe requests detected are sent to a remote database through a web service with JWT authenticated messages in HTTP requests and then stored in a MySQL database. All the calculations and optimizations required for analyzing the captured data and provide the results through a web server to the clients, were done by the server-side components [11].

The Wi-Fi routers are connected to a VPN allocated on the server to allow their remote management, and the scripts processing the data interact with several external services and APIs.

With this passive Wi-Fi tracking system, named Beanstalk, it is possible to collect Wi-Fi requests from devices nearby as talked in the previous subsection.

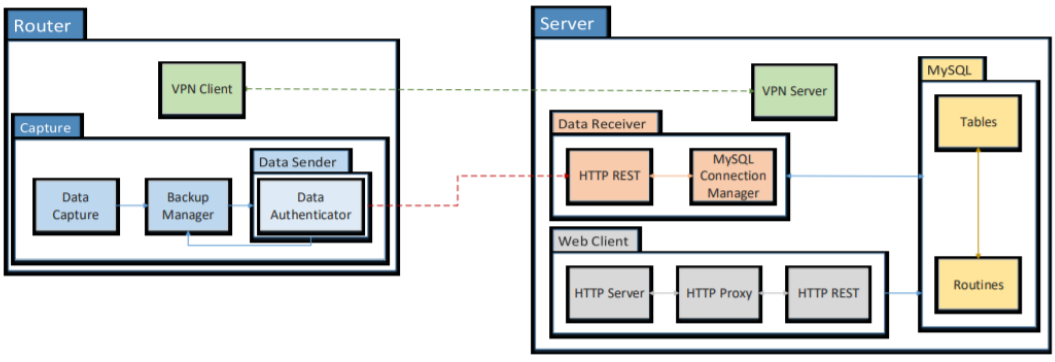

Fig. 5 System Architecture. [11]

Chapter 4

Data Analysis

4.1 Introduction

With this increase of human mobility also comes more complex analysis of human behaviour, such as understanding why people choose to move to a certain place instead of other place with similar characteristics. In this section we will explore this idea by doing a more in dept analysis of Madeira. We will start with an analysis of the distributions of population and POIs across the island in order to categorize by theme all areas.

After categorizing all island, we use the Wi-Fi data to complement our initial analysis. And here we explore how Covid-19 impacted human mobility and which places maintained their mobility and why, we did these by analysing two different periods one in 2019 pre-Covid and one in 2020 during Covid.

4.2 Data Description

In this section, we will give an overview of the data used for our analyses, mentioning the sources and the structure of the different datasets.

4.2.1 Shapefiles of Madeira

The shapefiles are the necessary files that provide the geographic boundaries of the concerned region. The shapefiles are captured in the form of geographic information system (GIS) files. Geographic information system (GIS) is a computer system for capturing, storing, checking, and displaying data related to positions on Earth's surface [22] [23] [19].

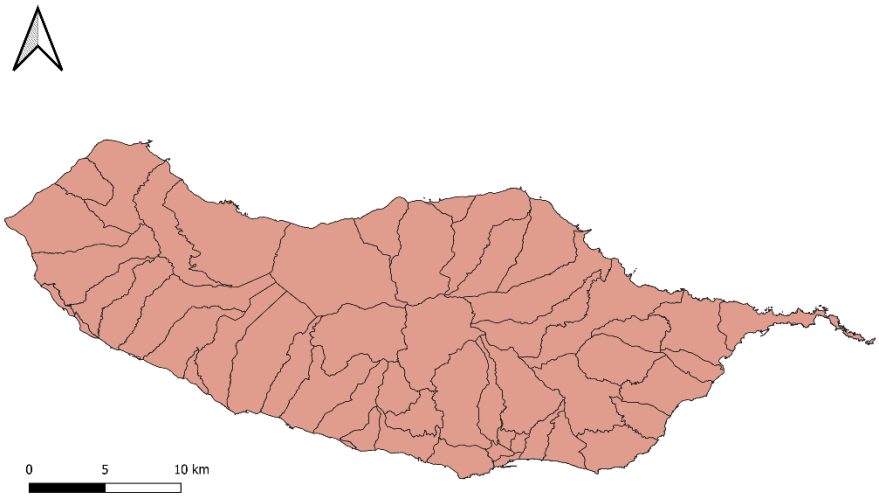

Fig. 6 Madeira Shapefile

Madeira shapefiles (Fig. 6) can be downloaded from CAOP (Carta Administrativa Oficial de Portugal) and were compiled in 2019 [27]. In the one used for our research, Madeira was divided into parish using polygons instead of lines as the geometry of the shapefile, where each polygon as its own ID, Freguesia (parish), Concelho and area in square kilometres, with these we get 200 different polygons.

In order to do a more in dept analysis of the areas with more concentration of points of interest, we can do an even more detailed map than a one divided into parish by creating a hex-grid in Madeira and get a map with a higher level of preciseness. The primary advantage of a hex map over a traditional square grid map or an parish map is that the distance between the centre of each and every pair of adjacent hex cells (or hex) is the same, which makes the distribution more even between different locations [47].

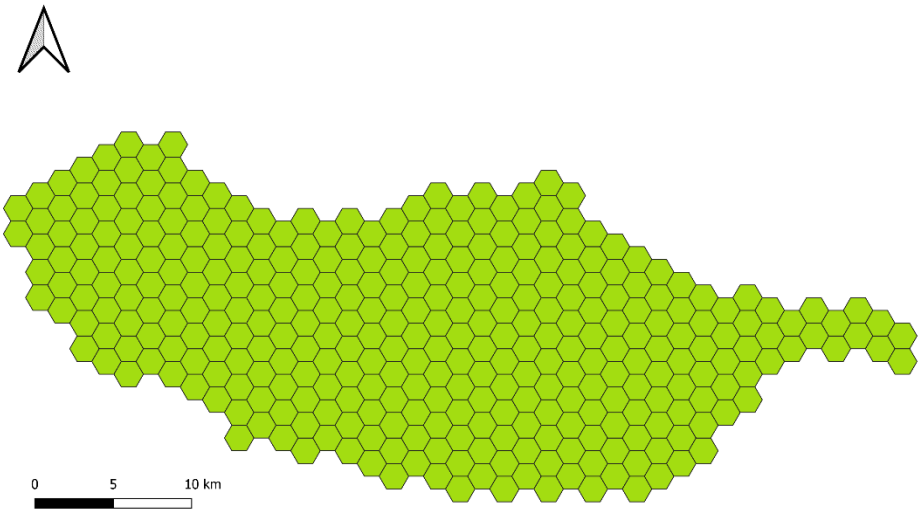

Fig. 7 Madeira Hexagonal Grid

These maps and the creation of the hex-grid were done in QGIS (Geographic Information System), an open-source software that supports viewing, editing, and analysis of geospatial data [40]. So, with the help of this Software we created a hex-grid as in Fig. 7 with 331 hexagons, ending up with a new shapefile where we have “hex_id” the id of each hexagon, “hex_area” (2.7 km2) the area of each hexagon, ”freguesia” which refers to the parish in which the hexagon is located and “concelho”

4.2.2 OpenStreetMap Data (OSM) for POIs

A Point of Interest (POI) is a specific location that someone may find useful or interesting [43], such as hospitals, banks, shops, restaurants, etc. And it is with these points that we can categorize the type of location we are exploring.

There are several sources from where we can obtain this kind of data related to Points of Interest, the one chosen for this thesis was OpenStreetMaps (OSM). To download all the data needed, we used two java command lines, OSMOSIS [28] and OSMCONVERT [29]. The first command line was to restring the area to be analysed only in Madeira:

(1) osmosis --read-pbf file=portugal-latest.osm.pbf --bounding-polygon file=madeira.poly --write-xml madeira.osm

Where portugal-latest.osm.pbf [30] represents the OSM file from Portugal and the madeira.poly [31] has the polygon of Madeira which will narrow the selection only to the Island region, from this command we end up with a madeira.osm that has all the POIs in Madeira Island.

The second command creates the csv file with the location and category of the location:

(2) osmconvert madeira.osm --all-to-nodes --csv="@id @lon @lat amenity building healthcare shop tourism" --csv-headline --csv-separator=";" -o=madeira.csv

The madeira.osm file is the output file from the first command (1), "@id @lon @lat amenity building healthcare shop tourism" are the columns that we selected for our csv, where the @lon and @lat are longitude and latitude respectively and the other columns are the categories [32] that we wanted to select.

4.2.3 Population

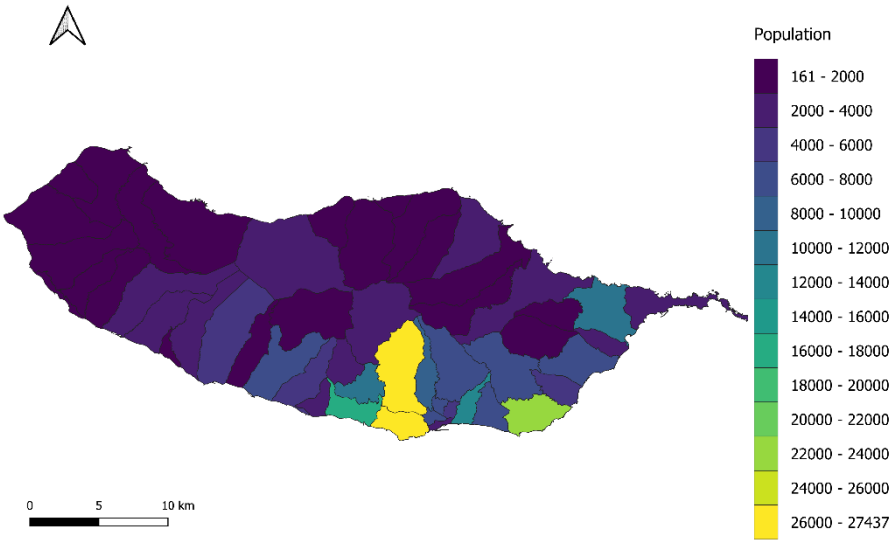

Fig. 8 Population per Parish

To get the population per parish in Madeira Island we used 2011 CENSOS [42], that were compiled by Instituto Nacional de Estatistica. Once we had this data, we created a map with the population values as in Fig. 8.

4.2.4 Wi-Fi Data

In order to understand the people movements and the areas with more mobility, we will use the counts of the routers distributed through the island, as explained in section 3, where for each router there is an hourly count of devices that entered the range of the router for each day. It will be with this data that we will do our analysis on the impact that COVID-19 had in the mobility on the island by comparing the counts from 2019 and 2020. We will analyse six months, from April to September, both in 2019 and 2020.

4.2.5 Telecom company Data

Another dataset that will be used in our analysis in mobility will be a dataset compiled by a telecom company which will be used as ground truth. Similarly, to Call Detail Record (CDR), with the use of cell phone towers it is possible to get the counts of devices that moved from one cellular tower to get associated to another cell.

These datasets are only available for 2020 when the pandemic started and will be used to validate the Wi-Fi dataset (the Passive Wi-Fi monitoring technology).

4.3 Data Processing

With all the datasets analysed and downloaded, it enters the next phase, the processing of those datasets and how to modify and filter the datasets to get the results expected.

The Processing phase will be divided into six phases:

- OSM Data Processing;
- Area Categorization;
- Population Processing;
- Router categorization;
- Wi-Fi Dataset Processing;
- telecom Dataset Processing;

The objective of these six phases is to filter and modify the different datasets, and compile them into new datasets, that will be used to feed the analysis methods.

First the processing, the data from OpenStreetMaps with all the information related to the establishments available. After explaining processing methods used on the OSM data, we will examine how to use this data to categorize an area according to the type of establishments available on that exact location. We then proceed to inspect the population dataset and right after we go into categorize the routers by inspecting the different POIs surrounding them.

Finally, we will process and filter the Wi-Fi and telecom Dataset, so that it can be used later in our model.

For the realization of this section, we used R programming language and Python for the filtering and processing of the datasets, and QGIS to create our maps. These softwares were chosen for the diversity of packages available that can analyse spatial data.

4.3.1 OSM Data Processing

As said previously, in order to categorize an area, we need to understand the type of establishments available nearby. And to get those establishments we use OpenStreetMap, where we downloaded all the POIs for the Madeira Island, as explained in section OpenStreetMap Data (OSM) for POIs.

id	lon	lat	amenity	building	healthcare	shop	tourism
4461385990	-16.8530156	32.6441512				greengrocer	
4461916353	-16.9020873	32.6469912					
4461916354	-16.9020787	32.6472637					
4461916355	-16.9023904	32.6472646					
4461916356	-16.9023883	32.6473028					
4461916357	-16.9018633	32.6476498	waste_basket				
4461916358	-16.901652	32.6499137					
4461916359	-16.9016152	32.6499758					
4461916360	-16.901571	32.6500153					
4461916361	-16.9015311	32.6499865					
4461916362	-16.9014578	32.6500477					
4461916365	-16.9014129	32.6500697					
4462367692	-16.8521695	32.6434246				gift	
4462367693	-16.8515941	32.6434035					
4462602016	-16.9239999	32.6493166					artwork

Fig. 9 Initial Dataset from OpenStreetMap

Once we got all the locations in OpenStreetMap, we ended up with a dataset, as in Fig. 9, with 730089 rows where each row had the longitude and latitude of the location as well as the type of building. To describe the type of building we had 5 different variables (columns), amenity, building, healthcare, shop and tourism. Amenity is for describing useful and important facilities for visitors and residents [33], for example schools, banks and prisons. The building key is used to mark areas as a building [34] such as offices and churches. The key healthcare is used to map a facility that provides healthcare (part of the healthcare sector) [35]. Shop contains all the commercial establishments and stores. Tourism is for the places and things of specific interest to tourists including places to see, places to stay, things and places providing information and support to tourists [36].

In Fig. 9, there are some rows which do not have any type of establishment, with this, the first step was to filter all the empty rows and ambiguous types. Once it was done, we ended up with 17037 different locations. When the filtering was done, two new variables (column) were created, *type* and *sub_type,* these variables are filled with the first non-empty value of the other columns in the following order, amenity – shop – tourism – healthcare – building, for example if one location has a value on column amenity and tourism, the new row with the new variables would have the corresponding coordinates (latitude and longitude), *type* = “amenity” and *sub_type* = value of the column amenity, here we chose the value as amenity instead of tourism because of the order that we considered previously.

id	longitude	latitude	type	sub_type
33993127	-17.0044439	32.6564703	tourism	viewpoint
60780252	-16.6976713	32.7489718	tourism	viewpoint
60780343	-16.6812209	32.739598	tourism	viewpoint
248797536	-16.8863253	32.7354094	tourism	information
255575356	-16.7932784	32.6890917	tourism	hotel
271689860	-16.9165519	32.6649724	tourism	hotel
271705628	-16.8501144	32.6371084	tourism	viewpoint
271716431	-16.7080996	32.7508885	tourism	viewpoint
275681647	-16.9427105	32.7588812	tourism	viewpoint
286711135	-16.9394112	32.6586625	tourism	viewpoint
286728213	-16.962948	32.710232	tourism	hotel
286734237	-16.8506323	32.6385891	amenity	place_of_worship
286757268	-16.764391	32.7142767	amenity	place_of_worship
286757271	-16.7656732	32.7188169	amenity	place_of_worship

Fig. 10 OSM Data after first processing step

After applying this processing, the new dataset can be compiled with the important variables being, *latitude*, *longitude*, *type* and *sub_type* as in Fig. 10. Once the new dataset was populated, followed the categorization of all locations into different groups, these groups consist of different establishments that offer a certain service to the community/people.

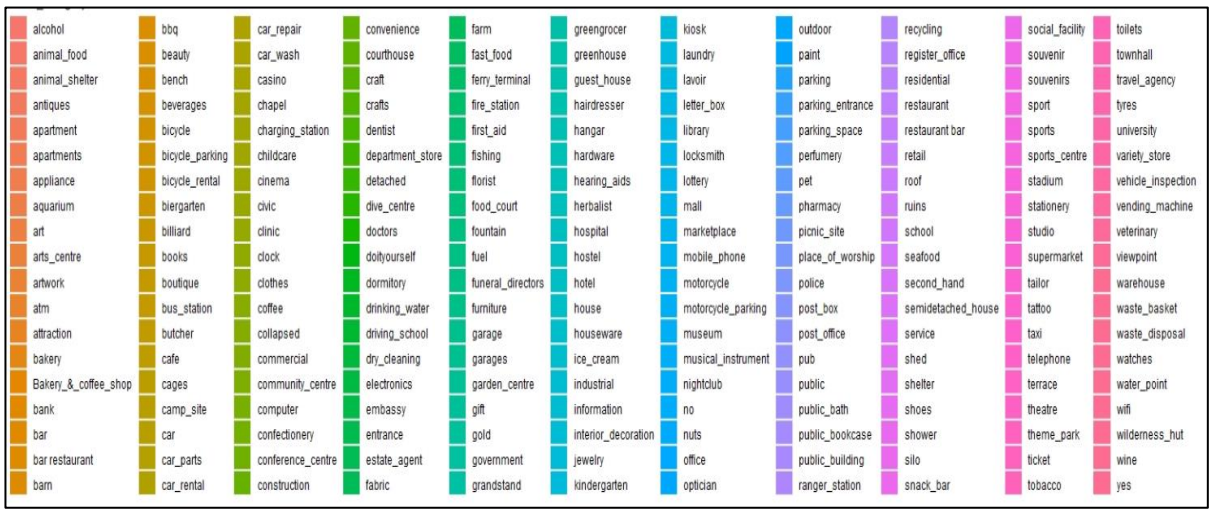

Fig. 11 Different establishments *sub_types* on OSM Data

The *sub_type* or *type* variables were not used for the categorization for the following reasons, if *sub_types* was used, there would 200 different categories as shown in Fig. 11, which would be very complicated to analyze, the same idea goes for the *type* variable which has only 5 values and the distribution is not categorize (mostly for the amenity *type*) and in order to do a more in dept analysis we needed more precision.

Table 1 Groups of Points of Interested summary

POI *Group*	**Count**	**Description**	**Examples**
Commercial	887	Commercial establishments	Shop, supermarket, mobile_phone
Community	528	Places that provide services to the community	sport_centre, place_of_worship, camp_site, picnic_site
Educational	217	Establishments regarding education	Kindergarten, school, university
Entertainment	238	Places related to attractions	Aquarium, museum, theatre
Financial	167	Establishments related to money	Atm, bank
Government	201	Establishments owned by the Government	Embassy, courthouse, police
Healthcare	213	Places that provide medical services	Dentist, pharmacy, hospital
Living	8425	Places with houses and apartments	Apartments, house, residential, dormitory
Sustenance	1550	Establishments with food and drinks	Bakery, bar, cafe, pub, restaurant
Tourism	1103	Places with tourism interest	Hotel, information, viewpoint
Transportation	1242	Any kind of transportation services	bus_station, taxi, car_rental

With that said, the new variable *group* has eleven different values being Commercial, Community, Educational, Entertainment, Financial, Government, Healthcare, Living, Sustenance, Tourism and Transportation. A description of each group can be analysed in Table 1.

id	longitude	latitude	type	sub_type	group
33993127	-17.0044439	32.6564703	tourism	viewpoint	Tourism
60780252	-16.6976713	32.7489718	tourism	viewpoint	Tourism
60780343	-16.6812209	32.739598	tourism	viewpoint	Tourism
248797536	-16.8863253	32.7354094	tourism	information	Tourism
255575356	-16.7932784	32.6890917	tourism	hotel	Tourism
271689860	-16.9165519	32.6649724	tourism	hotel	Tourism
271705628	-16.8501144	32.6371084	tourism	viewpoint	Tourism
271716431	-16.7080996	32.7508885	tourism	viewpoint	Tourism
275681647	-16.9427105	32.7588812	tourism	viewpoint	Tourism
286711135	-16.9394112	32.6586625	tourism	viewpoint	Tourism
286728213	-16.962948	32.710232	tourism	hotel	Tourism
286734237	-16.8506323	32.6385891	amenity	place_of_worship	Community
286757268	-16.764391	32.7142767	amenity	place_of_worship	Community
286757271	-16.7656732	32.7188169	amenity	place_of_worship	Community

Fig. 12 Final Dataset with the OSM Data

With this new variable *group,* we finalize our dataset (longitude, latitude, type, sub_type, group) with all the POIs available for Madeira Island as shown in Fig. 13. Each value of the variable *group* refers to a different establishment where each establishment has its own purpose symbolizing its *group*, for example, inside the category Commercial there are all kinds of shops and supermarkets whereas in Community we have churches, parks and football fields. For the Educational category the main establishments that represent this group are schools, Universities and kindergartens being places that give any type of tutoring. The distribution of the *sub_types* through the *groups* was done with help of OpenStreetMap wiki and common sense.

4.3.2 Area Categorization

With the new dataset with all the locations and their respectively category, we built maps to visualize the distribution of the different establishments on the Island. By using QGIS, we can create these maps and analyse the locations with more POIs of a certain group.

Initially we joined two different datasets, the one where there are all the polygons of Madeira Fig. 6 (which is divided by parish), and the dataset created from the last section, where there are the locations of all the POIs. One important aspect to take into account when using two different datasets of spatial data, is that the coordinate system needs to be the same in both of the datasets, for the purpose of this thesis we used WSG84 as the main coordinates system. Before passing the datasets to the QGIS software, we needed to create the shapefile with the geometry [26] of the POIs, this geometry is the variable that will allow QGIS to locate the point in the map. To transform the longitude and latitude into the geometry we used the geopandas package in python.

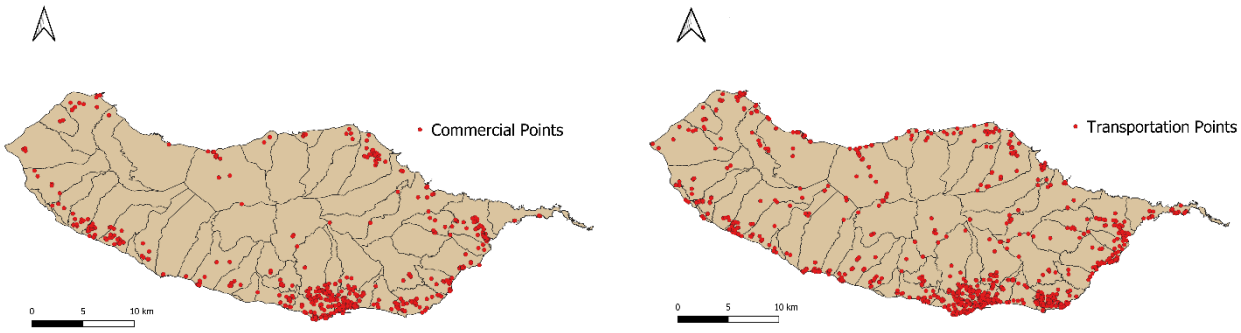

Fig. 13 Distribution of points of Commercial and Transportation

Once the geometry of the points is done, we can overlap the two shapefiles (polygon of Madeira and points of interest) in QGIS. To be easier to visualize, instead of overlaying all the points of all groups with the map of Madeira, we created one map for each group, creating maps like the one in Fig. 13 (all the maps can be visualized in A.3).

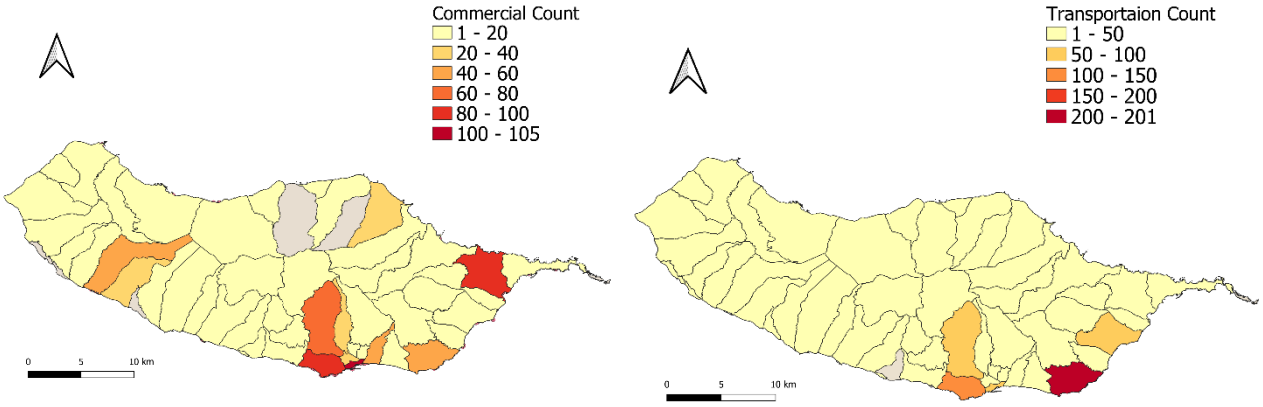

Fig. 14 Total of Commercial and Transportation establishments per Parish

To help with our analysis we needed a way to transform these distributions into numbers that can be utilized in our model so, for each parish we summed all the points of each group, as in Fig. 14 (the remaining maps are in A.2), ending up with a map where we have the number of locations available of each group inside each parish. With this analysis we can already get a good look on how the different groups are distributed through the island.

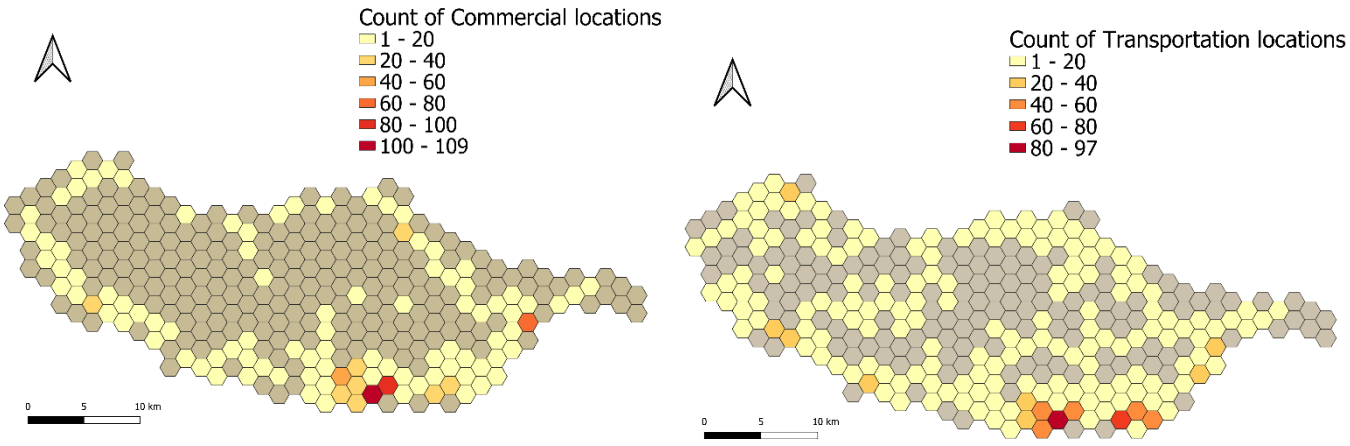

Fig. 15 Hex-Grid with Commercial and Transportation Distribution

Although the visualization by parish already gives us a good perspective on how all the POIs are distributed, there are some drawbacks with this analysis, one of them being that some parishes are too large, which may result on all the points of one parish being located only on one small spot, as in Fig. 13, which does not give us a precise distribution. To do a more in dept analysis for our model we used the hexagonal grid of Madeira, the same used for the population distribution, and applied the sum of the

POIs for each hexagon, ending up with a hexagonal map as in Fig. 15 for each group (remaining maps are in A.1). With these maps we increase the preciseness of the map giving us a more real distribution of the POIs, this verifies by visualizing the Commercial maps for example, when inspecting Fig. 13 (Commercial) it appears that most parishes have several distributed establishments however once we create the Hex-Grid we see that most of those establishments are in the coast, with these values our model should give more precise results than it would give if we did the calculations per parish.

4.3.3 Population Processing

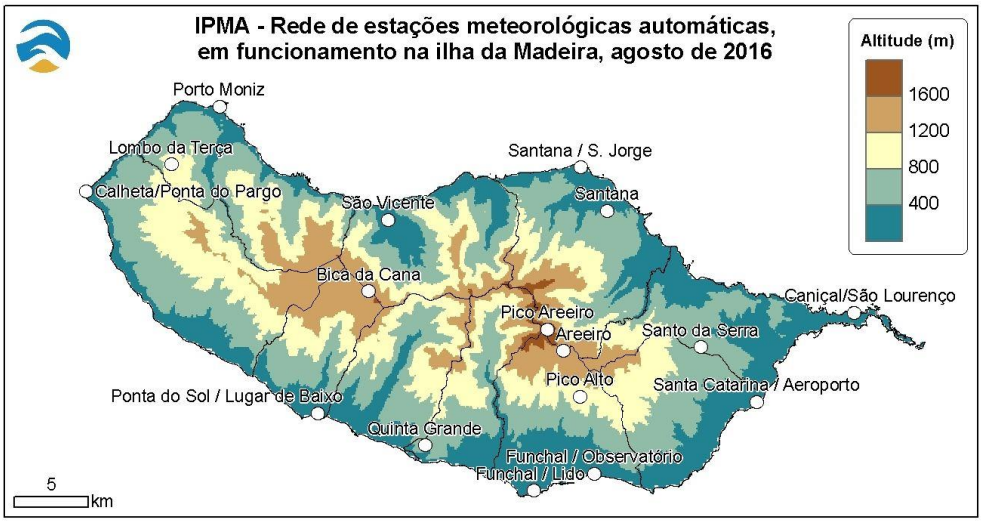

Fig. 16 Altitude map of Madeira Island [44]

As said on the previous section, we used the data gathered from CENSOS to get the population by parish. However, this distribution by parish is not the most precise and won't give us the best results at the end since, when talking about Madeira Island most of the population is located near the coast and furthermore, despite the existence of some population on the interior, most of these areas are natural reserves and mountains as shown in Fig. 16.

With these factors in mind, we had created a new map for the population, that allowed us to have a perception on how the population was distributed. To that end, we used our known Hex-Grid Map (Fig. 7), mentioned in Shapefiles of Madeira Section, with which we can better visualize where the population is located.

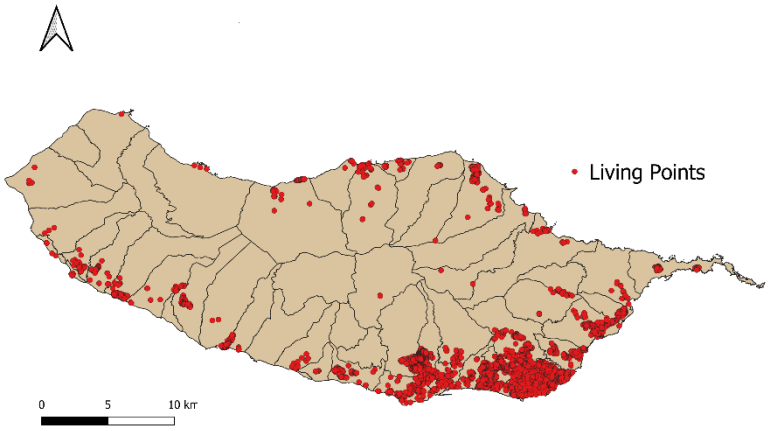

Fig. 17 Living Points

In the previous sub-section we described the different groups of establishments/buildings and their location, and one of them was the living group that represented locations with apartments and houses where people live (Fig. 17), with this information in mind we can detect the areas where the population is concentrated.

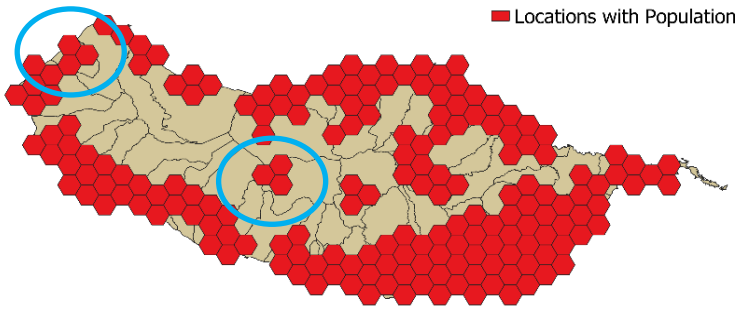

Fig. 18 Parish map overlaid with Hexagons with population

In other words, we can pinpoint the hexagons of the Hex-Grid map in which there will be population by verifying if there are any apartments or houses inside that hexagon, ending up with a map as in Fig. 18. For the map in Fig. 18, we had to add two new areas that were missing from the OSM data for the living group in order to get the most accurate results possible (highlighted in figure above).

The last step is to calculate the population itself per hexagon. To summarize, with the population by parish coupled with the hexagons in which there are living locations inside, we can slightly calculate the population per hexagon by dividing the population of each parish for all the hexagons inside that exact parish. For example, in Fig. 18 we can visualize that there are some parishes where only a small portion of it that has population, what we do is divide all the population of that parish by the hexagons on it, although there might be more population on the parish besides the areas we identify, the population per square kilometre on those location is too low (< 10 hab / km2) that we consider negligible. Once the calculation where done we ended up with a map as in Fig. 19.

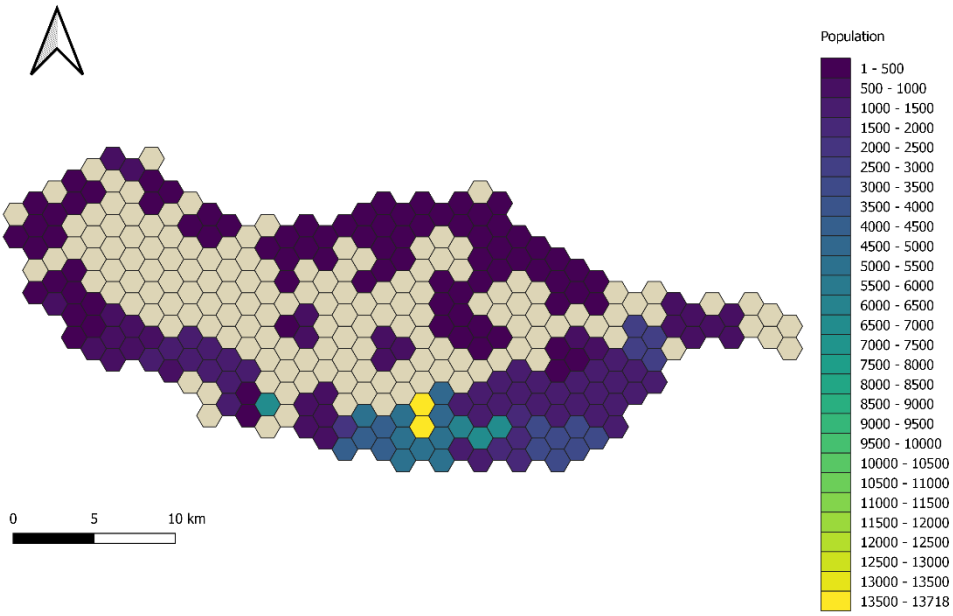

Fig. 19 Population per hexagon

4.3.4 Router Categorization

Following the same idea from the previous section on categorizing Madeira Island by creating maps with the distribution of the different groups of establishments. Now we created a visualization to help us understand the different establishments that exists around a router.

To implement this, we used R programming language and the library “geosphere” [45] which allow us to calculate the distance between two different geographic points (latitude and longitude). With all the locations of the routers and the POIs, we calculated, for each router, the quantity of each group of establishments there are in a radius of 500 meters of the routers’ location, ending up with a chart as in Fig. 20 (all the routers are in A.4), where the “router_id” refers to the actual router id and “Freq” is the number of establishments that exists from each group.

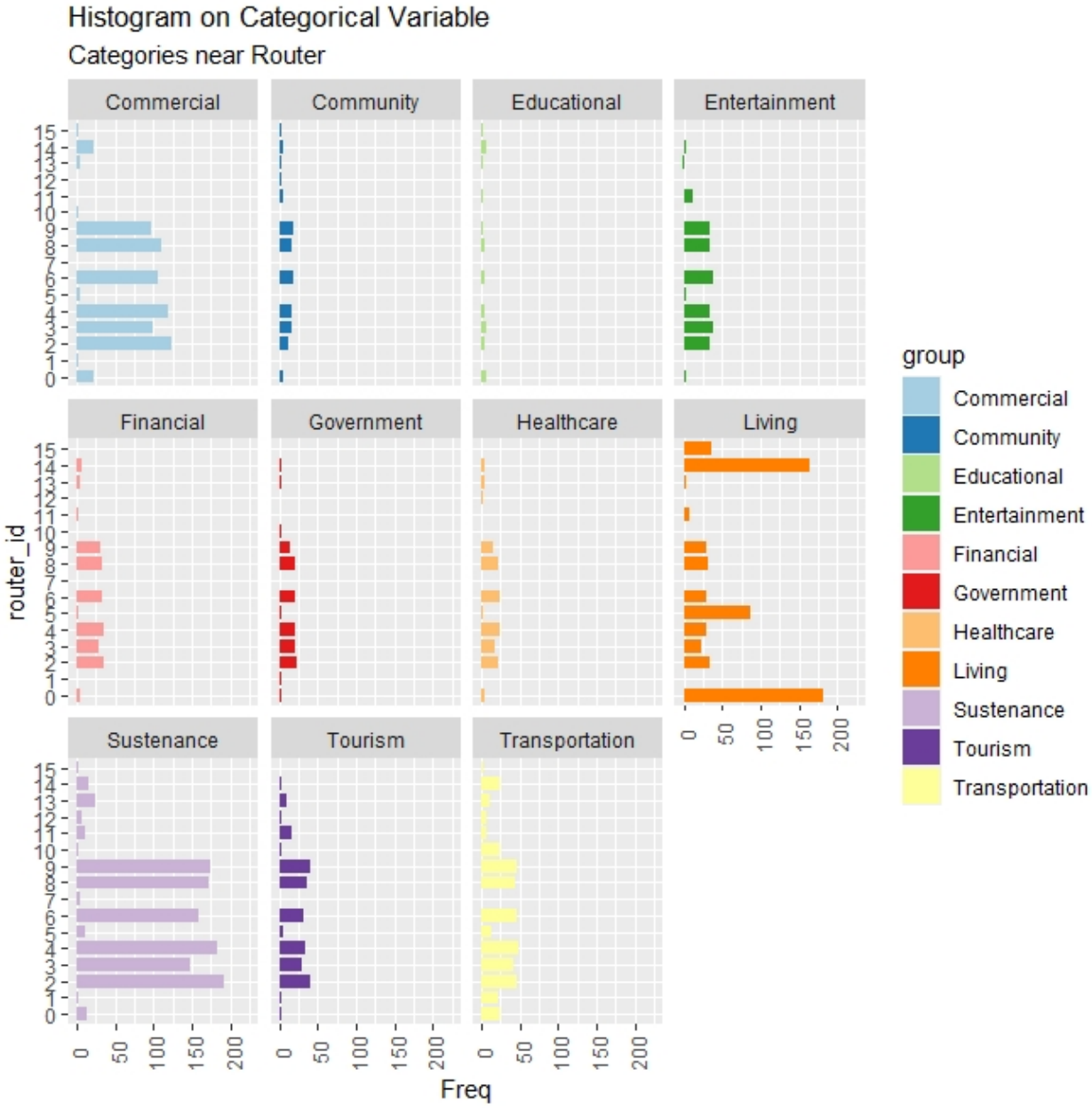

Fig. 20 Number of locations in a radius of 500 meters from each router

4.3.5 Wi-Fi Data Processing

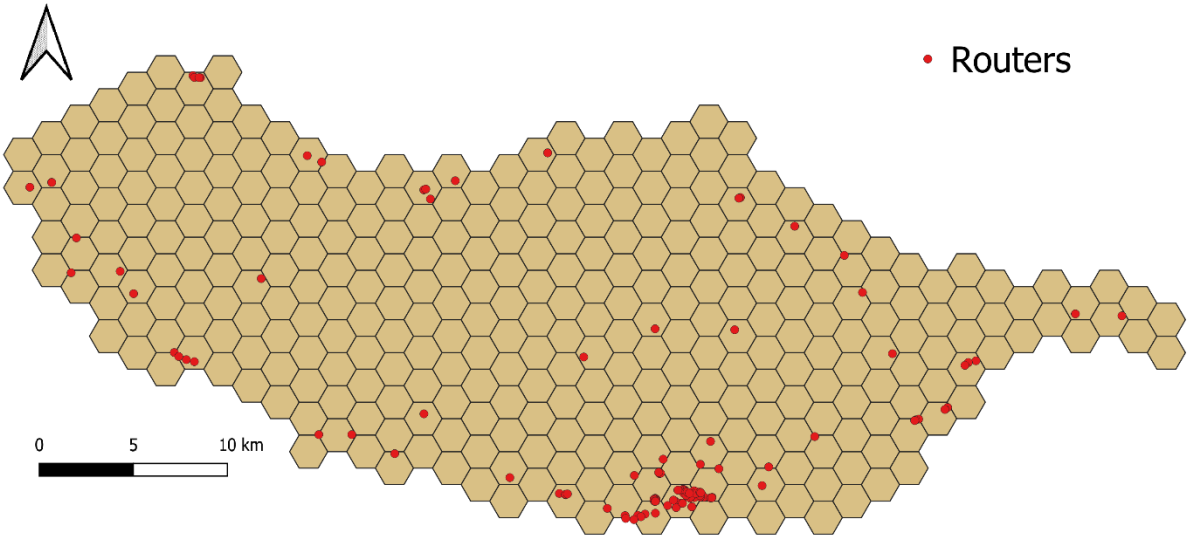

Fig. 21 Routers Distribution

To help us do a better categorize the locations of Madeira Island, we also used all the passive Wi-Fi infrastructure available to us, with more than 100 routers distributed through the Island (Fig. 21) as mentioned in Section 3. And with these routers, we had access to data regarding the counts of devices in certain areas.

The organization of this data is the following, there are hourly counts for each router, which have the number of devices per hour that entered a certain location within the range of the router, ending up with 24 different values for each router for one day. For this thesis we will analyse 6 months' worth of data from April to September both in 2019 and 2020. First, we need to prepare the data and filter it.

We started by using the raw counts of the routers to complement our analysis, however as we went further, we noticed some drawbacks on this method. The challenges on using the data as it is, is twofold: first there were to many outliers in the data, produced by the router itself when it got internet issues, for example it would send all the counts accumulated from three days, followed by days with zero counts; and second, depending on where the router is located the counts of devices can mean different things, for example, if we are in the city centre, one hundred counts at rush hour might not be much however for a router located outside the city centre where there are not many people, one hundred counts is actually a lot.

For these two reasons, we ended up processing the Wi-Fi data and make it more error free. And in order to do that we started by removing the outliers from all the Wi-Fi data available to us, using the "The IQR Method" algorithm which divides the data into three quartiles, defined by us, and then takes the values from the quartiles one and three and change them to a value in the border of quartile two, depending on if the value is higher or lower than the second quartile [46].

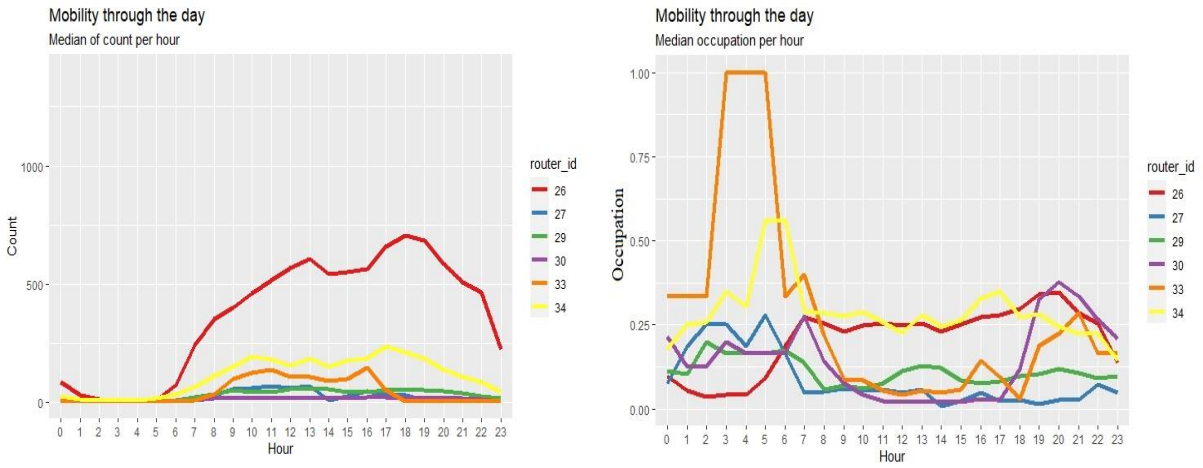

Fig. 22 Raw counts vs Occupation

Once the 4 years worth of data were pre-processed, we could tackle the second challenge of the passive Wi-Fi counts. The way we solve these discrepancies between one hundred count in the city centre and outside the city centre, was instead of using the raw count, a new value was created called *occupation* that refers to the percentage of occupation of that location according to a max (Fig. 22), this max is the higher value detected on a router since it was first turned on (we only extract this max after cleaning the data and not before, otherwise we would get wrong max values, for the reasons explained above). With this new variable, there is no longer the problem when comparing the counts from two different routers. Returning to the example above, there are two routers each with one hundred counts where one of the routers is in the city centre and the other outside, now with the variable the central router with one hundred counts actually transfer to 25% of occupation whereas the other router outside the city centre, one hundred counts is 85% of occupation. It will be this dataset with this new variable that will be used for the analysis.

With this processing and filtering we ended up with two different datasets, one for 2019 and 2020 with all the hourly occupation from April to September and other with the filtered counts (no outliers, no zero values, etc..) for the same time period.

4.3.6 Telecom Counts Processing

In order to validate the Passive Wi-Fi monitoring, it was given access to us several datasets regarding the movements between districts of Madeira. These movements agglomerate all devices that moved from one district to another and were possible to determine with the help of cell phone towers.

As this dataset was not compiled by us, the first step is to extract all the information needed with the preferred organization, in order to later use this dataset as a comparison. The initial datasets were roughly divided by week, for example, first week of April, second week of April, etc. For the purpose of our analysis, we used all the datasets from April to September.

We reduced the number of columns to 4, “Date”, “Time”, “District” and “Count”. Where “Time” is the the time of the day (this time has intervals of 2 hours, for example, 0h-2h, 2h-4h, etc.), and the “Count” is the sum of the values, from the initial dataset. Ending up with a dataset as in Fig. 23.

Date	Time	District	Count
2020-04-09	2.0002e+10	CALHETA	103
2020-04-10	2.0002e+10	CALHETA	108
2020-04-11	2.0002e+10	CALHETA	117
2020-04-12	2.0002e+10	CALHETA	119
2020-04-13	2.0002e+10	CALHETA	118
2020-04-14	2.0002e+10	CALHETA	106
2020-04-15	2.0002e+10	CALHETA	93
2020-04-16	2.0002e+10	CALHETA	101
2020-04-17	2.0002e+10	CALHETA	107
2020-04-18	2.0002e+10	CALHETA	101
2020-04-19	2.0002e+10	CALHETA	112
2020-04-20	2.0002e+10	CALHETA	111
2020-04-21	2.0002e+10	CALHETA	86
2020-04-22	2.0002e+10	CALHETA	96
2020-04-23	2.0002e+10	CALHETA	114
2020-04-24	2.0002e+10	CALHETA	107

Fig. 23 Final telecom dataset

Chapter 5

Exploratory Analysis

In this section, the datasets processed from the previous section will be used to do an analysis on Madeira Island from different perspectives.

The main purpose of this analysis is to detect patterns and relation between different variables such as population, routers count, etc. With that said we will divide this section into four different analyses:

- Relation between population and groups;
- Affinities between groups;
- Validation of Passive Wi-Fi Monitoring data with telecom data as Ground Truth;
- Mobility analysis before and during COVID-19 with Wi-Fi data;

5.1 Relation Between Population and Groups

When opening a new establishment with a certain purpose for the community, the stakeholders tend to have in mind several factors before starting the new business in a new location, such as, age and gender of population that lives near, the amount of population itself, interests and necessities of the general community, and many others, however one important factor is always the population.

We will attempt to detect this assumption and understand if the population in a certain place is correlated with the number of different establishments. These establishments are divided into groups, the ones discussed in OSM Data Processing Section being Commercial, Community, Educational, Entertainment, Financial, Government, Healthcare, Living, Sustenance, Tourism and Transportation, which represent different kinds of establishments and services available to the community. To get this relation, we used the Hex-Grid maps for the population and categories in the previous section and applied a linear regression to each hexagon in the map, ending up with eleven different scatterplots, as in Fig. 24.

Examining closely these results, we can get some interesting conclusions about the distribution of the different categories and the influence of population on the number of establishments available. For example, in the scatterplots the categories Commercial, Community, Sustenance, Tourism, Living and Transportation are the ones with a higher slope on the linear regression, meaning that there is a tendency as the population increases the more establishment of that category exists. Not only can we predict these assumptions with our data, as we can also suppose, with real world examples, why these categories have a higher relation with the population, if we think for the Living and Transportation category it is only logical to assume that the more population there is in one area the more of both categories should exist on that exact area. As for the Commercial, Community, Sustenance and Tourism, one of the reasons for this higher relation might just be out of necessity for the services provided for those establishments, for example, for supermarkets, restaurants, churches, and hotels (Commercial, Sustenance, Community and Tourism examples, respectively) there is also a tendency for the number of these establishments increase as the population increases too.

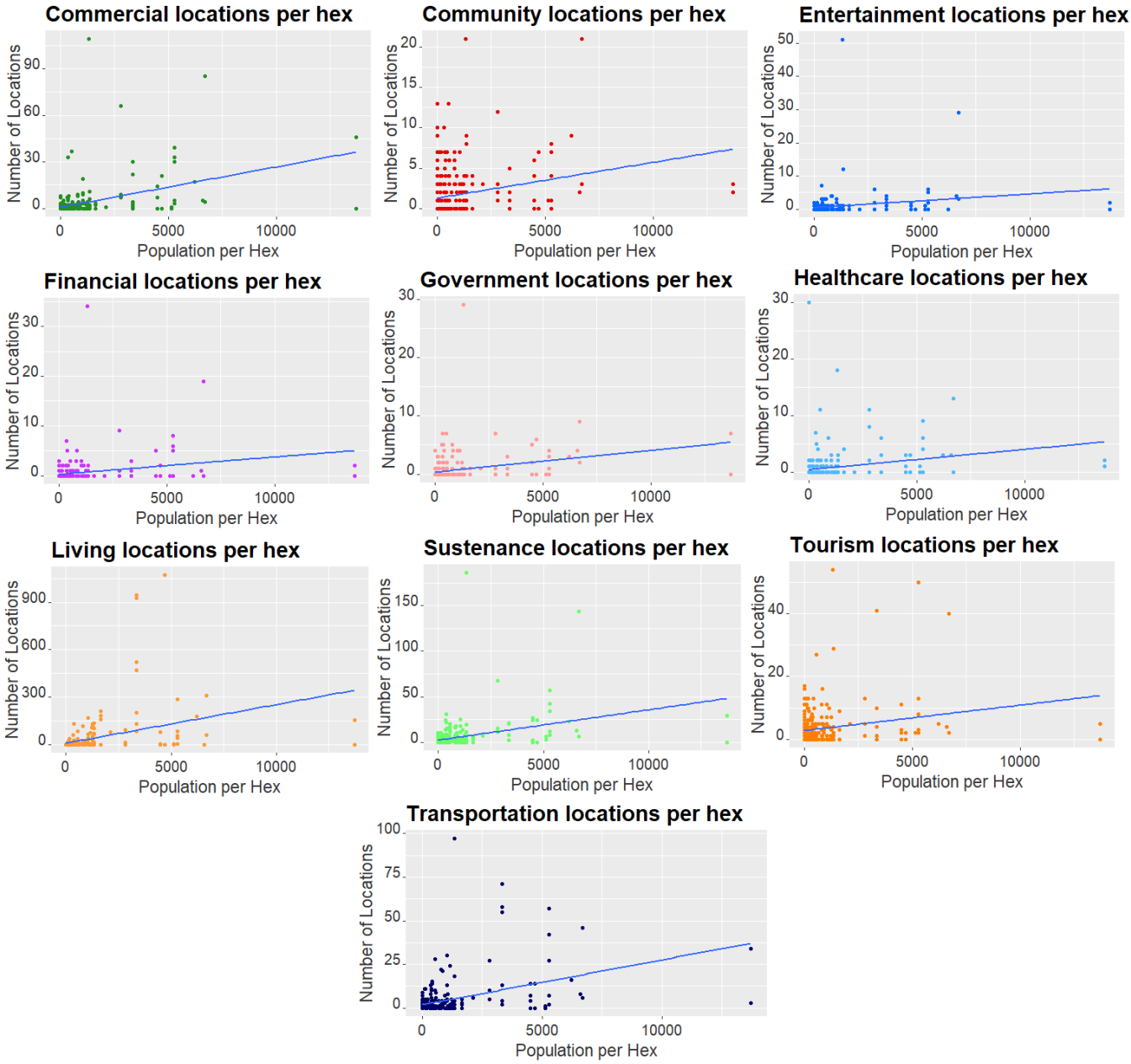

Fig. 24 Linear Regression between population and number of establishments

5.2 Affinities Between Groups

Having the number of different establishments in one location is related to the population, so we can also express the affinities between each category and understand rather or not one category tends to appear in the same areas as another category.

This was executed by using the hexagonal maps created in Area Categorization Section and joining them into one dataset, where each hexagon has the number of establishments for each group. Once this was done, the Pandas python library was used to do a Pearson correlation [37], which is a measure of the linear relationship between two features where in order to calculate it is needed to do the covariance of the two variables divided by the product of their standard deviations, as in the formula below (1). The result of this correlation can have several meanings depending on the value, such as:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other.
- Zero means that for every increase, there is not a positive or negative increase. The two just are not related.

$$\rho_{x,y} = \frac{COV(x,y)}{\sigma x \, \sigma y} \quad (1)$$

Looking at the heatmap created, it shows six values higher than 0.9, the relations between Financial – Entertainment, Financial – Sustenance, Financial – Commercial, Government - Financial, Commercial – Sustenance and Sustenance – Entertainment, which means in places where we find one of the categories there is a real good probability of also finding one of the other categories. We can also explain these affinities by translating to real world examples, we know that the Financial group is mostly ATM, so it makes sense that this Category appears in the same places of Entertainment, Sustenance, Commercial and Government (mostly banks and courts) since people might need to withdraw money to be used in these establishments. The explanation for the other two relations might be the fact that since Sustenance Category represents mostly Restaurants and Bars, it is normal that these establishments are normally situated in the same locations of Commercial and Entertainment for the service that Sustenance provides to the people that attend these locations. In the range from 0.8 to 0.9 we also have relations like Commercial – Entertainment, Commercial – Transportation, Government – Entertainment, Government – Sustenance and Government – Commercial.

Although we already described the higher valued relations there are some interesting results, that should be mentioned, which roughly represents the real world and how it works. Looking at the Living Category, immediately there are two values that stand out Transportation and Educational, although they are not very high, there is a larger discrepancy between these two and the other relations (these two values are higher than 0.5), which makes these two categories the ones with a higher affinity with the Living Category. When translating this to the real world it means that in places where people live should exist more schools and public transports. Also, by inspecting the Healthcare category and its relations we notice a slightly constant distribution between the other categories, which could translate to the evenly distribution of healthcare services in the island.

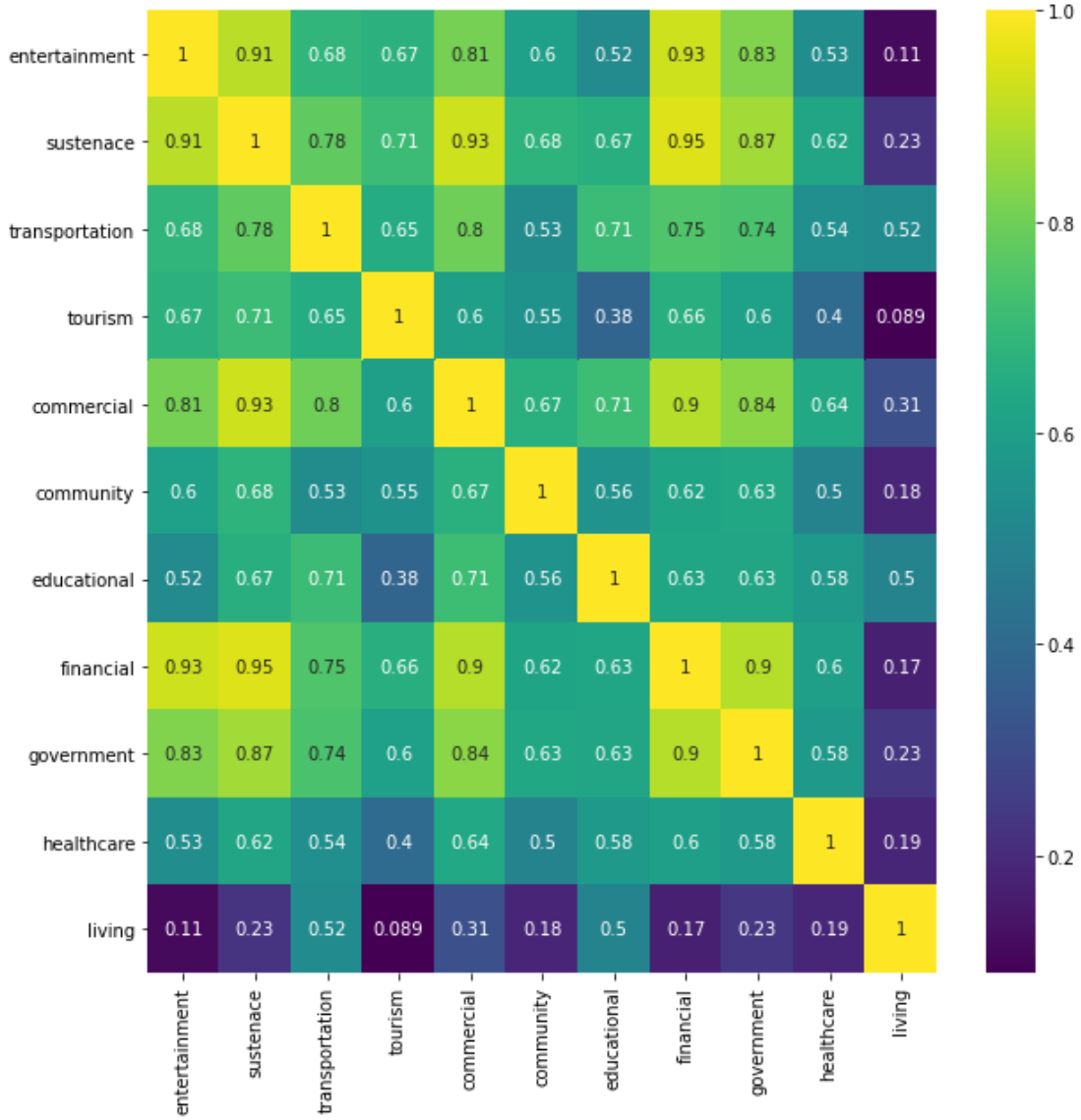

Fig. 25 Affinities between different groups

5.3 Validation of Passive Wi-Fi Monitoring data with telecom data as Ground Truth

Now that we did a more superficial analysis of Madeira Island regarding population and establishments, we can incorporate the Wi-Fi data to our analysis and understand if the different groups of establishments have any interference with the people mobility, in other words, if the locations with more mobility have anything in common between themselves and vice-versa. However before we proceed with this analysis, we need to validate our system, the Passive Wi-Fi monitoring system, and to accomplish that we will use the data from telecom provider as ground truth (Telecom company Data) that was already filtered and processed in Telecom Counts Processing Section.

Before we start with this analysis, we need to prepare both our datasets in order to validate one with another, because while for our Wi-Fi dataset (From the Passive Wi-Fi Monitoring System) we have the counts per router where each router has a precise location within a small area of a district. In the telecom dataset all counts are grouped by district, which has less preciseness regarding the area where the device was counted. Madeira Island is constituted by 11 districts as shown in Fig. 26, however with the router available from the Passive Wi-Fi monitoring System we can only analyse 9 of them, being Calheta, Porto Moniz, São Vicente, Santana, Machico, Santa Cruz, Funchal, Câmara de Lobos and Ribeira Barava.

Fig. 26 All districts of Madeira Island

Firstly, we grouped our Wi-Fi counts data (mentioned in Wi-Fi Data Section) by district, in order to get a single value per hour per district, and to do this we verified in which district each router is located and once we done this, we can calculate a new count for that district hourly.

Secondly, as mentioned in Section 3, this Wi-Fi infrastructure is maintained by the community itself and since the routers are distributed by public places, it may lead to some deactivate routers during night-time, with that in mind, to compare both datasets, we need to select a time in which both systems are fully operational, and after some analyses the time period selected was from 06:00h to 24h:00. Still in a filtering perspective and following the previous idea, as some routers may be deactivated during night-time it may also happen that there are some days in which both systems, the Wi-Fi and the telecom system, have no counts at all, so we need to filter the datasets to get only the dates where both systems are operational. With all these filtering we ended up with two datasets, as in Fig. 27, that we can use for our comparison, where the left one is the Wi-Fi dataset and the right one is the telecom dataset.

date	district	count
2020-04-09T23:00:00Z	CALHETA	158.33333
2020-04-09T23:00:00Z	CAMARA DE LOBOS	126.00000
2020-04-09T23:00:00Z	FUNCHAL	571.50000
2020-04-09T23:00:00Z	SANTA CRUZ	81.00000
2020-04-09T23:00:00Z	SANTANA	58.00000
2020-04-09T23:00:00Z	SAO VICENTE	469.00000
2020-04-10T23:00:00Z	CALHETA	86.33333
2020-04-10T23:00:00Z	CAMARA DE LOBOS	108.00000
2020-04-10T23:00:00Z	FUNCHAL	404.11111
2020-04-10T23:00:00Z	SANTA CRUZ	98.00000
2020-04-10T23:00:00Z	SANTANA	46.50000
2020-04-10T23:00:00Z	SAO VICENTE	333.00000
2020-04-11T23:00:00Z	CALHETA	393.00000
2020-04-11T23:00:00Z	CAMARA DE LOBOS	135.00000
2020-04-11T23:00:00Z	FUNCHAL	885.36842
2020-04-11T23:00:00Z	SANTA CRUZ	102.00000

date	district	movement
2020-04-09T23:00:00Z	CALHETA	6119
2020-04-09T23:00:00Z	CAMARA DE LOBOS	60132
2020-04-09T23:00:00Z	FUNCHAL	67425
2020-04-09T23:00:00Z	SANTA CRUZ	38607
2020-04-09T23:00:00Z	SANTANA	5582
2020-04-09T23:00:00Z	SAO VICENTE	2981
2020-04-10T23:00:00Z	CALHETA	6911
2020-04-10T23:00:00Z	CAMARA DE LOBOS	64247
2020-04-10T23:00:00Z	FUNCHAL	72381
2020-04-10T23:00:00Z	SANTA CRUZ	40801
2020-04-10T23:00:00Z	SANTANA	6124
2020-04-10T23:00:00Z	SAO VICENTE	3294
2020-04-11T23:00:00Z	CALHETA	6915
2020-04-11T23:00:00Z	CAMARA DE LOBOS	64133
2020-04-11T23:00:00Z	FUNCHAL	72286
2020-04-11T23:00:00Z	SANTA CRUZ	40806

Fig. 27 Wi-Fi Data (Left) and telecom Data (Right)

With the first two phases of filtering and processing completed, we can begin to examine if the Passive Wi-Fi monitoring system identifies the ups and downs on the mobility count according to the telecom dataset, in other words if the Wi-Fi count increases/decreases when the telecom counts increases/decreases too, with that said the purpose is not for the counts to be similar but for the “shape” to be similar. So, the next step is to normalize the counts from both datasets (“counts” column for the Wi-Fi dataset and “movement” column for the telecom dataset), since there is a clear discrepancy between the two counts. This happens because there are not enough routers to cover an entire district area, whereas the telecom infrastructure can cover a much larger area with the cell phone towers, these is why we need to normalize the values first. And to do that we used the StandardScaler library from sklearn.preprocessing in python.

Now that all the datasets are ready, we can verify if the Passive Wi-fi Monitoring System can give us true information about the mobility on the island. And to check the veracity of the previous statement, we will do two analyses that compares the Wi-Fi counts with the telecom counts and discuss the results of the different analyses, being:

- Line-chart Analysis, with time series and the counts of both Wi-Fi and telecom;
- Spearman Correlation (also known as Spearman's rho) between both counts;

It is important to consider that the Passive Wi-Fi monitoring System does not have the same coverage in all districts, in other words, we do not have the same number of routers in all districts, so it is only normal to some districts have better results than others. With that said, one of the objectives with these analyses is to gather information to help us improve our infrastructure, for example, increase the numbers of routers in certain districts, etc.

5.3.1 Line-Chart Analysis

This initial analysis is for us to get an overview on the behaviour of both counts. And will also provide us important insight to determine in which districts the Passive Wi-Fi monitoring system may have more accuracy and why.

When creating this graph, our main goal was to get a visualization that allowed us to get a time perspective of the counts. So, since our datasets are already filtered, which means we only consider the days that both systems are working simultaneous, we can create a line chart with two lines one for the Wi-Fi and other for telecom, where the x-axis is the time series and the y-axis is the scaled counts, ending up with one chart per district as shown in Fig. 28.

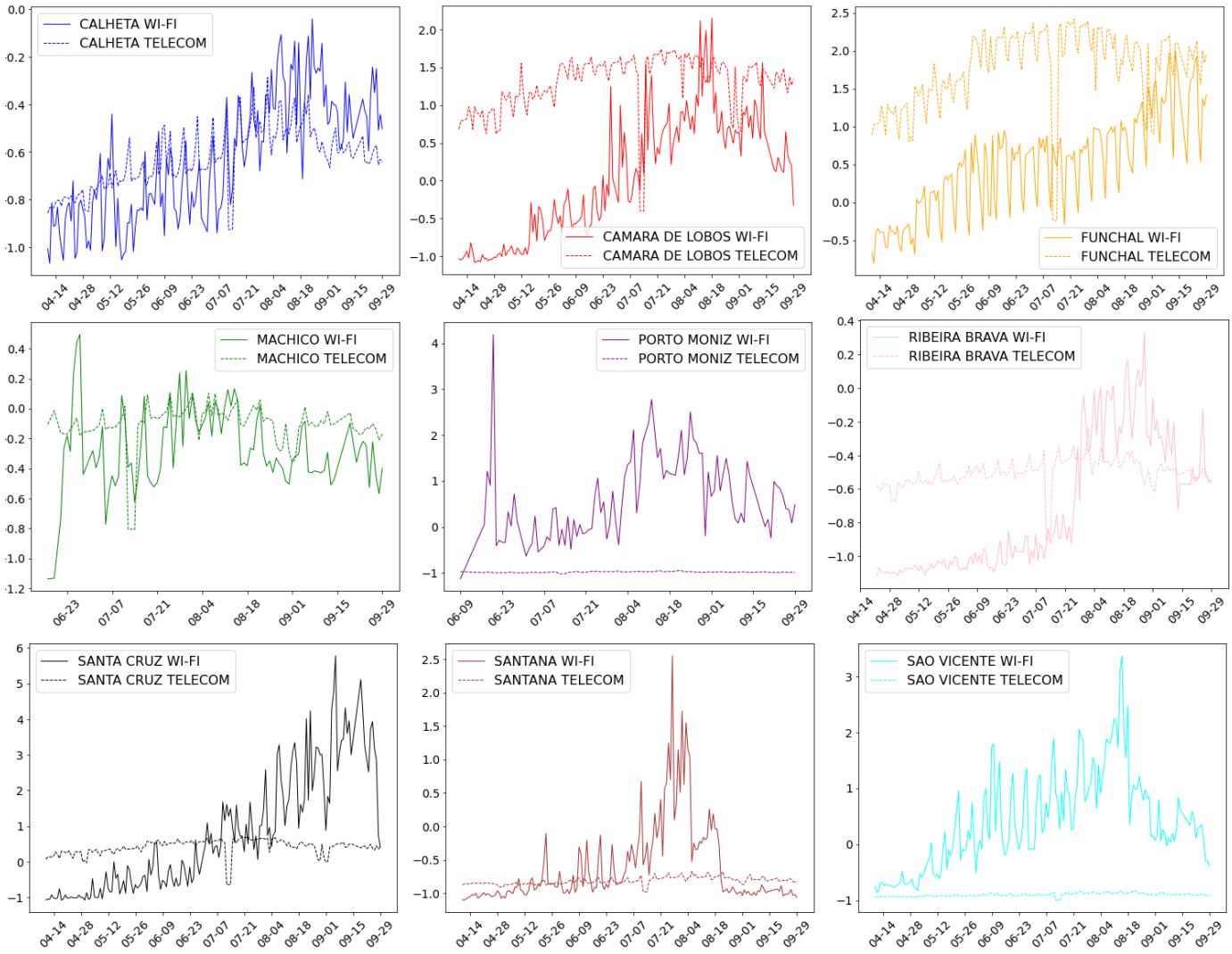

Fig. 28 Line charts for all districts: x- date, y- counts

5.3.2 Spearman Correlation

For our second analysis to validate the Passive Wi-Fi monitoring system with the telecom dataset, we proceed to do a Spearman Correlation (also known as Spearman's rho) [38]. This correlation is used when the relationship between two variables is not linear, in other words instead of measuring the strength of a linear relationship as in Pearson, it measures the strength of a monotonic relationship between paired data.

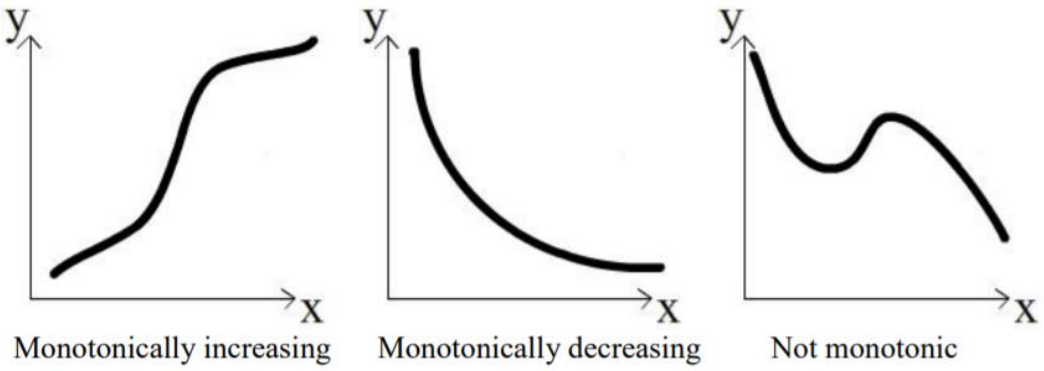

Fig. 29 Monotonic function examples [38]

A monotonic function is one that either never increases or never decreases as its independent variable increases. For example:

- Monotonically increasing - as the x variable increases the y variable never decreases;
- Monotonically decreasing - as the x variable increases the y variable never increases;
- Not monotonic - as the x variable increases the y variable sometimes decreases and sometimes increases.

With this correlation what we are trying to accomplish is to determine if the variables tend to move in the same relative direction, however not necessarily at a constant rate, because as we said previously on the beginning of Section 5.3 we cannot cover all the districts areas with the routers available to us, so it is only logical that the increase of both variables are not in the same proportion, nevertheless what we might get with this correlation is if they move in the same direction (if they increase or decrease together). Spearman Correlation can go from -1 to +1, where each limit represents a perfect monotonic relationship, whereas -1 represents a negative monotonic relationship (Monotonically decreasing), +1 represents a positive monotonic relationship (Monotonically increasing).

Now that we explained how the Spearman Correlation works, we can apply it to our data. To do this, we joined the two datasets (Router's data and the telecom data) and applied the function *corr* available in Pandas library in python which allowed us to choose the method we want (method="spearman"). To create the heatmap as in Fig. 34, we used seaborn library also available in python.

5.3.3 Discussion of the Results

With these two previous analyses we can extract some important information regarding the validation of the Passive Wi-Fi Monitoring System. With the first analysis on Line-Chart Analysis, we can already get an overview of the behaviour of both systems. On the second analysis in Spearman Correlation, the objective is to inspect in more dept the values and how the increase/decrease one variable affects the other.

Looking at the first analysis Fig. 28, by doing a quick observation on all graphs there are 5 districts that stand out positively from the others for having a very similar curve between both systems, being Calheta, Câmara dos Lobos, Funchal, Machico and Ribeira Brava (Fig. 30).

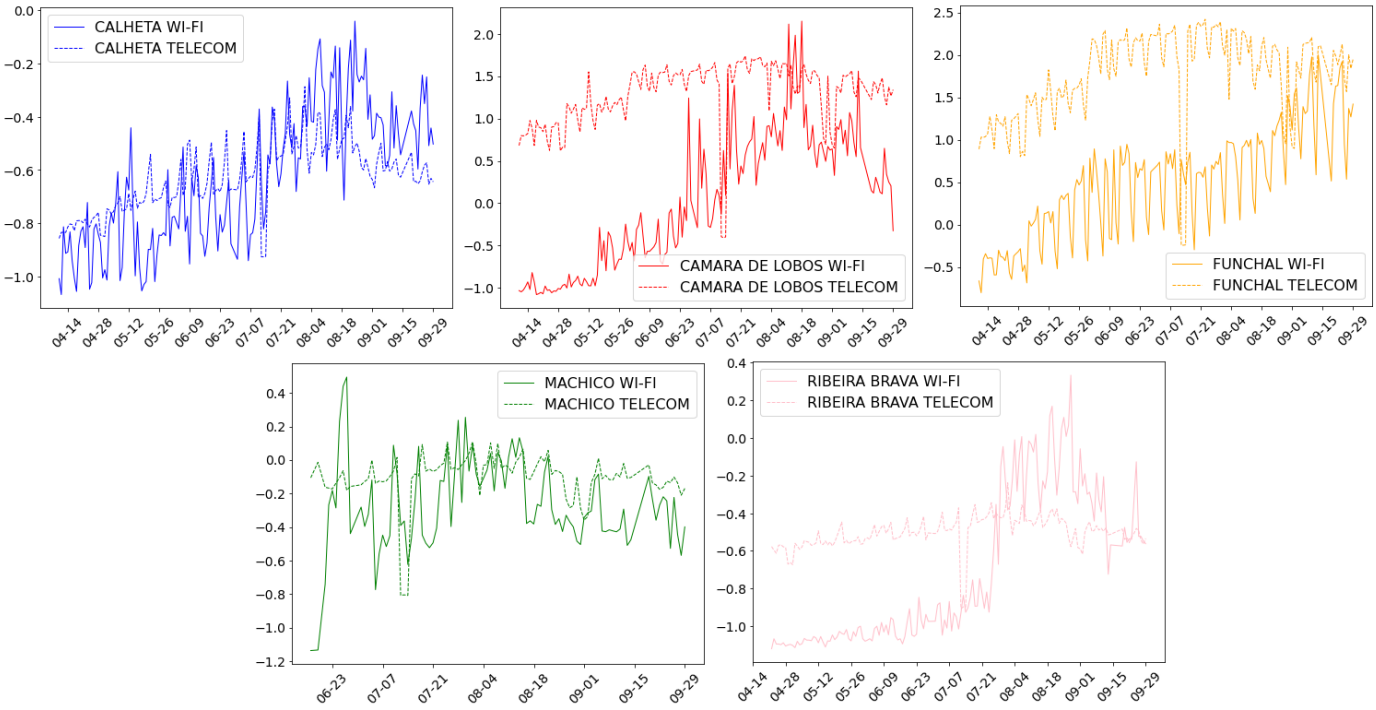

Fig. 30 Line charts with similar lines shapes between telecom and Wi-Fi

By having these similar lines, we can identify correspondent peaks and lowest points on both lines which may be a positive indication that the Passive Wi-Fi Monitoring System can translate the real mobility within each district. However, we are yet to analyse the other 4 districts (Fig. 31) because despite at first glance it appears that there is no correlation, we can still see some peaks, although small, in the telecom line and we need to verify if these peaks have any correspondence with the router's peaks.

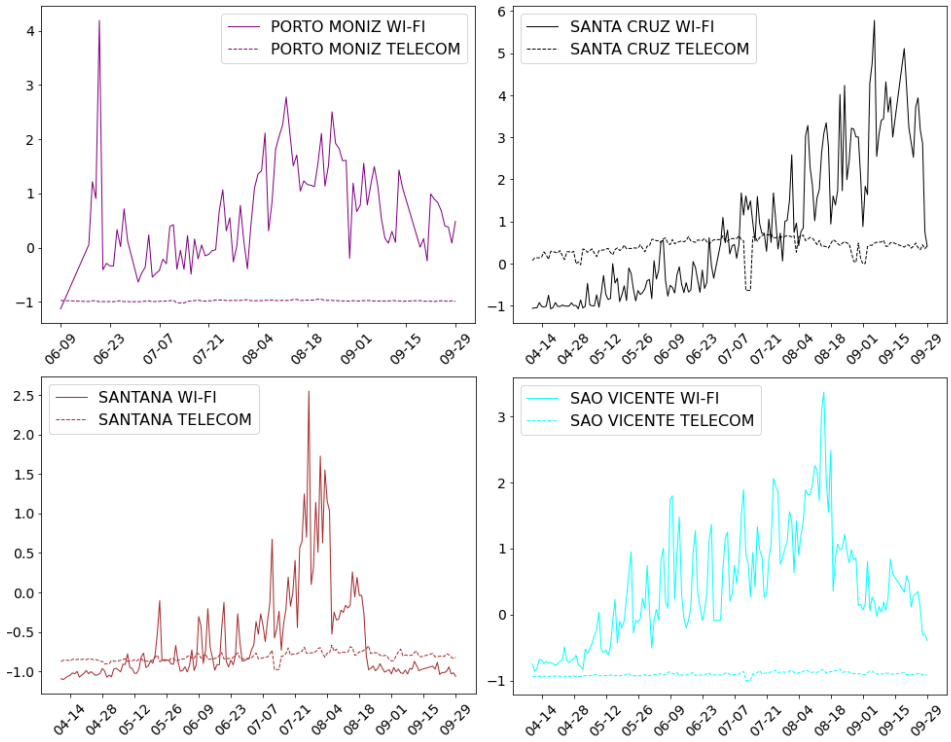

Fig. 31 Line charts with different lines shapes between telecom and Wi-Fi

One of the reasons we cannot identify the shapes of the lines on Fig. 31 is the scale. The scale on the charts is not adequate to identify peaks of the telecom lines, because of the small changes they have whereas in the Wi-Fi line we can clearly identify those peaks. To overcome this drawback, a second cluster of charts was created with a secondary axis, having two y-axes, one scaled accordingly to the Wi-Fi data (left axes) and other to the telecom data (right axes). Ending up with an all-new set of charts as in Fig. 33.

Looking back at the districts from Fig. 31, we can already identify a similar line shape between both systems, for example on Fig. 32, for São Vicente district with the new secondary axis it is easier to visualize the ups and downs of both systems and thus it is easier to validate the Passive Wi-Fi Monitoring system.

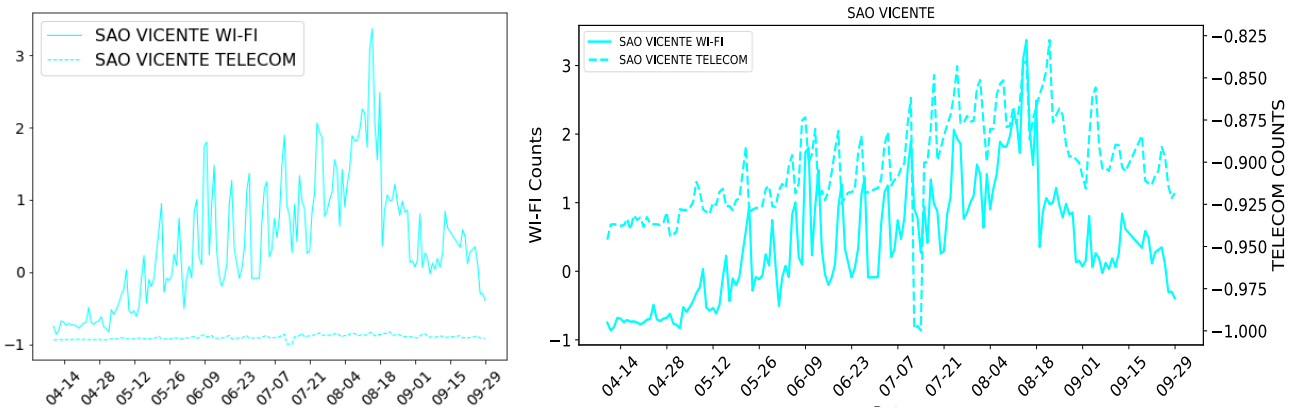

Fig. 32 Before (left) and after(right) adding a secondary axis

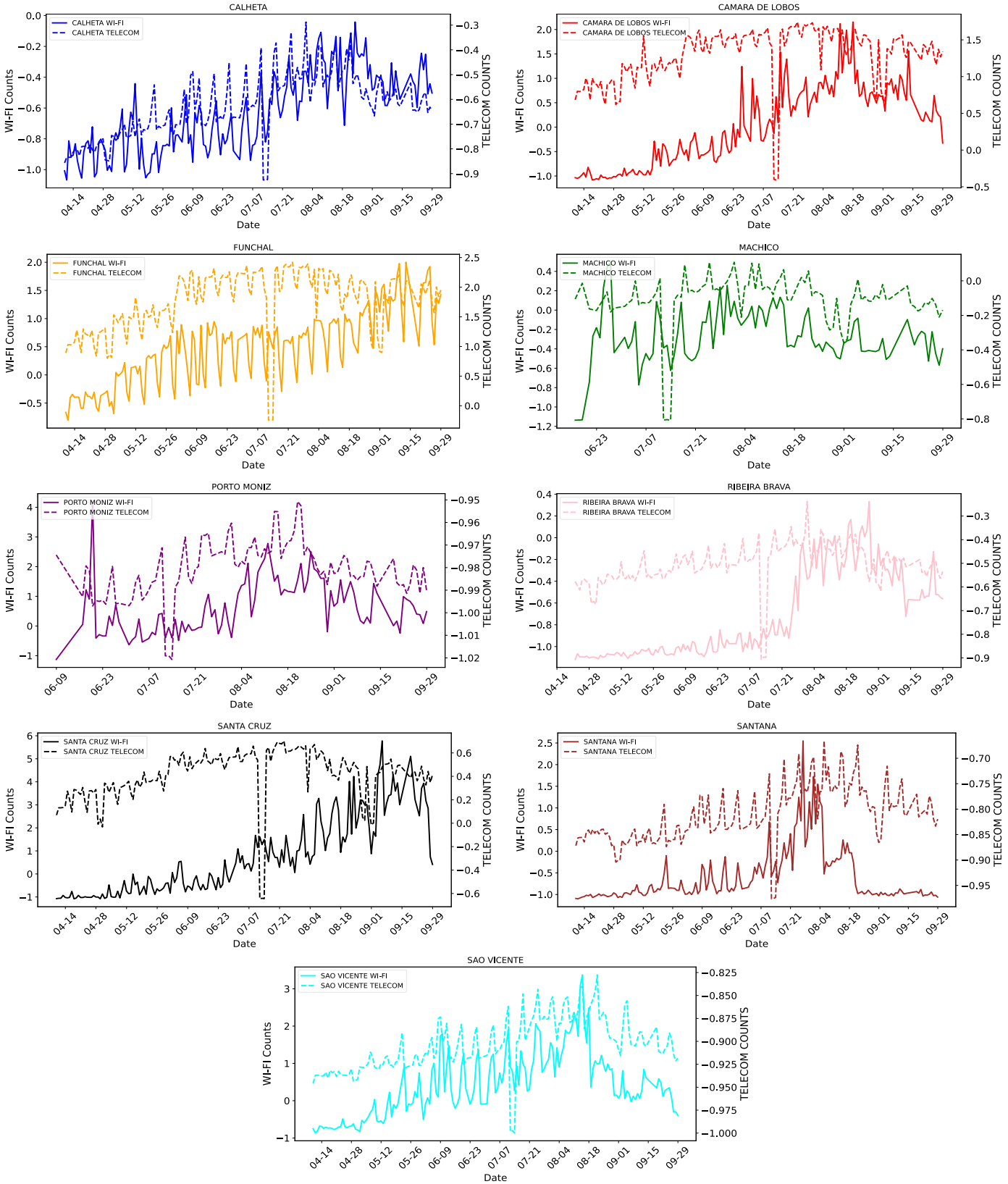

Fig. 33 Line charts with secondary axis

With this last analysis in mind, the Spearman Correlation from 5.3.2 becomes useful by giving us additional information, on how the Wi-Fi values and the telecom values relate to each other, which we cannot detect by doing a simple visual analysis. With the Spearman correlation, as mentioned above, we can do a core analysis to the values its selves and verify if there are no districts which have 0 or negative values for the correlations between both systems. By applying the Spearman Correlation to our values, we ended up with a heatmap as in Fig. 34.

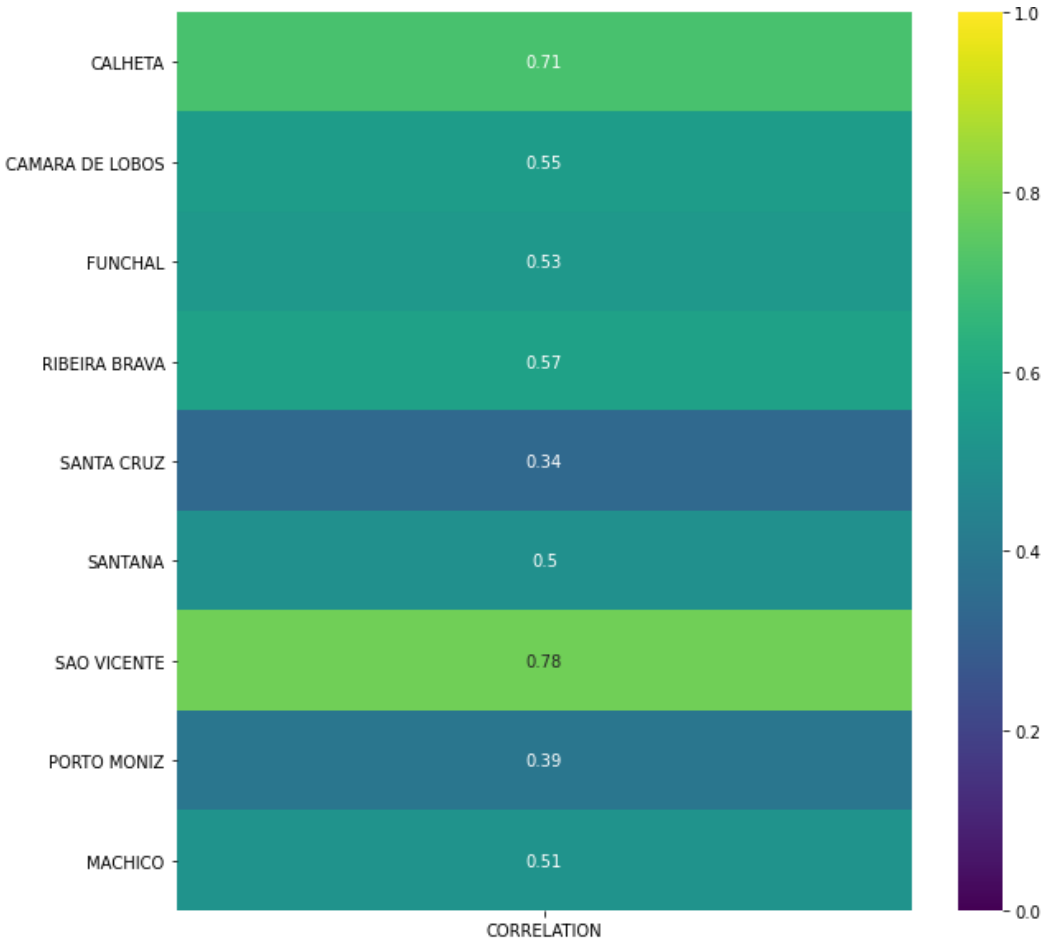

Fig. 34 Heatmap with Spearman Correlation between telecom and Wi-Fi counts

Looking at the heatmap, there are no negative values for the correlation, which by itself it is already an assertive indicator for the validation of the Passive Wi-Fi monitoring system, however there are higher values than others. In average most districts, such as Câmara dos Lobos, Funchal, Ribeira Brava, Santana and Machico, have a correlation between 0.5 - 0.6 which we can already consider a strong correlation, meaning that on those districts the Passive Wi-Fi Monitoring System translates the existent mobility with an accuracy higher than the average. Then there are two districts with the highest correlation values, Calheta and São Vicente where the system (Passive Wi-Fi Monitoring) has a good accuracy in determining mobility patterns.

The reason behind these results can be several nevertheless some important factors are the coverage

we have with routers within a district area and of course the location of the router itself. Likewise, imagine a system of buoys that detects tides and tsunamis, we cannot have only one buoy on Atlantic Ocean, we need several of them to do a precise analysis, the same applies to our routers, we need enough coverage on areas where people might pass by. One thing the districts with a correlation value higher than 0.5 have in common is this exact coverage, the routers are distributed throughout the areas with population where there is also more probability to have mobility. It is important to keep in mind that there might be other factors as the installation location of the router if it is inside a building or not.

Following the previous idea if we analyse the two districts, Porto Moniz and Santa Cruz (highlighted in Fig. 35), with the lowest values of correlation we can see some important elements, mentioned above, that might interfere with the accuracy of our system. If we look at the routers location (top left map from Fig. 35) and the population areas (bottom map from Fig. 35) and narrow our analysis only for the districts with the lowest correlation values, we can affirm with factors such as coverage and location as backup to our affirmation because in Santa Cruz we do not have enough routers to cover such a big area and in Porto Moniz the routers could be more spread out through the population areas.

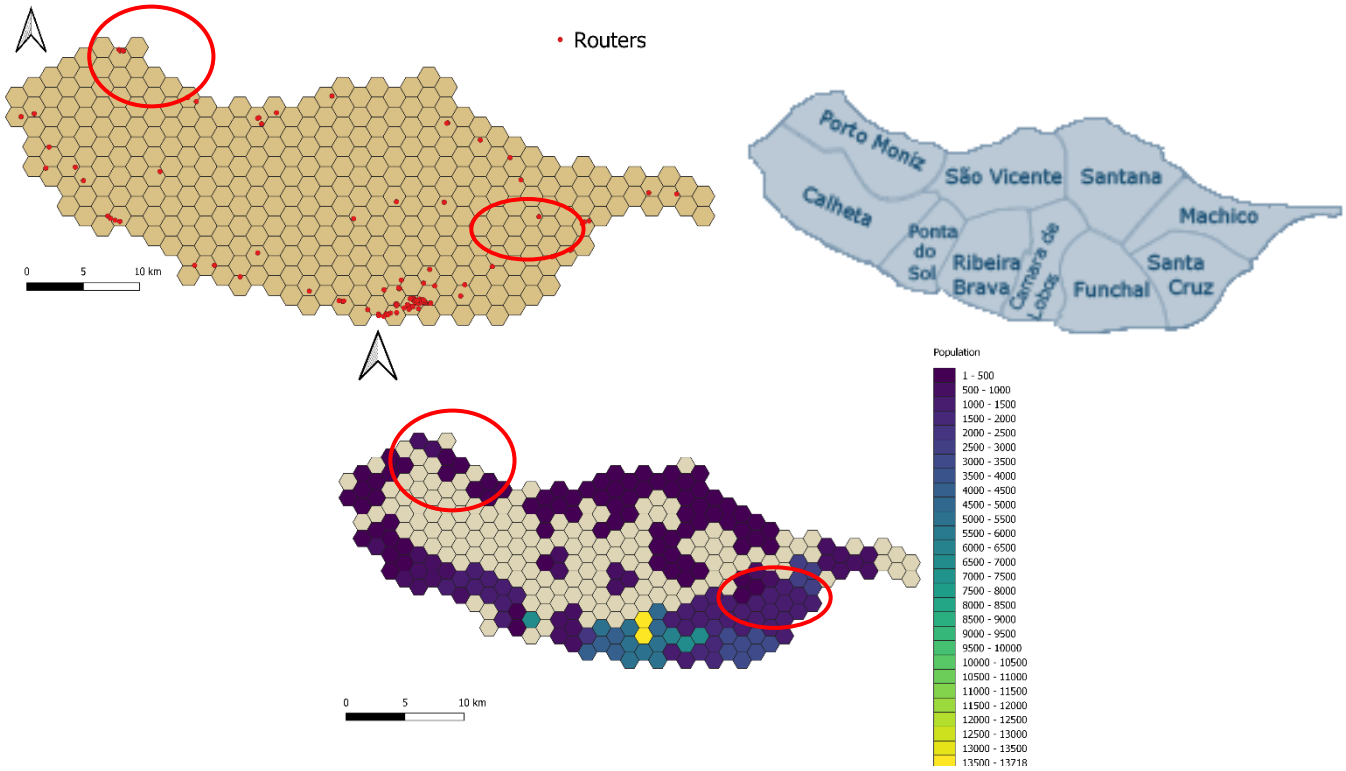

Fig. 35 Top Left - Routers location (with Santa Cruz and Porto Moniz highlighted), Top Right – Districts areas, Bottom – Population Distribution (with Santa Cruz and Porto Moniz highlighted)

To summarize, with these analyses we can state that the Passive Wi-Fi Monitoring System can successfully translate the behaviour of mobility within the island, despite two districts do not have the same accuracy as the others we still can withdraw essential behaviours from peoples' mobility within these districts. With that said we will proceed to use the Wi-Fi Counts to do a new analysis on understanding mobility patterns of people before and during the pandemic.

5.4 Mobility Analysis before and during COVID-19 with the Wi-Fi data

With the validation of the Wi-Fi data, this dataset coupled with the establishments (POIs) dataset can help getting another understanding in the people mobility by inspecting the types of establishments which are more popular. Also, since we have the data from 2019 and 2020, we will analyse the differences, if there are any, between the mobility without the pandemic and with the pandemic on the island.

To do this analysis we will use a method called Principal Component Analysis, Principal component analysis (PCA) in many ways forms the basis for multivariate data analysis. PCA provides an approximation of a data table, a data matrix, X, in terms of the product of two small matrices T and P'. These matrices, T and P', capture the essential data patterns of X. Plotting the columns of T gives a picture of the dominant "object patterns" of X and, analogously, plotting the rows of P' shows the complementary "variable patterns" [24]. The starting point in all multivariate data analysis is a data matrix (a data table) denoted by X. The N rows in the table are termed "objects". These often correspond to chemical or geological samples. The K columns are termed "variables" and comprise the measurements made on the objects. For our case, our variables will be mobility and the number of each kind of POIs group, and the objects are all the mobility observations in time, with this we aim to get this "patterns" between variables, more prominently the correlations between the mobility variable and the POIs groups, in order to identify some differences in the places that people more frequent attend before and during the pandemic. From here, PCA reduces the high-dimensional interrelated data to low-dimension by linearly transforming the old variable into a new set of uncorrelated variables called principal component (PC) while retaining the most possible variation (statistical information) [25].

As mentioned previously, the objective here is to find any kind of relationship between mobility and the POIs groups using the PCA, however there are 11 different variables that represent different kind of establishments being Commercial, Community, Educational, Entertainment, Financial, Government, Healthcare, Living, Sustenance, Tourism and Transportation (Section 4.3.1). That is to say that there are too many variables to feed to PCA and if we want better results, we need to do a variable reduction first. And to do that we need to know which variables can be grouped together provided that we know the correlation between themselves, and since this analysis was done in Section 5.2 we can use these results to combine our variables. Looking at Fig. 25 and our analysis we can do 4 different clusters as in Table 2, one called *Services* which will agglomerate Financial, Government, Sustenance, Entertainment, Commercial and Community, another called *Lifestyle* with all the Transportation, Educational and Living points, and finally we will have two different clusters one for *Tourism* and other for *Healthcare*, since they are the ones that are more distributed through the hexagons.

Table 2 Division of the Groupers per the Clusters

Cluster	List of groups
Services	Financial, Government, Sustenance, Entertainment, Commercial and Community
Lifestyle	Transportation, Educational and Living
Tourism	Tourism
Healthcare	Healthcare

Similarly to the POIs hexagons distribution, we will also need the counts of the routers from the Wi-Fi dataset per hexagons instead of per router. In this analysis we will use the occupation that was processed in Section 4.3.5, so when we mention counts, we mean occupation of the router. To get the occupation per hexagon we sum all the occupation per router and then do an average per hexagon, ending up with a dataset where to each day has an occupation per hexagon.

With the dataset as in Fig. 36, the next step before feeding the PCA model is normalize the datasets since there are different scales in the variables and PCA is very sensitive to differences in the values. To do that we will use the StandardScaler from the sktlearn library (x minus mean divided by standard deviation) in our dataset (Fig. 36 right).

mobility	services	tourism	lifestyle	healthcare
1.8255	5	3	2	0
2.09009	16	9	8	1
1.40962	38	16	26	1
0.99335	17	5	16	1
0.759351	27	10	102	0
0.362644	19	5	2	0
1.45027	44	7	42	2

mobility	services	tourism	lifestyle	healthcare
-0.261314	-0.603552	-0.780832	-0.854018	-0.636208
-0.184035	-0.496281	-0.378943	-0.790968	-0.420332
-0.382785	-0.281738	0.0899272	-0.601819	-0.420332
-0.504367	-0.486529	-0.646869	-0.706902	-0.420332
-0.572713	-0.38901	-0.311961	0.196812	-0.636208
-0.688582	-0.467025	-0.646869	-0.854018	-0.636208
-0.370913	-0.223226	-0.512906	-0.433686	-0.204457

Fig. 36 Dataset (called X) to feed to PCA. Left – not normalize, Right – normalized

In PCA, we split covariance (or correlation) matrix into two parts, the scale part (eigenvalues), which gives us the importance of each Principal Component, and the direction part (eigenvectors). We may then endow eigenvectors with the scale: loadings. So, loadings are thus become comparable by magnitude with the covariances/correlations observed between the variables, in other orders, gives us the importance of each variable to each Principal Component (PC). A basic assumption in the use of PCA is that the score and loading vectors corresponding to the largest eigenvalues contain the most useful information relating to the specific problem, and that the remaining ones mainly comprise noise. Therefore, these vectors are usually written in order of descending eigenvalues [24].

Once our dataset was ready and we knew how the PCA works we could perform the model and analyse its results. As said before, we want to analyse which kind of establishments suffered a higher impact with the pandemic by analysing the correlation between these establishments and the mobility. To perform the Principal Component Analysis will use the PCA method from sktlearn library [39] with “n_components=5” as arguments which will return five Principal Components. With these in mind, we will do two separate PCAs, one for 2019 and other for 2020, in order to determine the differences at the end.

5.4.1 Analysis for 2019 Wi-Fi data

The first thing to do before analysing the result is to choose the best PCs to do our analysis, and we do this by picking the ones with a higher eigenvalue (normally the first three). Looking at the array below with all the components eigenvalues, the top three higher values will be called PC1, PC2 and PC3, with the values 3.63100747, 0.87388791 and 0.2900972 respectively.

[3.63100747, 0.87388791, 0.2900972, 0.1873778, 0.02393479]

With these three PCs we get a variance of 95.78%, meaning we have roughly 96% of the statistical information from the original dataset, as show in Fig. 37.

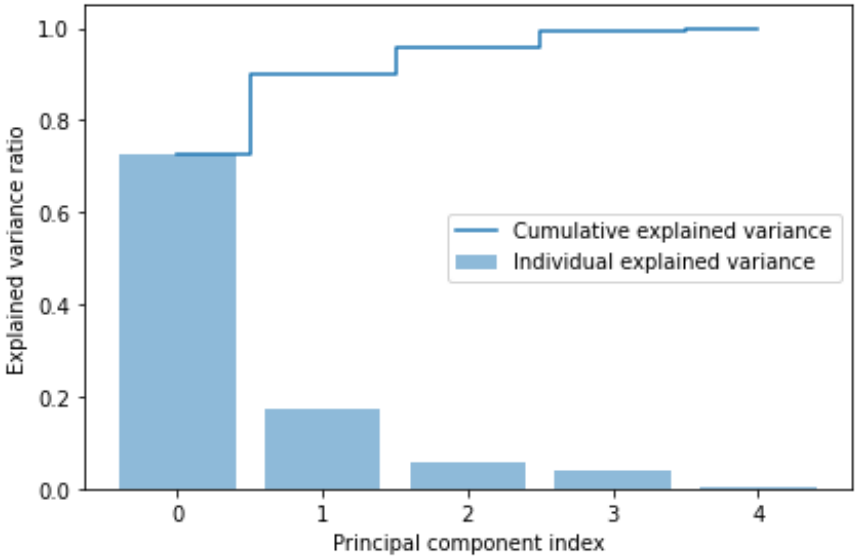

Fig. 37 Explained Variance of all PC for 2019

The next step is to create a visualization which will allow us to get a perspective on how each variable behaves with the mobility, and the best way to visualize this, is in a 2D chart with all the eigenvectors. And here is where the Principal Components enters, with these new variables (PC1, PC2 and PC3) we can now plot a 2D chart, with two of the three PCs and still have a good percentage of the original data information, which before was impossible because of the five variables we had. To choose the two PCs that will construct our chart we need to check the loadings of each variable on the Principal Components and see in which PCs the mobility has a higher impact. The loading as mentioned above, are the coefficients of the linear combination of the original variables from which the principal components (PCs) are constructed.

In Fig. 38, we can see the absolute values for the loadings for each Principal Component, and by doing a close inspection the two PCs with the highest loading in mobility is the PC1 and PC3. Ending up with 78.32% of the statistical information from the original data.

Index	mobility	services	tourism	lifestyle	healthcare
PC1	0.48357	0.503004	0.421748	0.285471	0.503767
PC2	0.0899512	0.163752	0.452376	0.871217	0.0378109
PC3	0.174476	0.291591	0.779309	0.350822	0.3926

Fig. 38 Loadings of each variable for the three highest Principal Components for 2019 (Absolute values)

With the two PCs, which will build our graph, chosen, the next phase is to plot the chart with all our observations, and also the vectors of the direction of each variable couple with a unit circle to better visualize then angle that each variable does with the mobility variable. If the angle is near 0 degrees, we are looking at a perfect positive correlation, where when one variable increases the other also increases, if the angle approximates to 180 degrees the correlation is also perfect however it is a negative one, where when one variable increases the other decreases, finally if the angle is close to 90 degrees there is no correlation between the variables. We will discuss these results side by side with the results from 2020 in Section 5.4.3.

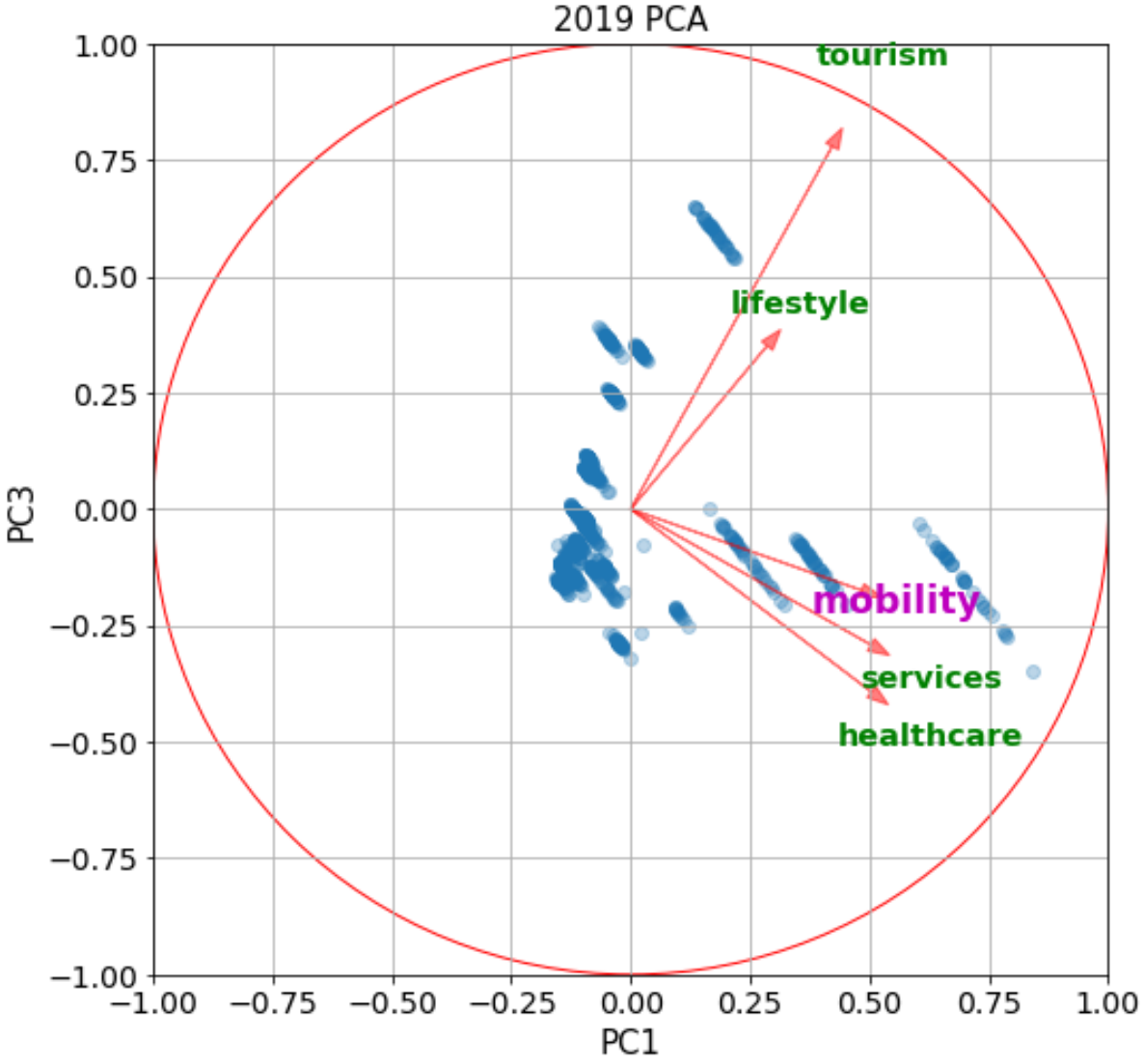

Fig. 39 Scatterplot with all the observations for 2019 and the loadings for PC1 and PC3

From this scatterplot alone in Fig. 39, we can already take important info regarding what makes people move to a certain area and at the same time extract some key factors that might be influencing the mobility on the areas with lower/higher mobility, such as accesses to certain areas (good/bad accesses), the location itself or even how well each establishment are distributed through the island. By looking at chart we see that tourism does not have a very high impact on mobility this might be because a high percentage of the mobility are locals and not tourists, and locals usually do not attend touristic places

as often as tourists, which may lead to this discrepancy. What both tourists and locals tend to move to is *services* locations, with restaurants, stores, and ATMs so it is only normal that these types of location have a higher correlation with mobility. We can verify that *healthcare* has a lot of correlation with mobility, two reasons for this could be, the fact that this cluster agglomerates all the types of medical places, from clinics, veterinarians, dentists to hospitals itself and people tend to have appointments all year, and second, being the waiting line and the people that accompany the person with the appointment. If we think about hospitals in general it tends to create a big waiting line and people usually do not go alone to an appointment (especially children), so instead of being one person in the hospital per appointment it can be 3 to 4 people, making a big slice on the mobility. Withal, the second factor might be the one that influences more the mobility, and it will also be the one that might make the difference in the 2020 analysis.

Regarding the routers infrastructure, this plot can also help us improve the infrastructure by giving us information where we might want to add some extra coverage with more routers, such as, locations with lifestyle points, because we might get lower mobility in these locations because of the number of routers we have on those areas, so by adding extra coverage we can eliminate this factor and get even better results.

5.4.2 Analysis for 2020 Wi-Fi data

As in the previous analysis with PCA for 2019 data, we first need to pick the PCs (Principal Component) with a higher eigenvalue (normally the first three). As we are using a different dataset for the mobility it is only normal that the results from the PCA are going to be different, so we need to do an all-new analysis. Looking at the array below, the top three higher values will be called PC1, PC2 and PC3, with the values 3.44352511, 0.83873089 and 0.40636178 respectively

[3.44352511, 0.83873089, 0.40636178, 0.26907177, 0.04805098]

With these three PCs we get a variance of 93.66%, meaning we have roughly 94% of the statistical information from the original dataset, as show in Fig. 40.

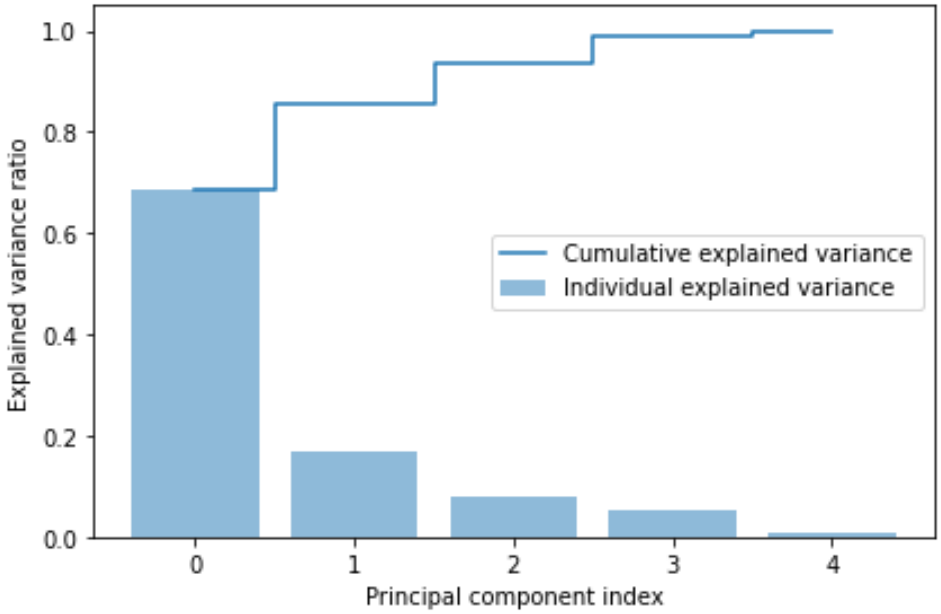

Fig. 40 Explained Variance of all PC for 2020

The next step here is again to create a 2D visualization in order to have a comparison to the chart created with the 2019 data. To choose the two PCs that will construct our chart we need to check the loadings of each variable on the Principal Components and see in which PCs the mobility has a higher impact. And again, as the dataset for the mobility is different the loading for each variable will be also different.

In Fig. 41, we can see the absolute values for the loadings for each Principal Component, and the two PCs with the highest loading in mobility is the PC1 and PC3. Ending up with 76.91% of the statistical information from the original data.

Index	mobility	services	tourism	lifestyle	healthcare
PC1	0.449019	0.521542	0.444909	0.274063	0.50331
PC2	0.115035	0.106588	0.335191	0.929	0.00351402
PC3	0.794763	0.0719487	0.588218	0.121894	0.0481411

Fig. 41 Loadings of each variable for the three highest Principal Components for 2020

With these loading we can plot the chart with the vectors for each variable and also all the observation from 2020.

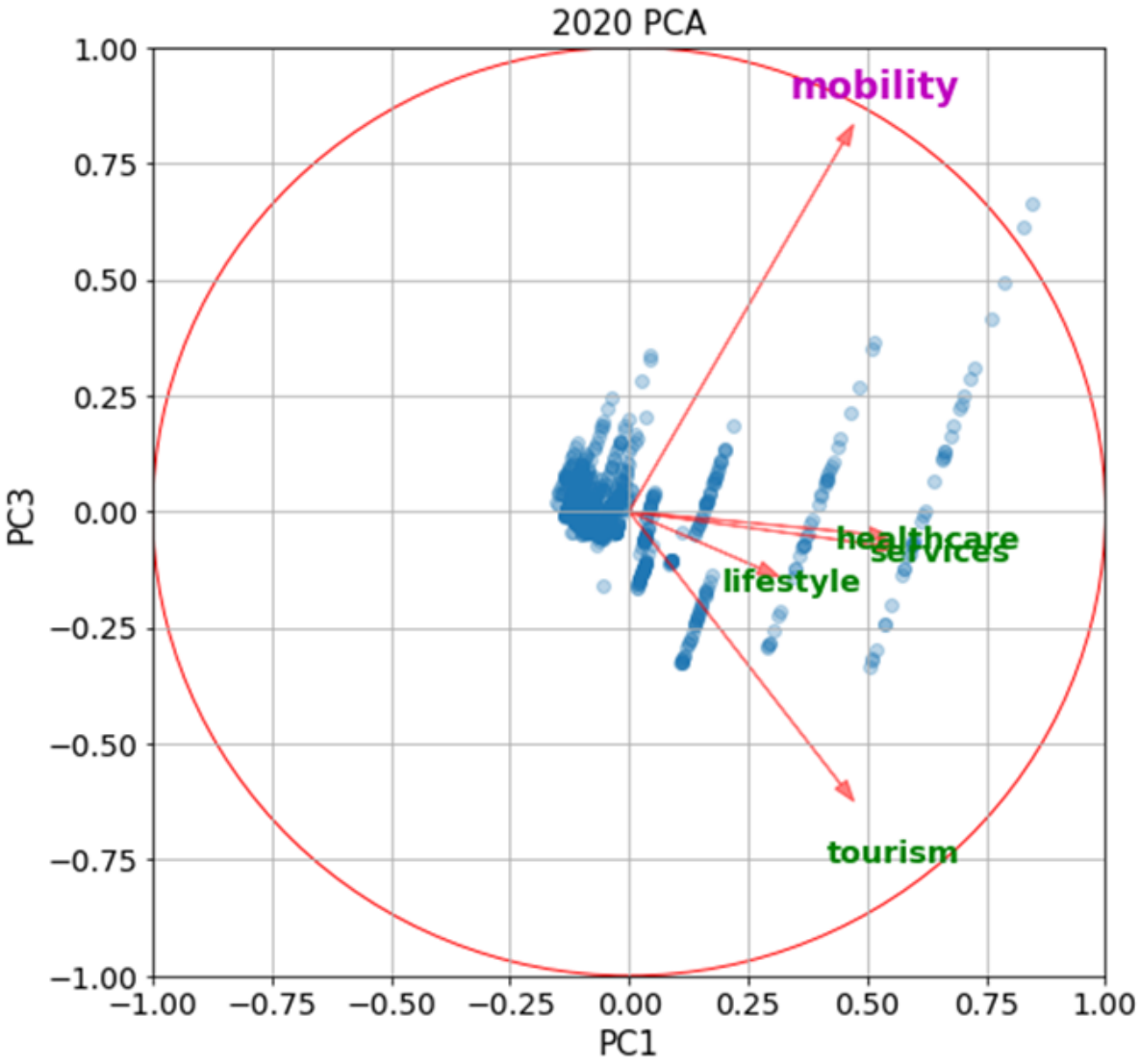

Fig. 42 Scatterplot with all the observations for 2020 and the loadings for PC1 and PC3

In the next section we will interpret both plots (from 2019 and 2020) side by side and discuss the results and analyse some causes that might have interfered with the differences in the mobility.

5.4.3 Discussion of the results from the PCA (2019 vs 2020)

With both PCA analyses performed to our data it is now possible to evaluate the results and compare them, in order to do this, both plots will be displayed below, side by side, for an easier interpretation.

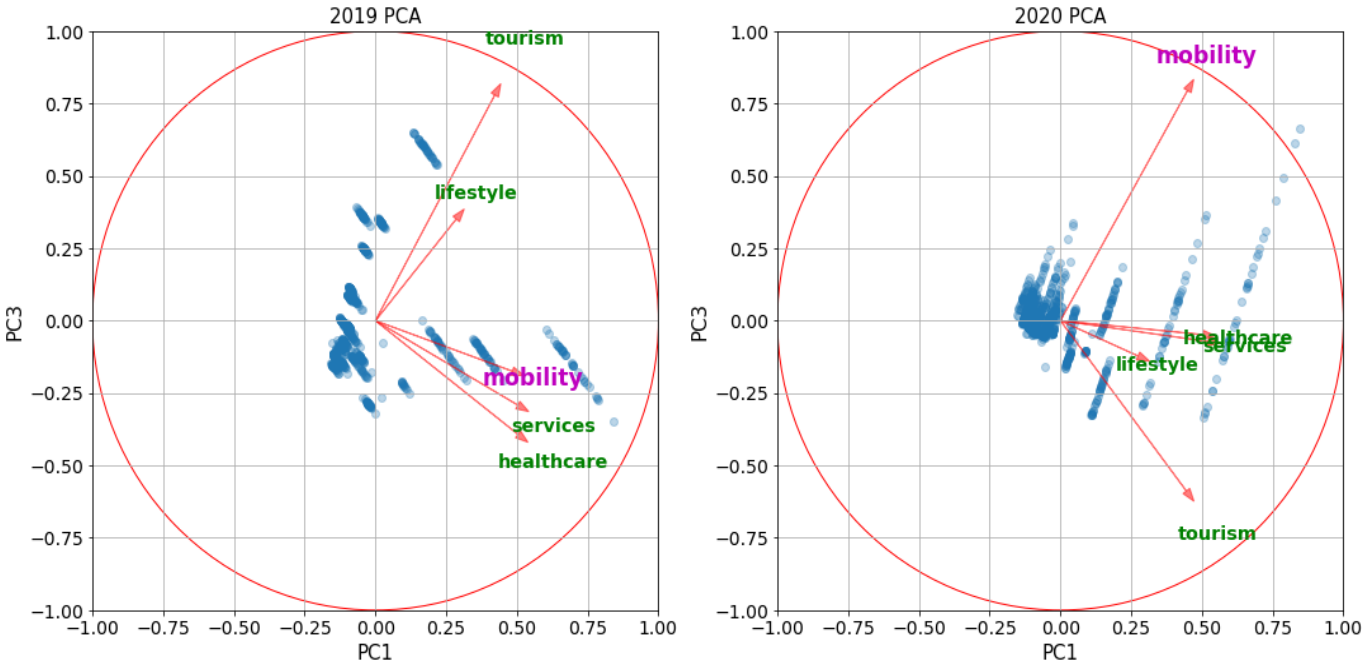

Fig. 43 Both PCA plots from 2019 (Left) and 2020 (Right)

Starting with the most evident differences, the two clusters that suffered more with the pandemic in terms of mobility are the *Healthcare* cluster which agglomerate the healthcare group, and the *Services* cluster which is composed by Financial, Government, Sustenance, Entertainment, Commercial and Community groups. Although there is still a positive correlation between these two clusters and the mobility, in 2020 this correlation is less strong than it is in 2019, this because, the angle that these two variables do with the mobility variable is much closer to 90 degrees. Meaning that these types of establishments suffered a huge loss in mobility from 2019 to 2020, these results can have multiple reasons being governments restrictions in the opening and closing schedule especially on the *services* cluster (i.e., closing time needs to be earlier) or even full closure of these establishments, on the other hand, the *healthcare* department could have also loss mobility for the fact that people stopped going to the Hospital unless they really needed with fear that they could caught Covid-19 on there or because, there was much less waiting line (restrictions of number of people inside the hospital, the schedule was tighten in order to prevent a big agglomeration of people) and people started to go alone to the appointments. Of course, these is only speculation, although with a little veracity, there could be multiple reasons for these results despite the pandemic, and it could be interesting to study these reasons in any future work.

Looking at the clusters that have suffered a bigger loss in the correlation with mobility, on the flip side there are those which correlation is not as evident, such as *tourism* and *lifestyle*, in which there are information to extract too. However, for these we need to know the exact angles that each variable does with the mobility vector, so we will use the Table 3 to compare all the angles.

Table 3 Angles between the Mobility variable and the other variables for 2019 and 2020

Variable	**Angle (2019) in degrees**	**Angle (2020) in degrees**
Lifestyle	70.70º	84.51º
Tourism	81.42º	113.43º
Services	10.26º	68.39º
Healthcare	18.09º	66.00º

By analysing the table above, especially the *Lifestyle* and the *Tourism* clusters we can extract some information on how the mobility on these locations was affected. Looking first at the *Lifestyle* cluster there is 70.70 degrees for 2019 and 84.51 degrees for 2020, meaning that in 2020 the *Lifestyle* cluster slightly lost its impact on mobility, because it got closer to 90 degrees, so in this year the mobility did not have a high correlation with locations mostly with Educational and Transportation establishments, this cluster might be the one that suffered less in terms of mobility with the pandemic because it is the one that is more redirect to the locals and not the tourists, which make a good percentage of the mobility. Secondly, for the *Tourism* cluster there was also a big change in the impact it had on the mobility, in 2019 although very low, it still had a positive correlation with the mobility (less than 90 degrees), meaning the tourism points increase with the increase of mobility. However, in 2020 this correlation became a negative one (more than 90 degrees), so the increasing number of points of tourism reflected in a lower mobility in that location. With these analyses it is possible to extract some of the impacts the pandemic had in the mobility, such as, locations where people used to agglomerate (Tourism locations) got shut down by government restrictions in 2020 resulting in less movement on those areas, transports and schools also were affected by restrictions, schools got closed and transports had to reduce the number of passengers.

We clearly observe the impacts of the pandemic on the daily basis of people and how they change their behaviour accordingly to the restrictions made by the government. We also can predict the impact of this decrease in mobility on small business across the island, since they all are in the *service* cluster, and this was one of the clusters that suffered more with the restrictions.

Overall, we determine the impact of COVID-19 and at the same time what drives people to move to a certain area, and if continued our analysis to 2021 and 2022 we could probably visualize the slowly increasing mobility on places as *services* and *tourism*.

Chapter 6

Conclusions and Future Work

In this section, we will summarize the approaches taken and review the research questions, the last point will be the future work where we discuss where our analysis could lead us, some questions that arose during this thesis and also what could be the next steps as well as other paths that could be more promising.

6.1 Conclusions

We analysed four different datasets for establishments, population, telecom and Wi-Fi, from April to September in 2019 and 2020 with the objective of understanding human mobility and the impacts of the pandemic as well as categorize Madeira Island according to different factor.

As we saw with this thesis, there are multiple factor that influence mobility, such as population, establishments availability, mobility government restrictions and much more. And were exactly these reasons that we wanted to analyse with our thesis as well as the relations between establishments and population.

To summarize our results, we will review the research questions made in section 1.2:

- RQ1: Does the population affect the type of establishments available on a certain location?

  As we saw in section 5.1, although not for all type of establishments, the population itself plays a huge role when deciding to open/build a new service on a certain area, and this information can be very helpful to policymakers and urban planners when choosing the location to build a new mall or bus station for example.

- RQ2 Does the types of establishments available influence the other establishments around them?

  We analysed this idea in section 5.2, and we can visualize that when inspecting a small area (Hexagon) there are some establishments that have a correlation higher than 0.8 with others, and normally these establishments with a higher correlation have some type of symbioses relation with each other, meaning that business owner might have in count the space around them before opening a business.

- RQ3: Can the Wi-Fi infrastructure (with the routers) translate the real mobility within a large area (an island in this case)?

  With the help of the telecom dataset as ground truth (section 5.3), we could prove that the Wi-Fi infrastructure can translate the real mobility within the island by analysing the shape of both lines (telecom and Wi-Fi) and the values higher than 0.5 in the Spearman correlation, this validation may lead to more complex analyses in the future.

- RQ4: Does the mobility have any relationship with the establishments?

  The main objective here was exactly to understand human mobility and the reason behind people movements. To accomplish that we did an analysis in section 5.4, that helped us understand the main services that people are looking for when they move from one place to another by giving us the type of establishments that had more relation with the mobility in 2019 as well as in 2020. Although we did this analysis also aiming to the impact of COVID-19 on mobility, the 2019 analysis is a “normal” year where we can extract insightful information about people movements to help governments and policymakers with their decisions for the improvement of public transports, land planning and crowd control for example.

Although we answered all these questions, there is still room for more analyses using these types of datasets since this line of study have an immense range of different types of analyses. Nevertheless there are always some limitation in these datasets/analyses that makes it more challenging or even impossible to make other analysis, for example, one of the drawbacks of using data compiled by the community (OSM), more specifically the establishments data, is that there might be some locations missing from the data which can impact the results, also, since we are talking about anonymous data (from the Wi-Fi data) we cannot predict the jumps from one place to another which could be really interesting in the mobility aspect.

6.2 Future Work

Regarding the future work, we already mentioned that these types of analyses have a large scope of approaches, in this section we will mention some follow up ideas that can be used using this thesis as a baseline as well as some new analyses that can be done.

Starting with the establishment’s distribution and their correlation there is room to do a deeper analysis on the distribution of these groups and try to understand if it is possible to improve the lifestyle of the people living in Madeira Island, especially by inspecting the living, educational and transportation categories which are the one that might have a bigger impact on people lives. For example, better distribution of the public transports station, new places to build schools in order to everyone have easy access to education, among other ideas. However, these can be applied to any category with a different goal, small business owners can use this data to understand what the best location is to open a new establishment, governments can also use this data to better manage Island, etc.

There is also the possibility to use our results and do an all-different analysis on new locations to install new routers, as we saw along this thesis, some of the drawbacks that we encounter were done by the lower number of routers, although there are districts where we have good coverage there is always room to improvements. And this coverage growth will open new windows on new analyses with much more accuracy and much more complexity.

Regarding the correlation between mobility and establishments, instead of doing an analysis on the impact that the pandemic had in people lives, there is the possibility to understand how people move depending on the time of the day. In this thesis we did an analysis where we analysed 6 months, yet it can be done an all-new analysis per time of the day, for example, morning, afternoon, and night and understand the differences between them.

One interesting idea that could be explored is the real time analysis of the data from which we could predict in real time the location with more movement and the type of establishments that might be congested. Another approach could be the creation of an API with all this thesis data regarding establishments and make it public to the community or even mobile application so that it can use to complement the application data and give the users a visualization with all the information, this application could be a tourism application, transports application, etc...

This thesis leaves open doors to utilize these datasets in several areas of study, tourism, mobility, city planning, etc , being for visualization, exploration, new analysis or even expand the dataset with more metadata.

Annexe 1

Map analyses of Madeira Island

Here we display all the maps regarding our analysis of Madeira Island from:

- Hexagon Maps with the distribution of each group
- Parish Maps with the distribution of each group
- Maps with the locations of each group

A.1 Hexagon Maps with the distribution of each group

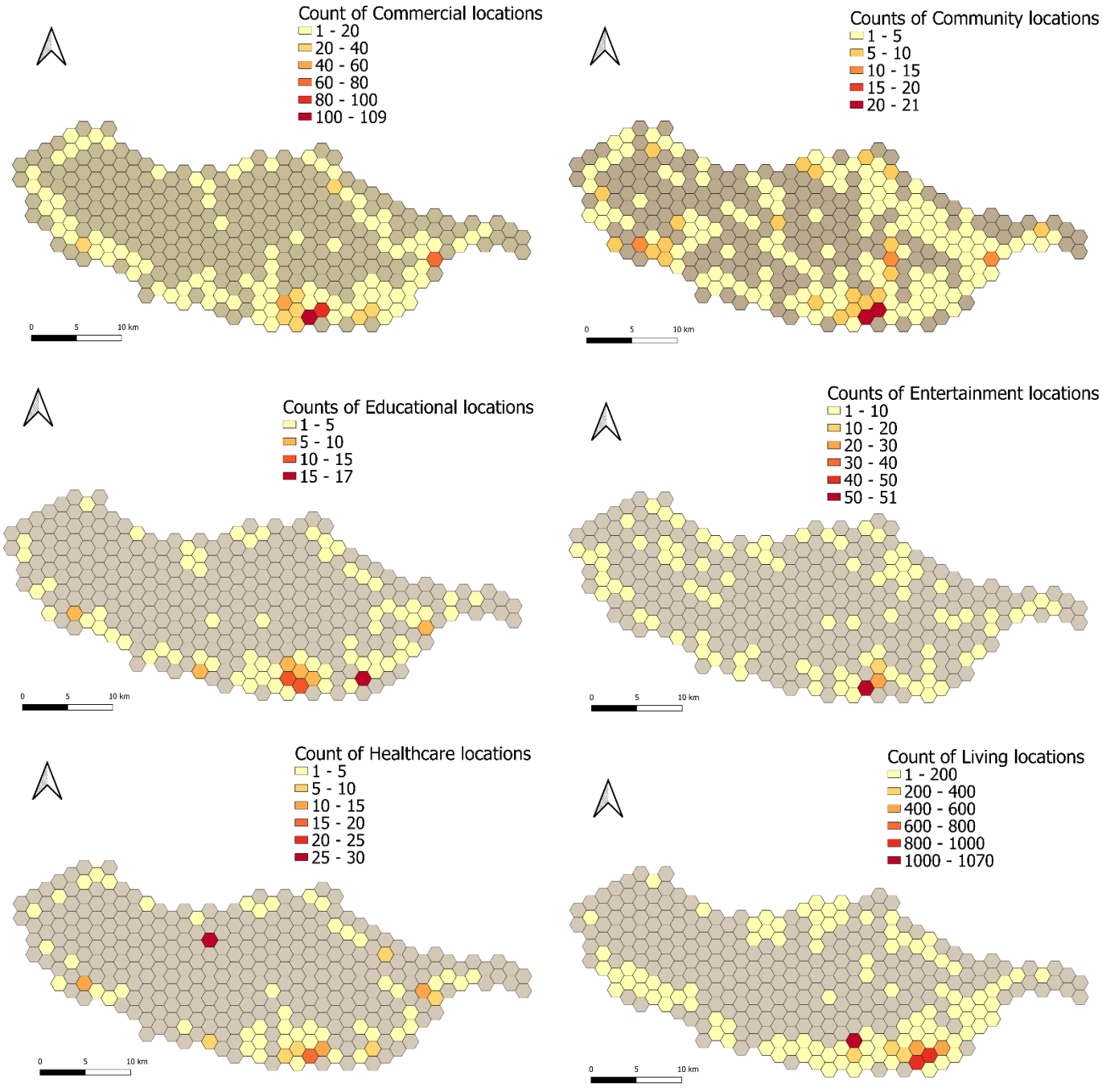

Fig. 44 All Hexagon maps for the POIs distribution (Part 1)

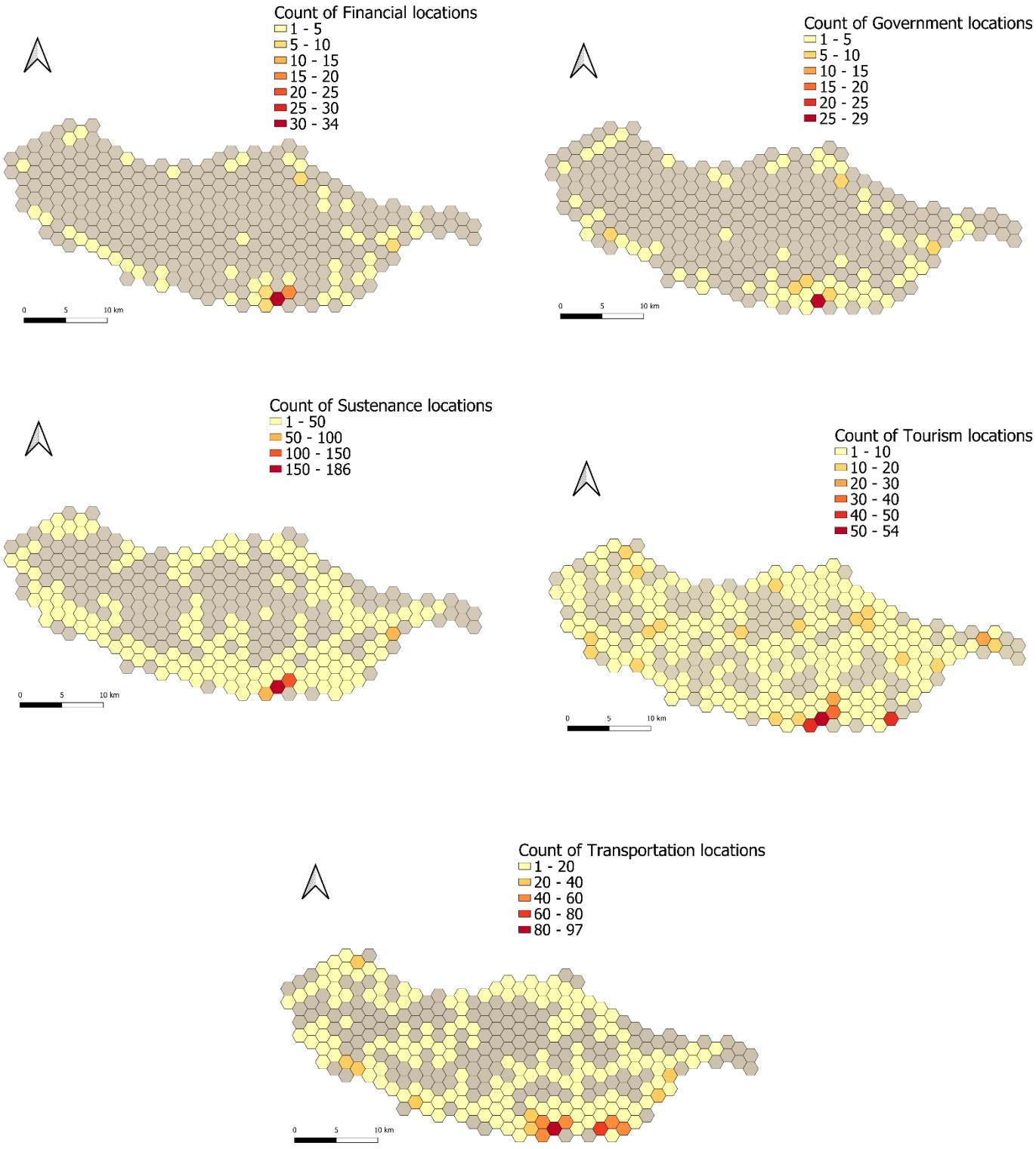

Fig. 45 All Hexagon Maps for the POIs distribution (Part 2)

A.2 Parish Maps with the distribution of each group

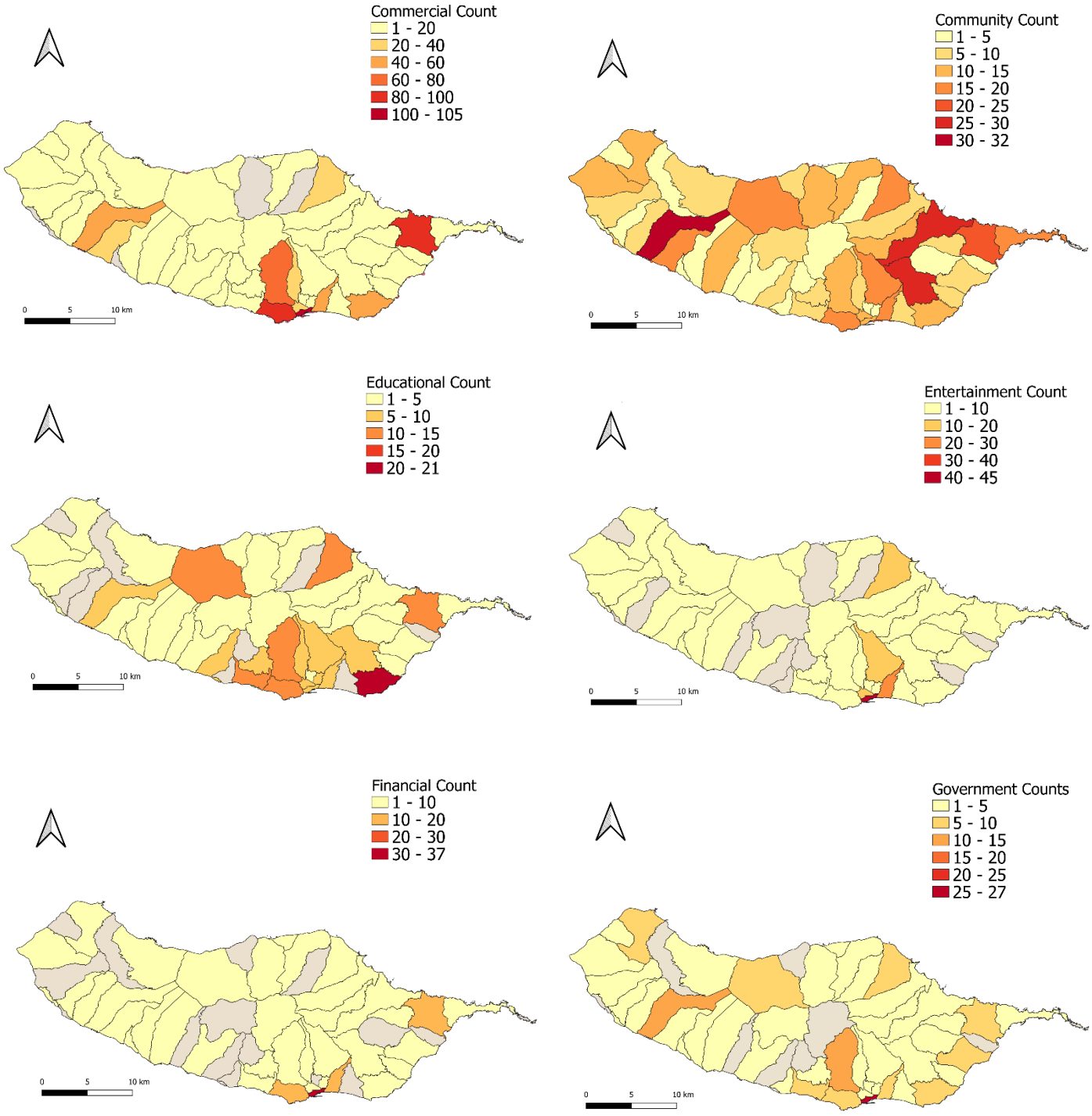

Fig. 46 All Parish Maps for the POIs distribution (Part 1)

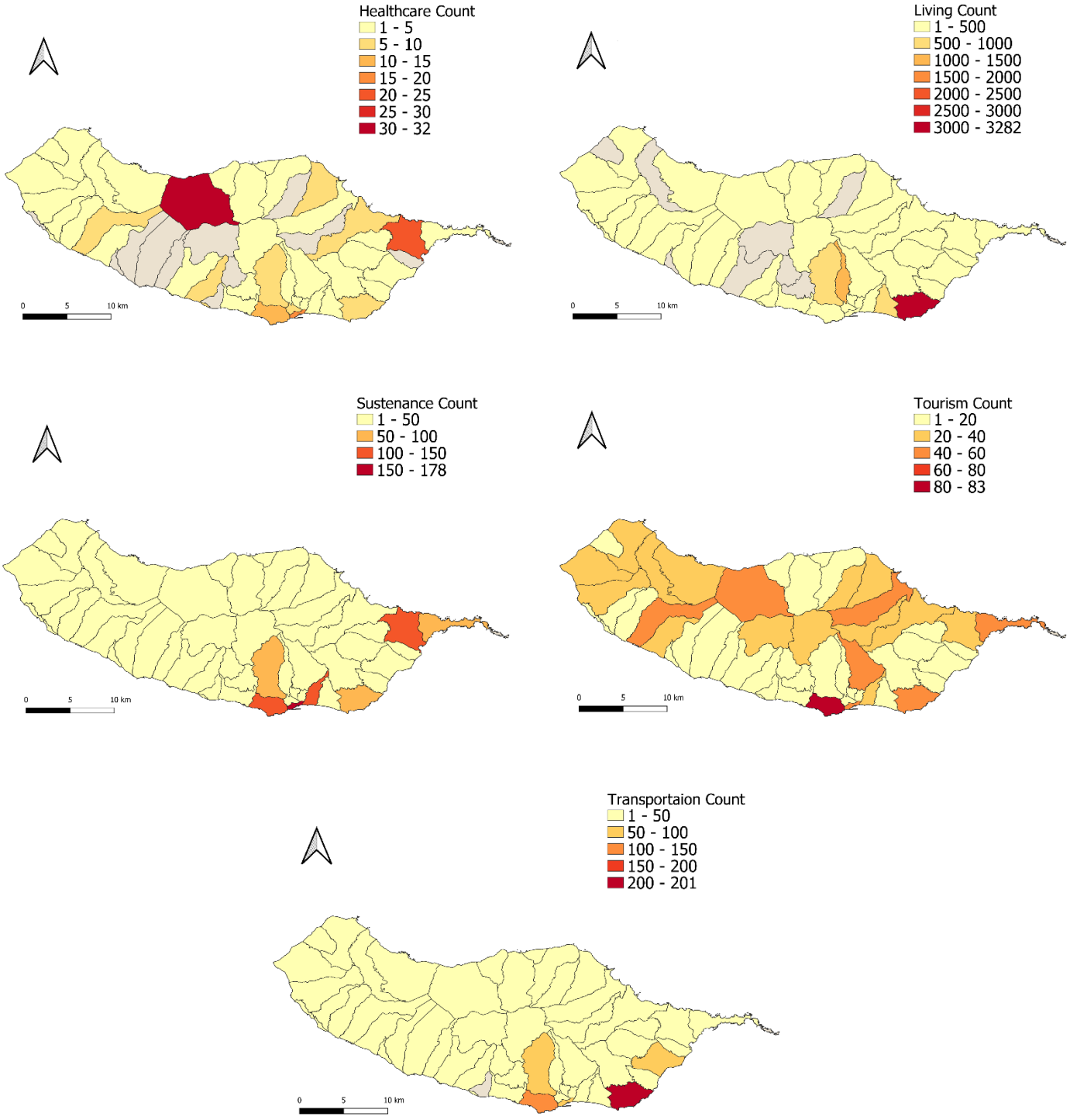

Fig. 47 All Parish Maps for the POIs distribution (Part 2)

A.3 Maps with the locations of each group

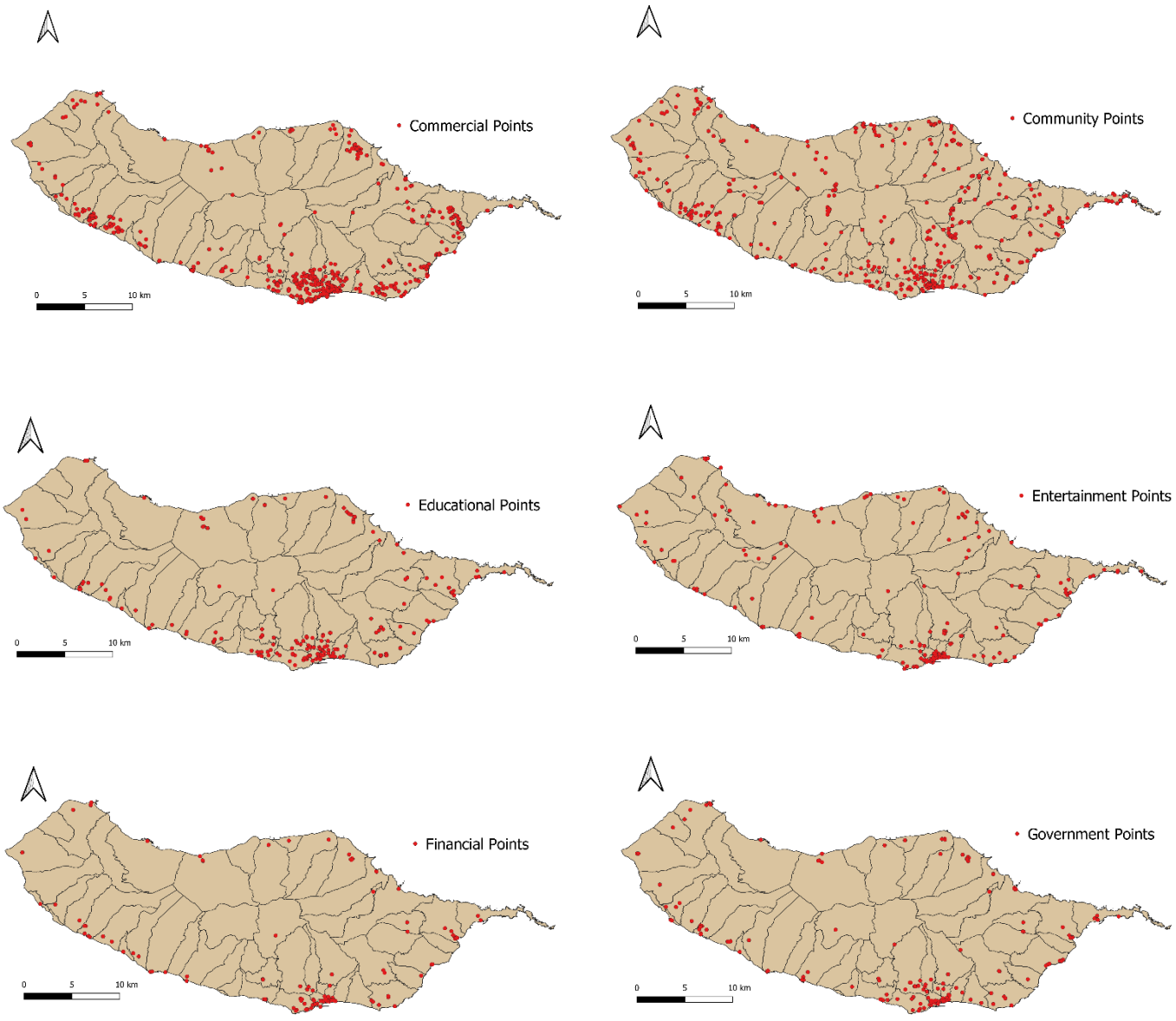

Fig. 48 All Maps with the points of each POI (Part 1)

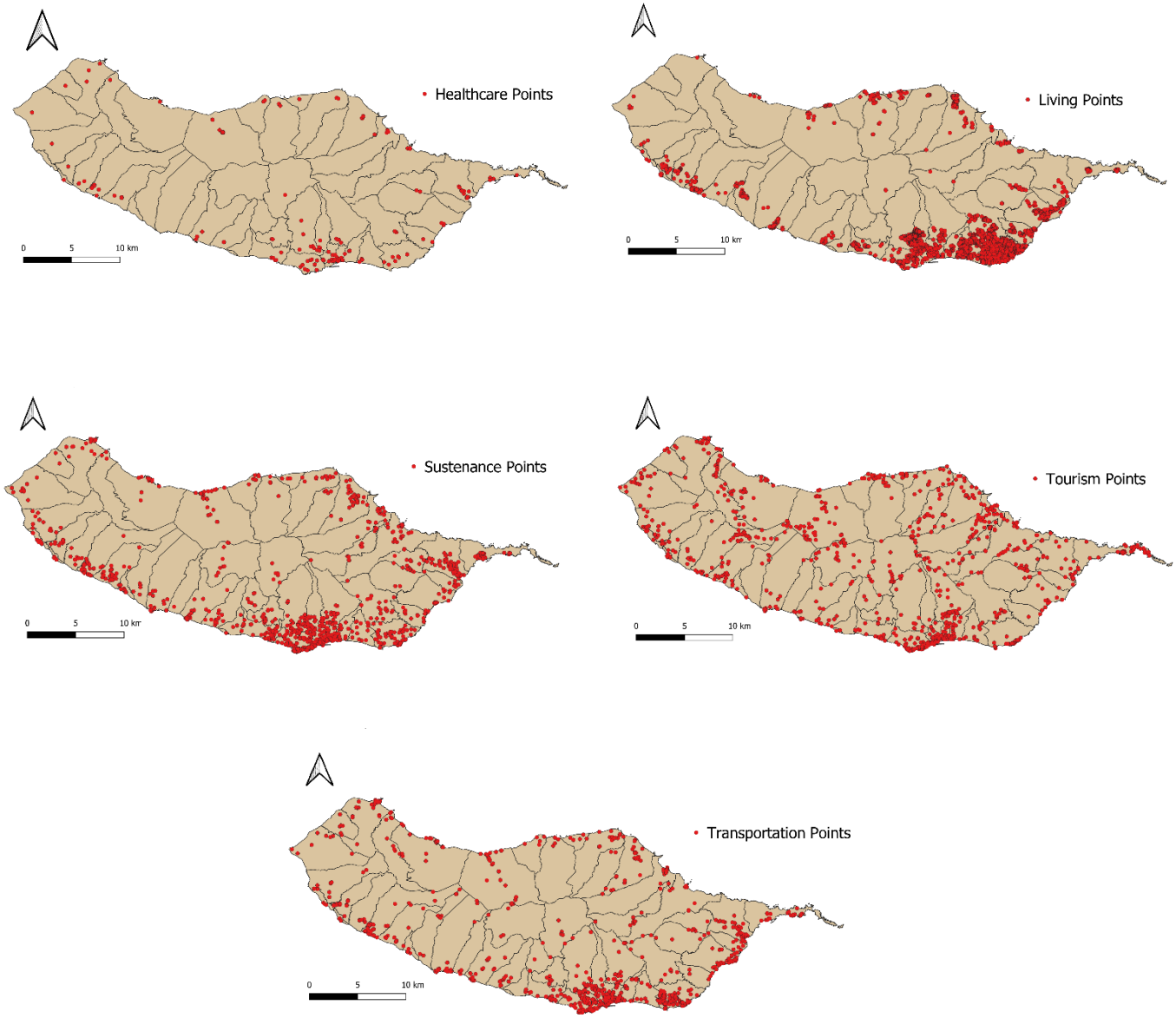

Fig. 49 All Maps with the points of each POI (Part 2)

Annexe 2

Closer POIs to each router

Nearest POIs to each router

A.4 Number of POIs close to each router

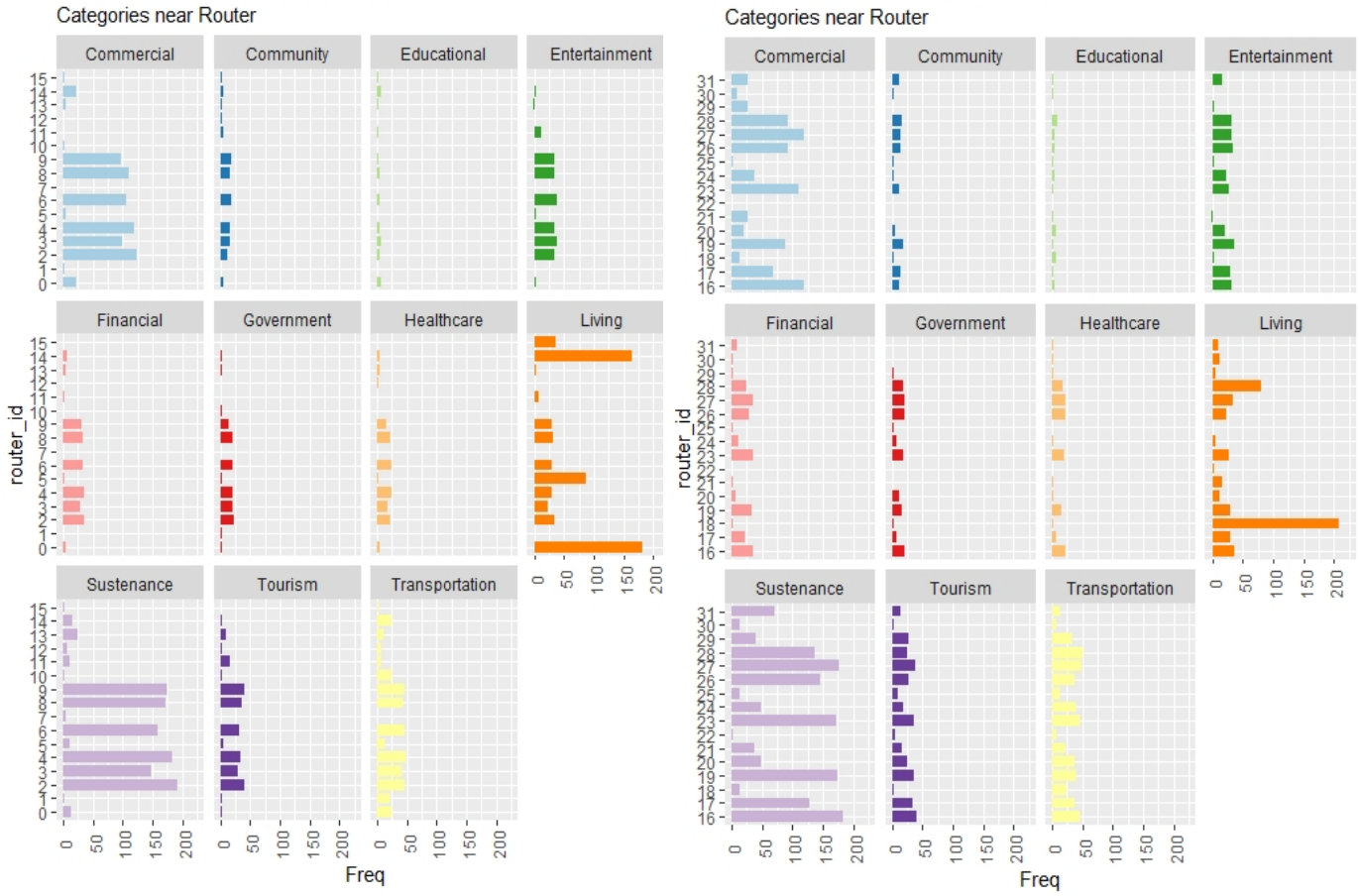

Fig. 50 Nearest POIs to all routers (Part 1)

Fig. 51 Nearest POIs to all routers (Part 2)

References

1. Agrawal, R., Imieliński, T., & Swami, A. (1993, June). Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).
2. Agung, M., & Kistijantoro, A. I. (2015, November). High performance cdr processing with mapreduce. In 2015 9th International Conference on Telecommunication Systems Services and Applications (TSSA) (pp. 1-6). IEEE.
3. Barrat, A., Cattuto, C., Tozzi, A. E., Vanhems, P., & Voirin, N. (2014). Measuring contact patterns with wearable sensors: methods, data characteristics and applications to data-driven simulations of infectious diseases. Clinical Microbiology and Infection, 20(1), 10-16.
4. Bonné, B., Barzan, A., Quax, P., & Lamotte, W. (2013, June). WiFiPi: Involuntary tracking of visitors at mass events. In 2013 IEEE 14th International Symposium on" A World of Wireless, Mobile and Multimedia Networks"(WoWMoM) (pp. 1-6). IEEE.
5. Chen, N. C., Xie, J., Tinn, P., Alonso, L., Nagakura, T., & Larson, K. (2017). Data Mining Tourism Patterns-Call Detail Records as Complementary Tools for Urban Decision Making.
6. Galí, N., & Donaire, J. A. (2010). Direct Observation as a methodology for effectively defining tourist behaviour. E-Review of Tourism Reseach.
7. Hartmann, R. (1988). Combining Field Methods in Tourism Research. Annals of Tourism Research 15 (1):88-105.
8. Herrera-Quintero, L. F., Vega-Alfonso, J. C., Banse, K. B. A., & Zambrano, E. C. (2018). Smart its sensor for the transportation planning based on iot approaches using serverless and microservices architecture. IEEE Intelligent Transportation Systems Magazine, 10(2), 17-27.
9. Keul A. & Küheberger, A. (1997). Tracking the Salzburg Tourist. Annals of Tourism Research 24(4): 1008-1012.
10. McKercher, B. (1999). A chaos approach to tourism. Tourism management, 20(4), 425-434.
11. Nunes, N., Ribeiro, M., Prandi, C., & Nisi, V. (2017, June). Beanstalk: a community based passive wi-fi tracking system for analysing tourism dynamics. In Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems (pp. 93-98).
12. Ribeiro, J. M. S. (2016). Human mobility tracking through passive wi-fi: a case study of Madeira Island (Doctoral dissertation).

13. Ribeiro, M., Nunes, N., Nisi, V. et al. Passive Wi-Fi monitoring in the wild: a long-term study across multiple location typologies. Pers Ubiquit Comput (2020).

14. Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science, 30(5), 881-906.

15. Vanhoef, M., Matte, C., Cunche, M., Cardoso, L. S., & Piessens, F. (2016, May). Why MAC address randomization is not enough: An analysis of Wi-Fi network discovery mechanisms. In Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security (pp. 413-424).

16. Versichele, M., De Groote, L., Bouuaert, M. C., Neutens, T., Moerman, I., & Van de Weghe, N. (2014). Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Belgium. Tourism Management, 44, 67-81.

17. Zheng, Y., Zhang, L., Xie, X., & Ma, W. Y. (2009, April). Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web (pp. 791-800).

18. Siła-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., & Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. International Journal of Geographical Information Science, 30(5), 881-906

19. Pappala, K. (2020). Investigating the Role of Points of Interest in Estimating Mobility Patterns in Cities: An extended Gravity model-London Rail

20. Srinivasan, S. (2000). Linking land use and transportation: measuring the impact of neighborhood-scale spatial patterns on travel behavior (Doctoral dissertation, Massachusetts Institute of Technology)

21. Macedo, M. A. (2019). Análise da evolução da rede rodoviária e das acessibilidades na Ilha da Madeira (Doctoral dissertation)

22. Vida Maliene, Vytautas Grigonis, Vytautas Palevičius, and Sam Griffiths. Geographic information system: Old principles with new capabilities, 3 2011. ISSN 13575317. URL www.palgrave-journals.com/udi/.

23. Keith C. Clarke. Advances in Geographic Information Systems. Computers, Environment and Urban Systems, 10(3-4):175–184, 1 1986. ISSN 01989715. doi: 10.1016/0198-9715(86)90006-2.

24. Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

25. Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

26. canva.com/design/DAEVSJ1kcno/share/preview?token=X9PB2m2B2MrY4iyRZr-7Ug&role=EDITOR&utm_content=DAEVSJ1kcno&utm_campaign=designshare&utm_medium=lin

k&utm_source=sharebutton

27. http://osgeo-org.1560.x6.nabble.com/CAOP-2019-em-Geopackage-td5432158.html
28. https://wiki.openstreetmap.org/wiki/Osmosis
29. https://wiki.openstreetmap.org/wiki/Osmconvert
30. https://download.geofabrik.de/europe/portugal.html
31. http://wiki.openstreetmap.pt/images/upload/madeira.poly
32. https://wiki.openstreetmap.org/wiki/Map_features
33. https://wiki.openstreetmap.org/wiki/Key:amenity
34. https://wiki.openstreetmap.org/wiki/Key:building
35. https://wiki.openstreetmap.org/wiki/Key:healthcare
36. https://wiki.openstreetmap.org/wiki/Key%3Atourism
37. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html
38. https://www.statstutor.ac.uk/resources/uploaded/spearmans.pdf
39. https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
40. https://en.wikipedia.org/wiki/QGIS
41. https://en.wikipedia.org/wiki/Router_(computing)
42. https://estatistica.madeira.gov.pt/en/download-now-3/social-gb/popcondsoc-gb/popcondsoc-censos-gb/popcondsoc-censos-publicacoes-gb/category/35-censos-publicacoes.html
43. https://en.wikipedia.org/wiki/Point_of_interest
44. https://www.ipma.pt/pt/media/noticias/news.detail.jsp?f=/pt/media/noticias/arquivo/2016/madeira-5-10-ago-2016.html
45. https://cran.r-project.org/web/packages/geosphere/index.html
46. https://en.wikipedia.org/wiki/Interquartile_range
47. https://en.wikipedia.org/wiki/Hex_map