Mobility Analysis & Tourism in Madeira Island

Pedro Barbosa
Lisbon, Portugal
pedro.a.barbosa@tecnico.ulisbo
a.pt

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. 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

Author Keywords

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

INTRODUCTION

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, so that 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 [40]. 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.

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 (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?

RELATED WORK

This section presents literature related to movement patterns and distinct ways to gather data about mobility using several technologies.

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.

Call Detail Record

We will look over a model based on Call Detail Record (CDR) to 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, 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.

Bluetooth

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 tourists 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 tourists' patterns and behaviour.

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;

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.

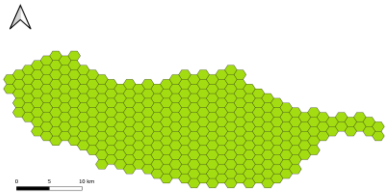

Fig. 1 Hexagon Grid

Taking this initial idea of analysing the city itself and the distribution of the POIs that we want to explore in this thesis, and by adding our Wi-Fi dataset, this combination will help us to extract important information about human mobility and examen the impact of COVID-19 on the island.

EXPLORATORY DATA ANALYSIS

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].

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.

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.

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].

Madeira shapefiles 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.

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 [39]. So, with the help of this Software we created a hex-grid as in Fig. 1 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".

OpenStreetMap Data (OSM) for POIs

A Point of Interest (POI) is a specific location that someone may find useful or interesting [42], such as hospitals, banks, shops, restaurants, etc. And it is with these points that we can categorize the type of location we are exploring.

Population

To get the population per parish in Madeira Island we used 2011 CENSOS [41], that were compiled by Instituto Nacional de Estatistica.

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, 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.

Telecom 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 5, they used these cell phone towers 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.

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 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.

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.

Once we got all the locations in OpenStreetMap, we ended up with a dataset, 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].

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.

With this new variable *group*, we finalize our dataset (longitude, latitude, type, sub_type, group) with all the POIs available for Madeira Island. 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.

Wi-Fi Data Processing

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 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 [43]. And 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, 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).

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.

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.

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.

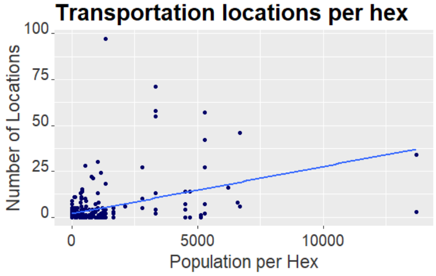

Fig. 2 Linear Regression Transportation-Population

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 scatterplot there is 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 this assumption with our data, as we can also suppose, with real world examples, why this category have a higher relation with the population, if we think for the Transportation category it is only logical to assume that the more population there is in one area the more of this category should exist on that exact are

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 Fig.1 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.

Looking at the heatmap created in Fig.3, 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

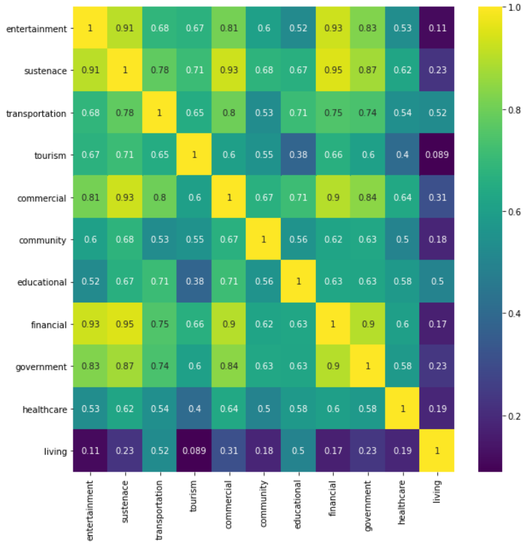

Fig. 3 Heatmap of correlation between Groups

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.

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 as ground truth (telecom 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, 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.

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, 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.

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. 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 already done in Affinities Between Groups we can use these results to combine our variables. Looking at Fig.3 and our analysis we can do 4 different clusters, 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

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 the 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 [38] 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.

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 in Fig.4, side by side, for an easier interpretation.

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.

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.

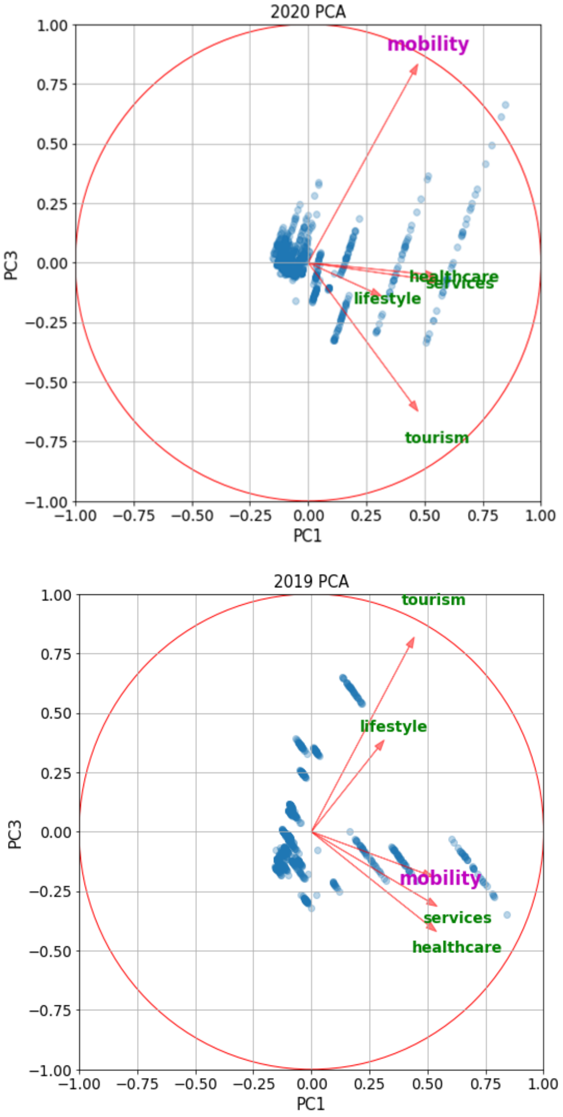

Fig. 4 PCA for 2019 and 2020

Conclusions

We analysed 4 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 Objectives and Research Questions:

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

As we saw in Relation Between Population and Groups, 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 Affinities Between Groups, 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, 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 0, 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

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.

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=link&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://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
39. https://en.wikipedia.org/wiki/QGIS
40. https://en.wikipedia.org/wiki/Router_(computing)
41. 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
42. https://en.wikipedia.org/wiki/Point_of_interest
43. https://en.wikipedia.org/wiki/Interquartile_range