Treball de Fi de Màster

Màster Universitari en Enginyeria Industrial

Mobility and COVID-19 spread pattern recognition in Barcelona area through Machine Learning techniques

MEMÒRIA

Autor: Gerard Franco i Panadés
Director: Immaculada Ribas Vila

Convocatòria: Setembre 2020

ETSEIB

Escola Tècnica Superior
d'Enginyeria Industrial de Barcelona

UPC

ABSTRACT

The past year 2020 was marked by the COVID-19 pandemic which had a great impact on many aspects of society, being mobility one of the most affected. In this context, this master's thesis was born with the aim of finding patterns of mobility and contagion of COVID-19 in the first crown of the metropolitan area of Barcelona. These patterns are useful to characterize the areas of the territory and to understand the temporal evolution of mobility and spread of COVID-19.

The motivation for this master's thesis arises from the opportunity presented by the institution of the Metropolitan Area of Barcelona to make use of cell phone big data provided by Orange to the Ministry of Transport, Mobility and Urban Agenda (MITMA). It is highly interesting to be able to carry out a study around these data complemented with other data sources, since they have not been exploited and the public administration has significant interest in them.

To identify these patterns, a search of academic documentation has been carried out, and a Python software has been developed to process all the data, to compute correlations and to cluster the areas using machine learning techniques for further analysis. In this subsequent analysis, the correlations have been analysed and the causalities that justify them have been sought. Finally, the clusters have been used to find patterns visually by plotting the characteristics of the zones. It is interesting to mention that the developed software is used to process all the data from the MITMA database which is still being updated to the present date.

The desired results have been obtained since the characterization of the mobility and contagions of COVID-19 by clusters of zones has been achieved. Interesting conclusions have been drawn, such as that women get more infected with COVID-19, areas with a higher percentage of population aged 0-15 years have a lower volume of mobility and COVID-19 cases, or that areas with more health, education and social services facilities have a higher volume of mobility and COVID-19 cases.

Summary

Index of figures

Index of tables

Glossary

Glossary of acronyms

ABS	Basic Health Area
AMB	Metropolitan Area of Barcelona
CARNET	Cooperative Automotive Research Network
CCAA	Autonomous Communities
COVID-19	Infectious disease caused by SARS-CoV-2
ICU	Intensive Care Unit
MITMA	Ministry of Transport, Mobility and Urban Agenda
ODM	Origin-Destination Matrix
PCA	Principal Component Analysis
WHO	World Heald Organization

Glossary of technical concepts

Artificially dichotomized variable	Variable which originally had more than two qualitative values and has been split into as many binary variables as qualitative values can take.
Causality	Relationship established between an event and its influence on the cause of another event.
Clustering	Unsupervised learning technique which divides samples into groups with no pre-defined categories/classes are available.
Correlation	In statistics, it refers to the relationship between two variables.
Granularity	Size of the areas that form a grid when studying mobility flows.
Machine Learning	Study of computer algorithms that automatically adjust their performance from exposure to information encoded in data
Orthodromic distance	The shortest distance between two points on the surface

	of a sphere.
Outlier	Statistically typical values.
Unsupervised Learning	Algorithms which learn from a training set of unlabelled examples, using the features of the inputs to categorize inputs together according to some statistical criteria.

1. Introduction

The year 2020 was marked by the appearance of the virus SARS-CoV-2 which causes the coronavirus disease 2019 (COVID-19), the respiratory illness responsible for the COVID-19 pandemic. To mitigate the damage that the pandemic was wreaking on people's health and sanitary systems, most governments around the world took steps to halt its advance.

Coronavirus disease 2019 (Covid-19), technically referred to as SARS-CoV-2, is an infectious disease that was first reported in Wuhan, the capital of the Hubei province of China, on 31 December 2019. The World Health Organization (WHO) declared the 2019-20 coronavirus outbreaks a Public Health Emergency of International Concern on 30 January 2020 and a pandemic on 11 March [1].

The virus spreads mainly during close contact and when those infected cough, sneeze or chat, through small droplets. During breathing, these small droplets can also be formed. During the first 4 to 6 days after the onset of symptoms, the virus is most contagious, while spread is possible in asymptomatic conditions and in later stages of the disease. Social distancing, mobility limitations, pro-active testing and isolation of detected cases are the recommended steps to contain the pandemic [2].

The measures taken by governments worldwide to curb its spread have led to partial and/or total confinement, restrictions on mobility, limitations on economic activity and social distancing in most populations around the world. In Spain, these restrictions led to the declaration of a state of alarm on 14 March taking to a house confinement for a large part of the population. This was maintained for more than a month before giving way to the so-called de-escalation phase. This was done asymmetrically depending on the needs of each autonomous community. At the end of the state of alarm on 21 June, the whole territory automatically returned to the “new normality”. During this period, the state restrictions enacted through the state of alarm were not applied. Nonetheless, this step did not mean the cessation of restrictions, as there were new measures at both state and autonomous community level. Since then, restrictions have been applied according to the number of infected people in each area and a night-time curfew has been in place for several months.

Territories with previous experience on pandemics are the ones who have had the best response to the situation [3]; and it is clear that mobility is directly linked to the spread of the SARS-CoV-2 (observing the actions of the states).

For this reason, the motivation for this thesis arises from the synergies between CARNET (a future mobility research hub) and the institution of the Metropolitan Area of Barcelona (AMB). Both entities were involved in a European project related to data analysis to promote

sustainable urban mobility in which the author of this master's thesis participated as a CARNET intern. When the AMB knew that the author wanted to develop the master's thesis related to the analysis of mobility data, they let CARNET know that they had a large source of mobile telephony mobility data (considered Big Data for its size) from Ministry of Transport, Mobility and Urban Agenda (MITMA) originated from the Orange telecommunications company, and that currently no resources were allocated to analyse it. On the other hand, CARNET is interested in generating knowledge about the mobility of the territory and hence this master's thesis was conceived. Its motivation is to identify mobility and COVID-19 spread patterns during the year (with different mobility restriction periods) across the different municipalities and districts from the territory of the first crown of the metropolitan area of Barcelona. In addition, the aim is to analyse those patterns to understand how the introduction of mobility restrictions to curb the pandemic really affects the different zones that conform the study area, taking as a source of data a huge population sample that represents 27.41% of the total population, as can be seen in Figure 1 [4].

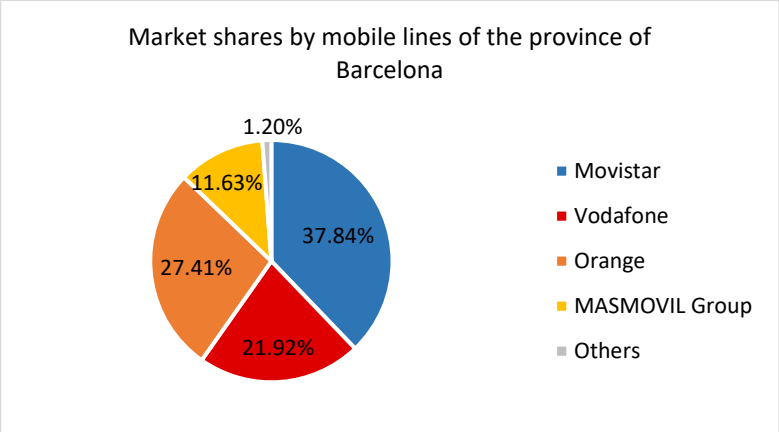

Figure 1. Representation of the market share of the telecommunication companies of the province of Barcelona from 2019 [4].

The study area and its zone segmentation are relatively unusual, as it covers neither the entire province nor the entire metropolitan area; it is the union of the municipalities of the first crown and the 10 districts of the city of Barcelona. Although it is a challenge to collect data in a unified way in these different zones with this level of detail (municipalities and districts), this area and segmentation has been taken because of its great interest, since it contains the largest volume of daily mobility in the whole Catalan territory. As mentioned, the mobility data source for the study comes from the telecommunications company Orange which gave its mobility data to the Ministry of Transport, Mobility and Urban Agenda (MITMA). They were first accessed through the AMB and then opened publicly for innovation projects. As it can be seen on Figure 1, Orange represents 27,41% of the market share of its sector [4], allowing to assume that the data is significant enough to draw conclusions.

The data on COVID-19 and on the territory have been obtained from different datasets from

the official open data sources of the Generalitat and Barcelona City Council [5].

1.1. Objectives of the project

The aim of this master's thesis is to recognise patterns of mobility and contagion of COVID-19 from 14 February to 30 November, relating them to their evolution through the different stages of mobility restrictions and the different zones that make up the study area. In order to achieve this main objective, the following specific objectives have been defined:

- Analysis of the data to characterise the mobility behaviour evolution through the territory.
- Analysis of mobility between Barcelona districts and the municipalities of the first crown to find significant mobility flows.
- Analysis of the data to characterise the evolution of the spread of the COVID-19 pandemic through the territory.
- Look for correlations and causalities by crossing all the existing data to find relationships between the different variables to analyse them later. Linking mobility and COVID-19 data among them and their characteristics and with demographic data and facilities of the territory.
- Group the different municipalities and districts that conform the study area into clusters using machine learning techniques to find out which municipalities and districts have similar behaviour in relation to mobility and COVID-19 propagation.
- In-depth analysis of the mobility and COVID-19 of the groupings made by data similarity using machine learning techniques to recognise marked patterns on its evolution through the mobility restriction periods.
- Verify non-obvious correlations in the study of clustered zones.

1.2. Scope of the project

This master's thesis analyses the mobility and COVID-19 infections to recognise patterns in 10 districts of the Barcelona city and the municipalities of the first crown of the metropolitan area of Barcelona. This first ring area has been divided into 44 zones, summing up the municipalities and districts of Barcelona.

The foundations of the study are the baseline data. Despite being collected from mobile phone data, the composition of the mobility dataset from the Ministry of Transport, Mobility and Urban Agenda (MITMA) are counts of total number of trips classified by zones and different characteristics. These data do not provide detail on human behaviour in compliance with the

European Data Protection Board's 'Guidelines on the use of location data and connection tracing tools related to the COVID-19 epidemic' (EDPB, 2020), but they do provide valuable insights into population movements. The data on COVID-19 and on the territory have been obtained from different datasets from the official open data sources of the Generalitat and Barcelona City Council [5].

2. Methodology

The following methodology, divided into six major steps, has been adopted to approach the study:

Existing literature

The first phase of the study has consisted of an extensive literature search in the four main fields covered by this master's thesis: mobility analysis, COVID-19 propagation analysis and its link to mobility, correlation and causality techniques, and machine learning techniques for clustering. The search has been divided into two main parts, starting with the first two fields, as they are essential for the initial treatment of the data. Subsequently, after a first analysis of the data, the existing techniques in the other two fields have been researched in order to use them as analysis tools.

For the search of academic articles, Google Scholar has been used, giving priority to the content supported by well-known institutions (e.g. Elsevier, MDPI, Cambdrige, Springer, IEEE Xplore) which could have be accessed thanks to the virtual resources tool of the Unversitat Politècnica de Catalunya (eBIB button). In addition to academic articles, for the literature search on machine learning techniques, the books from the basic bibliography of the elective subjects *Pattern Recognition and Machine Learning* [6] and *Scientific Python for Engineers* [7] that the author took as part of the optional subjects of the master's degree of this master's thesis have also been used.

Data collection and treatment

Starting later, but running in parallel, the treatment of mobility data from the MITMA data source has been carried out. COVID-19, demographic and facilities data of the territory have been sought from official sources that are useful for characterising the first crown of the metropolitan area of Barcelona, these data have also been further processed.

There are multiple software's that allow this processing to be carried out. Among all possibilities, Python has been chosen because of its current widespread use for data processing applications, its philosophy oriented towards the readability of the code, which allows a better understanding for those who wish to use the developed code, and because it has an opensource licence that makes it free and open to everyone.

First analysis of the data to understand it

Once the mobility and COVID-19 data have been processed, a study is carried out for each of the databases to gain a preliminary understanding of how the data sets varies. For this purpose, the temporal evolution has been analysed throughout the study period and the zones have been analysed by dividing the territory arbitrarily into four clusters.

Feature correlation and causality

All data collected in the second step have been processed and grouped according to their common features. From these groupings, correlations have been calculated between all the variables considered interesting to correlate using the appropriate techniques studied in the first step. Finally, the most significant correlations have been identified, and their causalities have been sought.

Machine learning techniques to cluster areas

The collected data have been processed to give it the necessary structure to apply the corresponding machine learning techniques. The appropriate techniques resulting from the first step study have been applied, and a Python code has been developed to group the different municipalities and districts that form the study area. The aim of this grouping is to ensure that the data from the areas that share the same group are as similar as possible.

Analysis and characterisation of zones by clusters and existence of correlations

Finally, an in-depth analysis of mobility, COVID-19, demographic and facilities data for each cluster area has been carried out. In this analysis, the different segmentations of the data have been observed, and their patterns have been characterised in the different areas of the territory and during the time period of the study according to the stages of mobility restriction. Furthermore, it has been verified whether the correlations found in step four can be observed for the different clusters.

Although the ultimate goal is to find correlations and their causalities, and to characterise the areas of the territory according to the recognised patterns, it is essential to go through all the previous phases by dedicating them the time needed. It is imperative to understand what tools are usually used to carry out this type of studies; to ensure that all existing reliable data sources are taken into consideration; to obtain a robust data treatment that avoids errors that may lead to erroneous conclusions later on; and to carry out a preliminary analysis of the data in order to understand whether the later results are consistent and to detect errors.

3. Mobility data analysis

The starting point of this section has been the review of the most used tools to represent and analyse mobility data. This review includes not only the representation of mobility data, but also its characteristics. Finally, the analysis at hand is presented with the comparison of mobility patterns during pre-confinement, confinement, and post-confinement periods.

3.1. Mobility analysis state of the art

Mobility patters are normally analysed by considering the origin and destination of trips. This information can be represented by means of an origin-destination matrix or an origin-destination flow map.

3.1.1. Origin-Destination matrix

In the origin-destination matrix (ODM) each cell represents the number of trips from an origin (row) to a destination (column) from geographical reference areas over a reference period. In general, as it is stated in [7] an ODM is structured as showing:

- Reference period (date and, eventually, time)
- Area of origin
- Area of destination
- Count of movements

To achieve the aforementioned parameters, the study area is defined, and a grid is created to define the source and destination areas. To refer to the size of the grid, the concept of granularity is used. Granularity refers to the area size study (i.e., neighbourhoods, municipalities, provinces, countries, etc). Depending on the one chosen, these origin and destination areas will group a greater or lesser number of journeys [8].

Once the zoning is done, as it can be seen in Figure 2, the mobility flow can be represented as a mobility graph which represents each origin-destination area as nodes, and the flow of paths between them as directional arrowheads.

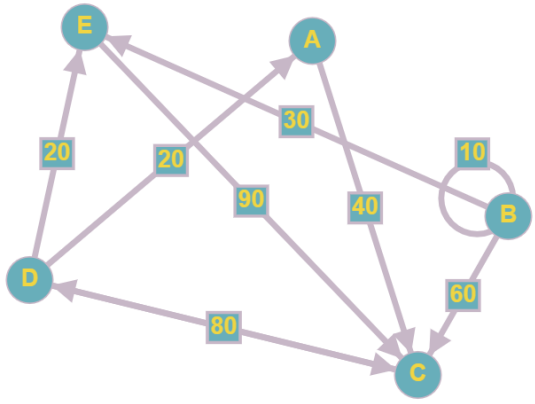

Figure 2. Example of a mobility graph. Source: Own

Figure 3 shows the origin-destination matrix, which contains the numbers that describe the mobility graph from Figure 2.

Nodes	A	B	C	D	E
A	0	0	40	0	0
B	0	10	60	0	30
C	0	0	0	30	0
D	20	0	80	0	20
E	0	0	90	0	0

Figure 3. Origin-Destination Matrix from the mobility graph example of the Figure 2.

To obtain more information from the ODM matrix, mobility indicators are used to carry out a more in-depth analysis. The mobility indicator used in this master's thesis, is the result of grouping all the trips by zones instead of origin-destination pairs, as in done in [8]. Instead of characterising the different origin-destination pairs, each zone and all its origins or destinations are reduced to only 3 categories: internal trips that do not leave the zone, trips that leave the zone (outwards) and trips that arrive in the zone from another zone (inwards).

Figure 4. Origin-Destination Matrix representing the mobility indicator by colours [8].

The ODM shown in Figure 4 illustrates the calculation of this mobility indicator. The rows of the matrix correspond to the origins, the columns to the destinations, and the content of each cell is the number of movements from a given origin to a given destination. A single grouping of greater granularity would be obtained, for example, taking k_1, k_2 and k_3 as 3 zones that belong to the same grouping K by adding up their rows and columns in a single one. The internal value of the indicator will be the sum of the ODM cells in orange (eq. 1). The inward indicator will be the sum of the cells in green: these are the movements that start outside K and end in K (eq. 2). Finally, the outward indicator will be the sum of the cells in blue: movements that start in K and end outside K (eq.3). Hence, the mobility indicator is the sum of orange, green and blue cells (eq. 4).

$$M_K^{int}(t) = \sum_{i \in K} \sum_{j \in K} ODM_{ij}(t) \quad \textit{(Eq. 1)}$$

$$M_K^{inw}(t) = \sum_{i \notin K} \sum_{j \in K} ODM_{ij}(t). \quad \textit{(Eq. 2)}$$

$$M_K^{outw}(t) = \sum_{i \in K} \sum_{j \notin K} ODM_{ij}(t) \quad \textit{(Eq. 3)}$$

$$M_K^{total}(t) = M_K^{int}(t) + M_K^{inw}(t) + M_K^{outw}(t) \quad \textit{(Eq. 4)}$$

Although part of the information is lost by eliminating the relationship between origin and destination zones, this indicator is useful to analyse each area of the territory individually while maintaining the characteristics of mobility and being able to associate them with other characteristics of the territory. The creation of this indicator does not detract from the fact that the origin-destination relationships can be analysed separately [7].

3.1.2. Origin-Destination flow maps

Using a static image, flow maps visualize movement and display not only which places have been influenced by movement, but also the direction and volume of motion. The graph drawing community has defined aesthetic requirements applicable to the development of node-link diagrams, although there are no empirical usage studies performed by cartographers on which to base flow map design decisions [9].

Study in graph drawing attempts to establish techniques for the two-dimensional representation of graphs. The difficulty of representing a graph - a set of nodes linked by edges - is related to the issue of mapping origin-destination flows for two reasons [9]:

- Origin-destination flows form a graph; the nodes are the starts and ends of flows and the connecting edges are the flows.
- The geometry of links does not need to be accurately described in both graph drawings and origin-destination flow maps but should be modified for a better readability.

Multitude of design choices have been studied, also known as aesthetic standards, which are applicable to cartographic origin-destination flow maps evaluating users of studies in graph drawing perception. However, only some of the graph drawing aesthetic requirements are applicable to flow maps since the location of geographic nodes in flow maps can usually not be changed, unlike node-link diagrams.

From a quantitative content analysis of 97 flow maps, the following design principles have been identified [9]:

- To show various amounts, almost every quantitative flow map differs in line width.
- Sometimes, flows are curved, but sharp bends are avoided in flow lines and symmetric curves are preferred.

- The number of flow crossings is minimized.
- For acute angles, wider intersection angles are preferred.
- Both maps showing the direction of flow use arrowheads.
- The path indicator tapered flow distance, which was recently recommended for graph drawing, is not used.

Scaled flow width best reflects quantity. Varying colour brightness may also display quantity, applying larger values to dark colours and smaller values to bright colours as seen in Figure 5. Varying luminosity reveals different quantities. Readability can also be improved by a small gradient of hue or saturation.

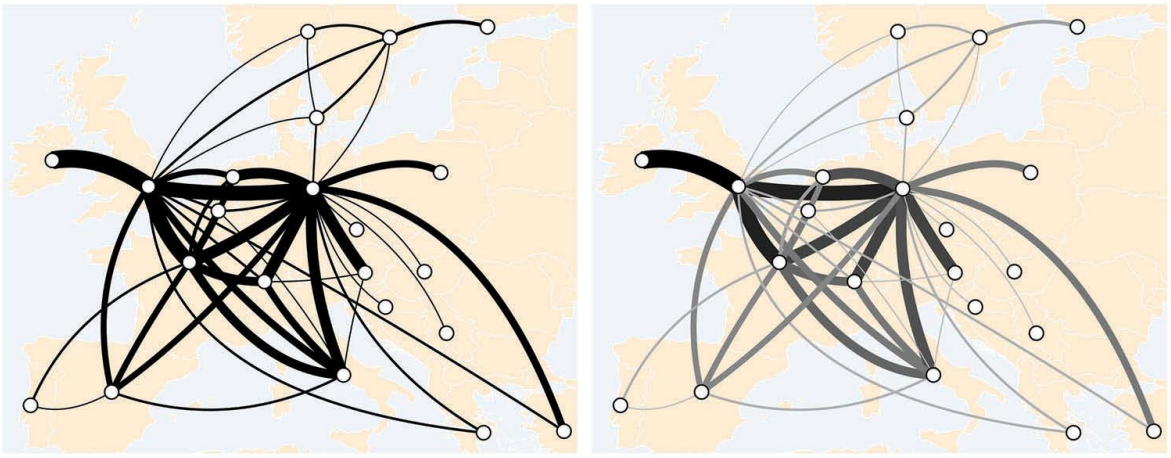

Figure 5. Quantities represented by width only (left) and width and brightness (right). Source: [9].

Direction is best indicated with arrowheads. The arrowhead size is calibrated for flow widths. To enhance readability, the size of smaller arrows is increased as the examples in Figure 6. Overlaps should be prevented between arrowheads and flows.

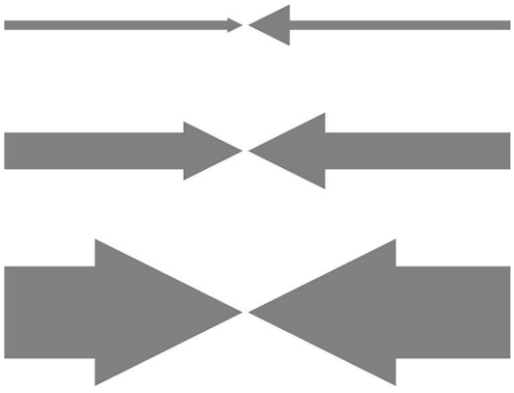

Figure 6. Size of arrowheads proportional to line width (left) and enlarged for thin flows to increase readability (right) [9].

3.2. Mobility dataset description

The Ministry of Transport, Mobility and Urban Agenda (MITMA) made available to the public, as open data and for free download, the data corresponding to daily mobility at national level.

This data has been obtained using mobile phone positioning information as the main source, complying in all its processing with Organic Law 3/2018, of 5 December, on Personal Data Protection and guarantee of digital rights. The Ministry, for its achievement, has the support of the technology company NOMMON, and with the data of the operator Orange Spain. The results offered are extrapolated to the total national population.

All journeys of more than 500 metres made by residents to and from Spain have been collected. For this purpose, the national territory has been divided by districts, in some cases grouping the districts into municipalities or aggregation of these (in order to always comply with current legislation on personal data protection). The number of journeys between and within these areas is available on a daily basis, differentiated by time periods and distances.

Information can be downloaded for the period from 14 February 2020 onwards and up to the present. It is available both on a daily and monthly basis. Two types of content are available for download, on the one hand, the daily journeys made between each source zone and each destination zone and, on the other hand, the number of journeys per person made in each zone. It also includes the zoning file used and the relationship file between this zoning and the municipalities, as well as the methodological document.

3.2.1. Data processing from the MITMA

The first sub-process consists of the extraction and pseudonymisation of the mobile phone records from 13 million mobile lines provided by a mobile operator in such a way that it is impossible to perform the process in reverse. The information generated and delivered to the Ministry is already aggregated, anonymised, and expanded to the population universe [10].

The generation of the mobility data characteristics has been carried out by the MITMA using specialised software developed for this purpose. This software has been used in more than 80 projects in different countries where anonymised mobile phone data have been used for the characterisation of urban and interurban mobility, both for public and private clients.

In relation to spatial resolution, location information is available at telephone cell level - coverage area of each antenna - which means a spatial precision of tens or hundreds of metres in cities and up to several kilometres in rural areas, which provides an idea of the error introduced in the determination of the position depending on the areas analysed. From these records, an origin-destination matrix with areas of a similar size to the zip codes is created.

However, they do not correspond exactly, it is a zoning of its own, as can be seen in Figure 7.

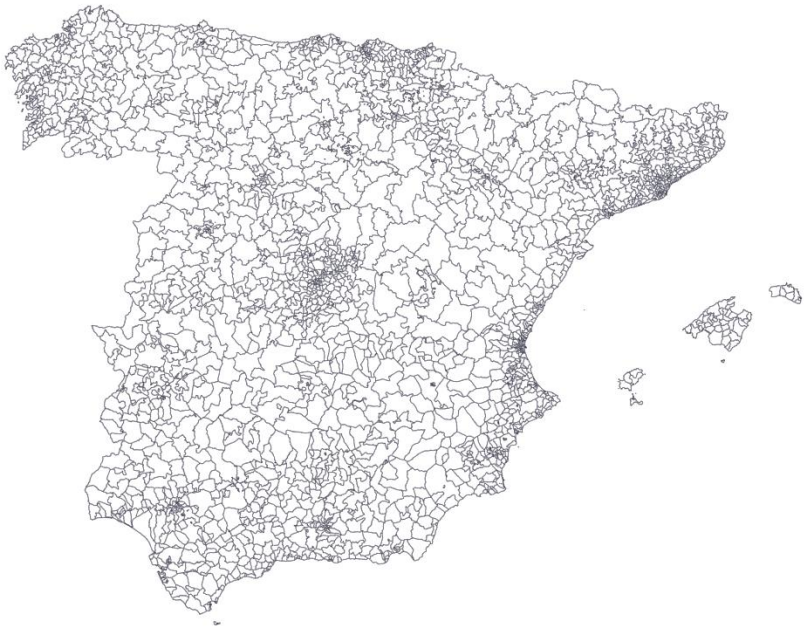

Figure 7. MITMA zoning for the entire Spanish territory.

The initial data obtained from the MITMA processing corresponds to a row-wise record of trip counts between an origin and a destination area according to a date and categorised as shown in Figure 9. This data between 14 February and 30 November corresponds to 89GB of characters in .csv format. The following origin-destination matrices segmentation is provided:

- Hourly sections according to the starting time of the trip.
- According to the orthodromic distance (as seen in Figure 8, the shortest distance between two points on the surface of a sphere) between origin and destination, distinguishing 6 distance ranges: 0.5-2 km, 2-5 km, 5-10 km, 10-50 km, 50-100 km and more than 100 km.
- For each element of the matrix, the total number of passengers-km corresponding to that origin-destination pair is provided, according to the orthodromic distance between origin and destination.

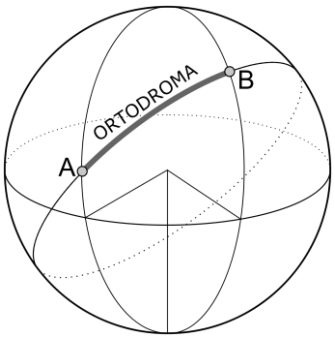

Figure 8. Orthodromic distance between two points along a great circle on the surface of a sphere.

fecha	origen	destino	actividad_origen	actividad_destino	residencia	periodo	distancia	viajes	viajes_km	valid
20200915	0801501	0801501	casa	otros	8	00	0005-002	6.637	5.479	True
20200915	0801501	0801501	casa	otros	8	01	0005-002	6.637	5.881	True
20200915	0801501	0801501	casa	otros	8	02	0005-002	13.274	11.004	True
20200915	0801501	0801501	casa	otros	8	04	0005-002	13.275	9.383	True
20200915	0801501	0801501	casa	otros	8	05	0005-002	19.911	12.496	True
...	...	...	...	...	...	...	...	...	...	...

Figure 9. Fragment example of the raw dataset downloaded from the MITMA open data source.

These origin and destination zones, which can be seen in Figure 9, comprise the geography of the whole territory of Spain zoned aggregated by administrative units (districts, municipalities or groupings of municipalities, depending on the case) with a population, in general, greater than 5000 inhabitants and in no case less than 1000 inhabitants.

It is worth bearing in mind that the segmentation of trips does not provide socio-demographic data on travellers. Therefore, the information contains solely trip characteristics, details on age or gender are not offered.

3.2.2. Limitations of the baseline data

Taking into account that the data used corresponds to Orange, it could be agreed that the sample of users consistent with one of the three main operators in each area of the territory and for each socio-demographic stratum approximates reasonably well to random sample of the resident population in that area.

However, there are intrinsic limitations associated with the technology, such as the absence of very young children who do not own a mobile phone, or a lower representation of the elderly, some of whom are also not mobile line users or whose records have been eliminated due to the poor quality of their mobility data [10].

For this reason and the complexity in detecting different trips due to the wide variety of ways in which the population travels, the following study will be approached from a comparative, time series and percentage perspective, rather than an analysis of absolute mobility numbers.

3.3. Mobility data analysis

Since this study covers the mobility analysis of the Barcelona city and the metropolitan area of the first crown, the data has been imported into a data structure with Pandas in a Python-based working environment. It has been decided to use Python as the software to process the data as it is freely distributed software. In this way, anyone interested can use the code attached in the Annex to process the data taken from open sources. Using the code in Annex**,**

the desired areas have been filtered to obtain the data. This is a very extensive code, as the design is very segmented in parts because a conventional commercial computer with 8GB of RAM does not have enough capacity to store all the data of the study area according to the MITMA granularity.

The zones corresponding to the first crown of the Barcelona metropolitan area according to MITMA zoning, consist of the 88 areas shown in Figure 10 [11].

Figure 10. MITMA zoning from the Barcelona territory with the zones of the first crown indicated.

By filtering 1916 million records of trips according to date, origin, destination and characteristics for the whole of Spain, only 3.1 million records correspond to the study area.
These zones have been grouped to achieve a granularity equivalent to the municipalities and the 10 districts of the city of Barcelona. The result are the following 44 zones:

- Badalona
- Badia del Vallés
- Barberà del Vallès
- Ciutat Vella (Barcelona)
- Sant Martí (Barcelona)
- Eixample (Barcelona)
- Sants Montjuic (Barcelona)
- Les Corts (Barcelona)
- Sarrià Sant Gervasi (Barcelona)
- Gracia (Barcelona)
- Castelldefels
- Cerdanyola del Vallès
- Cervelló
- Corberà de Llobregat
- El Papiol
- El Prat de Llobregat
- Esplugues de Llobregat
- Gavà
- L'Hospitalet de Llobregat
- Molins de Rei
- Sant Andreu de la Barca
- Sant Boi de Llobregat
- Sant Climent de Llobregat
- Sant Cugat del Vallès
- Sant Feliu de Llobregat
- Sant Joan Despí
- Sant Just Desvern
- Sant Vicenç dels Horts
- Santa Coloma de Cervelló
- Santa Coloma de Gramenet

- Horta Guinardó (Barcelona)
- Nou Barris (Barcelona)
- Sant Andreu (Barcelona)
- Begues
- Castellbisbal
- Montcada i Reixac
- Montgat
- Pallejà
- Ripollet
- Sant Adrià de Besòs
- Tiana
- Torrelles de Llobregat
- Viladecans

3.3.1. Time series evolution

By observing the temporal evolution of journeys in the first ring of the selected metropolitan area from Barcelona, it can be seen how the effect on mobility of the different weekdays (e.g., weekday or weekend) has a great influence, obstructing the analysis of its long-term behaviour.

For this reason, a 7-day moving average has been applied [8]. The value for 14 February is the average of the journeys from 14 to 20 February and so on for all other days. Using this statistical technique to soften the changes it is possible to observe how mobility evolves over time without the abrupt changes that weekends, public holidays or special occasions entail. External mobility are those journeys which have a different origin and destination, while internal mobility signifies that users don't leave the area in which their journey starts.

Broadly speaking, as it can be seen in Figure 11, it is worth mentioning that external mobility is much larger than internal mobility (an order of magnitude higher) and for obvious reasons it is the one that is affected by perimeter mobility restrictions. Therefore, it can be extracted that the behaviour of total mobility is clearly marked by the internal mobility as it much bigger than the external mobility.

Figure 11 shows that the evolution of the number of trips is very consistent with what would be expected, in absolute terms. From 14 March onwards, there was a very sharp decline due to the house arrest of a large part of the population, which increased with the tightening of the measures over the following two weeks. Thereafter, the number of journeys gradually increased with the softening of restrictions until the summer. However, it is worth mentioning that in summer, a considerable part of the population prefers to enjoy summer holidays in areas outside of the first ring, therefore that could be the reason why mobility was reduced during August. In September, mobility gradually increased again until the arrival of new mobility restrictions (weekend perimeter confinement and curfew).

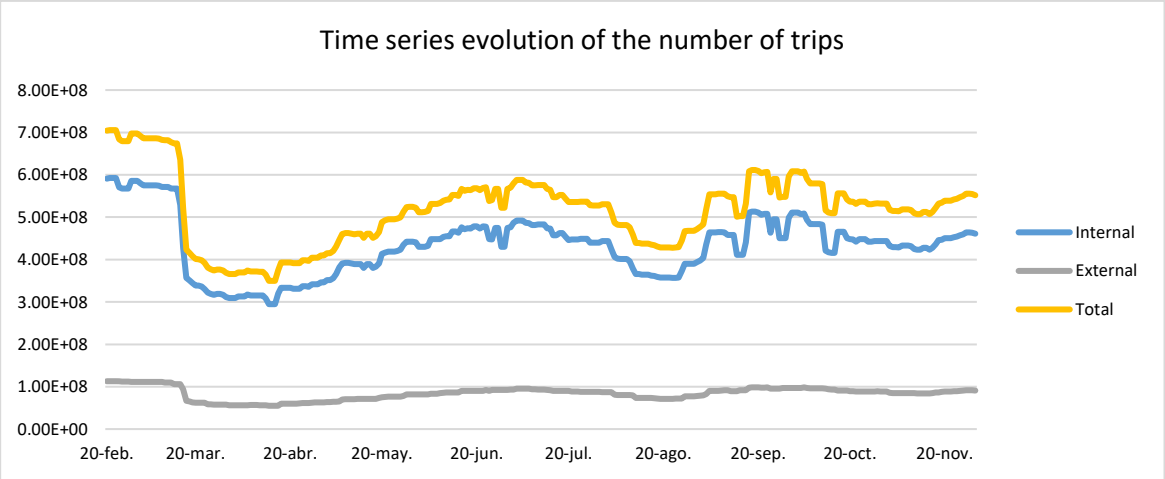

Figure 11. Evolution of the number of trips.

For the calculation of the trip reduction percentage, the following calculation shown in equation 5 is applied to the data in Figure 11 to obtain Figure 12:

$$\%\ trip\ reducction = \frac{7\ day\ moving\ average\ from\ 20\ February - 7\ day\ moving\ average\ from\ current\ day}{7\ day\ moving\ average\ from\ 20\ February}$$

(Eq. 5)

According to Figure 12 and taking as a reference the average number of journeys from 14 to 20 February, a substantial drop in mobility, quantifying in relative terms, of 50% can be identified in the first week of confinement and of 60% by the end of March, with the tightening of the measures. In July, mobility values corresponding to 20% below the reference week were regained, and in August they fell again to 40% compared to February. After summer, mobility increased to only 15% lower than in February, but with the introduction of new restrictive measures it fell back to 30%.

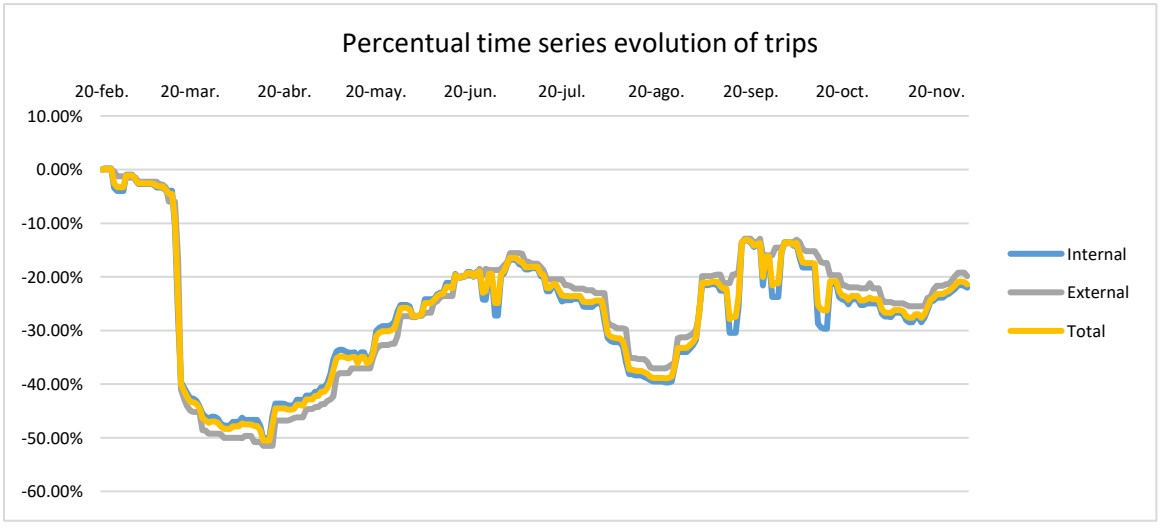

Figure 12. Evolution of the percentage of trips.

3.3.2. Evolution of trips by time periods

To evaluate different time periods, 30 days has been taken as the reference duration for the stages. Periods of this duration have been taken because no more pre-pandemic data are available. Then it has been considered which time periods have had mobility restriction measures that are more representative of the characterisation intended. Looking at these measures, it has been considered appropriate to divide into pre-pandemic, lockdown, and post-lockdown. Therefore, time periods with intermediate measures (deconfinement by phases, weekend perimeter restrictions, curfews, etc.) and times of the year with singularities (summer) have been avoided.

Bearing in mind the previous considerations, the following divisions have been taken for the temporal stages:

- Pre-pandemic: 14 of February to 14 of March
- Lockdown: 21 of March to 19 of April
- Post-lockdown: 14 of September to 14 of October

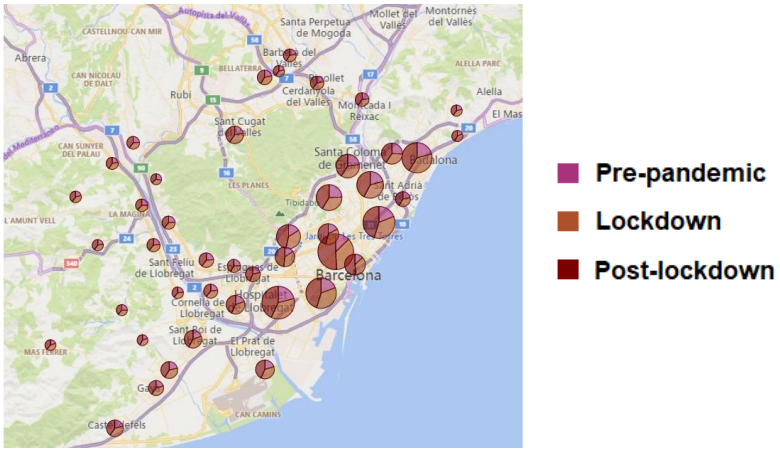

Figure 13. Proportion of trips by zones of the first crown of Barcelona during 30 days before the pandemic, during lockdown and post-lockdown.

Looking at Figure 13, areas with a larger graph have a much higher mobility of the area than the rest. In relation to the change in mobility by zones during the three established periods, all the zones have a very similar behaviour. Considering the total sum of the trips of the three periods as a reference, most of the zones comply with the following range of percentage distribution of the periods: pre-pandemic (39%-45%), lockdown (18%-23%) and post-lockdown (34%-39%). The areas to highlight as examples of values that fall outside the above ranges because the reduction in mobility has been more extreme are the districts of Barcelona: Ciutat Vella, Eixample, Les Corts and Sarrià-Sant Gervasi. In all of them, mobility was especially reduced during confinement. Figure 14 shows that the Barcelona districts of Sant Martí, Eixample, Sants-Montjuic, Horta-Guinardó, Sarrià-Sant Gervasi and Sant Andreu account for

51% of the total trips in the first ring, only 8 zones. In addition, Figure 14 also displays that 80% of the mobility of the whole area account for only the following 15 zones from the total of 44 zones considered:

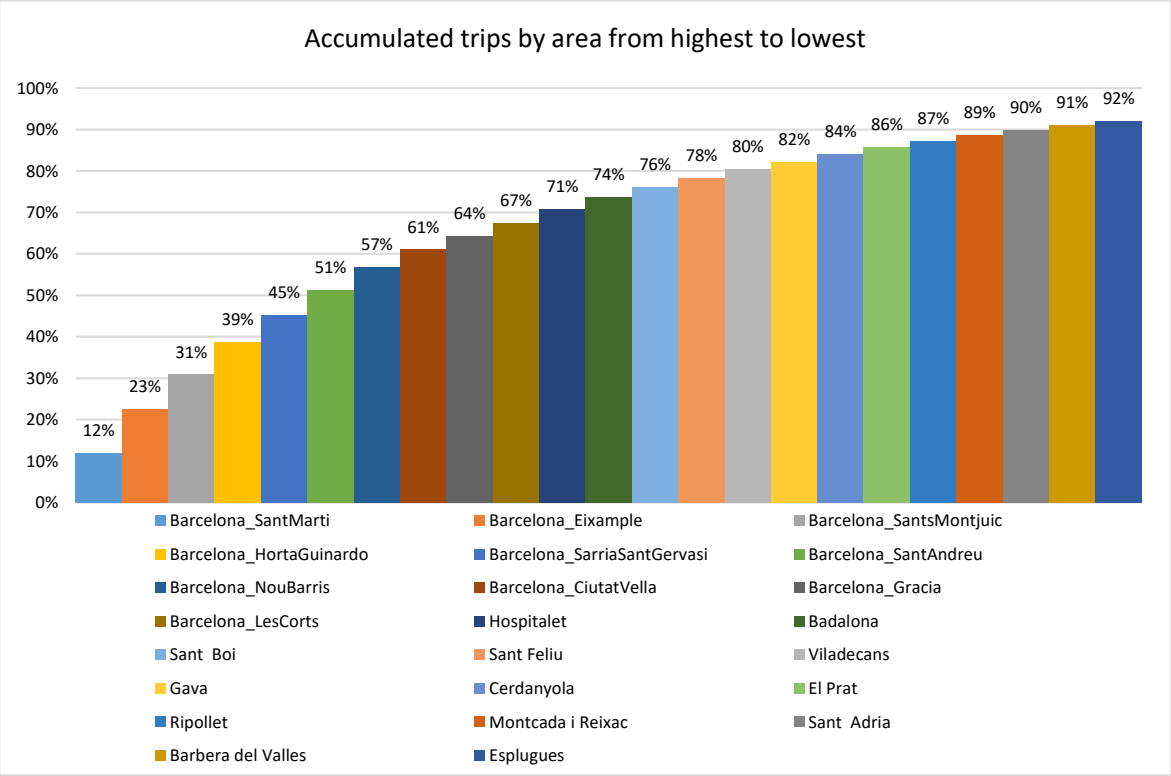

Figure 14. *Accumulated percentage of trips by area from highest to lowest.*

3.3.3. Inflows and outflows from the city of Barcelona

In order to analyse the entries and exits from Barcelona, the aforementioned areas have been grouped into 4 major clusters. These clusters have been determined by taking the 10 areas corresponding to the districts of Barcelona as the city cluster, and all the other areas grouped into 3 other clusters by closeness of *comarcas*.

Each cluster comprises the zones that can be observed in Figure 15:

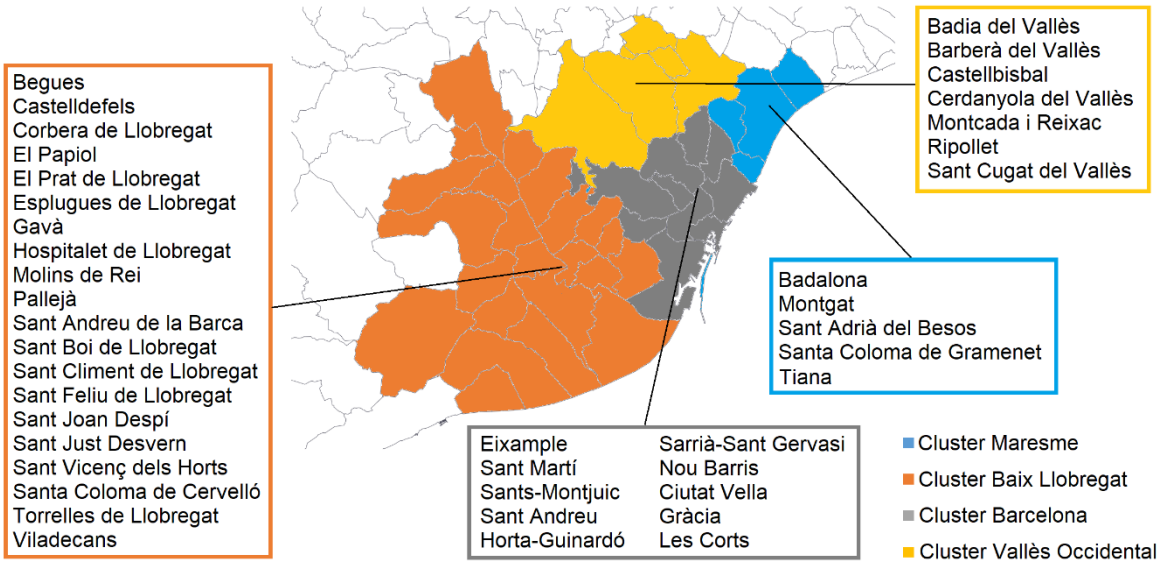

Figure 15. Map of the comarca-oriented clusters.

The absolute volume of journeys of the different zones have increased or decreased depending on the mobility restrictions as shown in previous sections, the entries and exits from the city increase or decrease the same amount for each year-time period

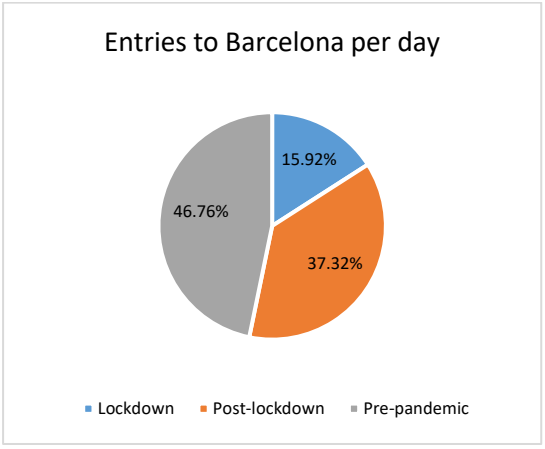

Figure 16. Percentage of trips that enter to Barcelona per day.

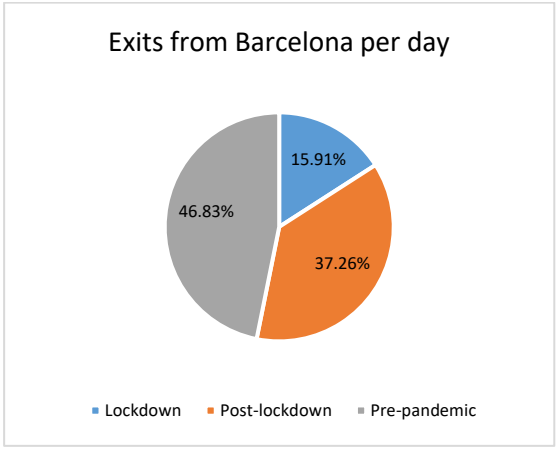

Figure 17. Percentage of trips that exit Barcelona per day.

Analysing the distribution of the average daily trips in and out of Barcelona by periods, it can be seen that the majority of trips that leave Barcelona re-enter and vice versa, as Figure 16 and 17 show. If this was not the case, very different proportions would be seen, since the sum of all the daily averages of the different routes would change greatly between entries and exits, and instead only vary by tenths of a percentage.

In addition, it is conspicuous that during the lockdown there was approximately 31% less

mobility in the entrances and exits of the city than during pre-pandemic. In the post-lockdown period, it increased but there was still approximately 10% less mobility than before the pandemic.

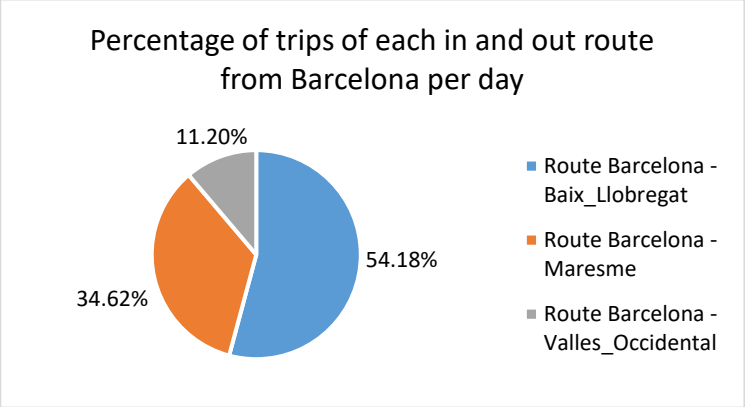

Figure 18. Percentage of trips by route from Barcelona per day.

Analysing the mobility that enters and leaves the city to go to areas of the first crown of Barcelona, according to Figure 18, it is the Barcelona - Baix Llobregat route and vice-versa the one with the highest number of travellers. The next busiest route is Barcelona - Maresme and vice-versa. Finally, the least frequented route is Barcelona - Vallès Occidental and vice-versa.

The distribution of passengers for each route can be understood by analysing the characteristics of the chosen clusters, in columns, in Table 1.

Characteristics for each cluster	Barcelona	Baix Llobregat	Vallès Occidental	Maresme
Gross value added. Industrial sector [million €]	5622.9	2496.00	671.1	581.1
Gross value added. Transportation and communications sector [million €]	9250.6	2558.0	836.2	230.2
Gross value added. Total [million €]	58239,3	15569.2	5035.1	4075.8
Number of family housing	594451	253969	75657	127290
Area [km^2]	101.36	339.31	146.86	42.86

Table 1. Characteristics of the zones picked for the four clusters mobility analysis of the first crown of Barcelona.
Source: [12] [13] [14] [15] [16] [17] [18] [19]

The route between Baix Llobregat and Barcelona represents more than half of the journeys in and out of the city towards the first ring of the city, probably because of its great industrial and transport sector activity, as it can be seen in the comparison of the Table 1. However, the whole area of the Baix Llobregat cluster is several times larger than the other clusters of the first ring. For all these reasons, it is justified that the route between Barcelona and Baix Llobregat contains more than half of the journeys made between Barcelona city and its first ring of municipalities [17].

The route between Barcelona and the Maresme region accounts for 35% of the mobility in and out of the city despite of its low level of economic activity. Observing Table 1, the Maresme cluster has many family housings compared to its area size. This means that there is a high daily mobility of people leaving the Maresme cluster to go to work to the city of Barcelona [19].

The route between Barcelona and Vallès Occidental only represents 11% of the mobility in and out of the city because, despite being a region with high industrial activity, the municipalities that make up the cluster of the first crown are not particularly industrialised. Moreover, this fact means that part of the daily commute to work is to other municipalities in the county rather than to the city of Barcelona [18].

(a) Pre-pandemic

(b) Lockdown

(c) Post-lockdown

Figure 19. Geographical representation of the OD matrix for 4 clusters.

All of the above observations can be seen in the origin-destination maps in Figure 19 (a), (b) and (c). Furthermore, it is clearly apparent how this distribution in the percentage of journeys in and out of the city of Barcelona is maintained for most of the routes for the three periods selected for the study (pre-pandemic, lockdown and post-lockdown), despite of the fact that the number of trips was significantly reduced during the lockdown and quite lower during the post-lockdown.

The proportion of trips in the OD matrices in Figure 19 should be complemented with the proportion on the same route according to the periods shown in Table 2. Otherwise, the legend

for the different periods a, b and c can be confusing, as they have different references.

	Pre-pandemic	Lockdown	Post-lockdown
From Barcelona to Baix Llobregat	48.12%	14.57%	37.31%
From Barcelona to Maresme	44.11%	19.18%	36.71%
From Valles Occidental to Barcelona	48.95%	12.29%	38.75%
From Baix Llobregat to Valles Occidental	42.16%	20.41%	37.44%
From Maresme to Valles Occidental	43.48%	18.96%	37.57%
From Baix Llobregat to Maresme	48.57%	11.84%	39.59%

Table 2. Percentage for the routes of the OD matrices from Figure 19.

It is interesting to note that during the period of the lockdown, the route between Barcelona and Maresme was the least affected by mobility restrictions. This can be seen in the legend to Figure 19 (b), as this route increases in thickness in relation to the rest. Looking at Table 2, it can also be seen that from the three inflow and outflow routes of Barcelona, it is the one that retains the highest percentage of journeys over the total of the three periods with 19.18% during the lockdown.

4. COVID-19 data analysis

This section is started with the description of the data-driven techniques used to analyse the spread of pandemic. The COVID-19 data collected is then explained and analysed to understand what information is available on COVID-19 in the study area.

4.1. Data-Driven techniques to analyse the pandemic spread

In terms of studies and visualizations, much of the data available on Covid-19 is used to explain the pandemic [20]. Although these methods are useful for highlighting the magnitude of the crisis, they are not sufficient to address and mitigate the issue [2]. These are also inadequate for decision-makers to anticipate the response to the spread of the virus and evaluate the effectiveness of the actions implemented.

In order to obtain mathematical models for epidemics, classic epidemic models are also useful. However, to estimate them accurately, many parameters of these models, such as infected rate and basic reproduction number, require data-driven approaches. Classic epidemic models, which are typically based on curve fitting techniques, often involve data to obtain parameters on different phases of the epidemic [2]. For these reasons, it is clear that more effective methods are required quickly in order to model and predict the spread and effects of the pandemic and afterwards assess the mitigation approaches that have been implemented.

It is possible to apply a range of data-based techniques, ranging from classical statistical and machine learning methods, such as linear regression and Bayesian inference, to advanced neural network-based models. To provide a good estimate, these techniques need adequate and high-quality data [2]. The amount of data will greatly differ from hundreds to millions of samples, depending on the technique used. In addition, to obtain an accurate model of a complex and dynamic system such as the COVID-19 pandemic, a wide variety of data can be needed.

Therefore, data from various disciplines is required, which hinders the task of data collection. Three pillars of data-driven approaches to the fight against Covid-19 are highlighted [2]:

- Informative variables for the development of an accurate model.
- The characterization of the pandemic of Covid-19. Clusters, models, forecasts, etc.
- Effective decision-making criteria.

4.1.1. COVID-19 common related variables

By bringing together different analyses and models studied in the literature a list of parameters is considered to observe the evolution of COVID-19 has been drawn up. The list of variables is large, since to develop accurate models, many aspects should be considered [2],[21]. Depending on their discipline, the variables considered can be divided into various categories.

COVID-19 variables (when possible, the data should be divided per gender, age range, etc.):

- Regional time series of the number of confirmed cases
- Suspicious cases
- Death
- Recovered
- Number of tests
- Hospitalized cases
- ICU cases
- Isolated positive cases
- Serology studies

Demographic variables:

- Population and density of population by location (for normalization of the rest of the variables)
- Age and gender
- Secondary health conditions related to higher COVID-19 mortality
- Socioeconomic status

Health system variables:

- Total number of ICU beds,
- Number of doctors and nurses,
- Personal protective equipment
- Respirators
- Number and types of tests

Policies

- Ban of mass gatherings
- School and kindergarten closures
- Contact restrictions
- Mandatory face masks

Government measures:

- Social distancing
- Movement restrictions
- Lockdowns

Weather variables:

- Temperature
- Relative humidity
- Radiation

Geographic variables:

- Locations of Covid-19 variables
- Nursing homes

Contamination variables:

- Air pollution (i.e., fine particulate matter PM2.5)

International and national mobility and connectivity:

- Number of international and national flights
- Number of international train connections
- Traffic patterns

Mobility:

- Retail and recreation
- Grocery and pharmacy
- Public facilities
- Workplaces
- Residential
- Transit stations

Numerous institutions of various kinds, such as governments, local entities, global institutions, institutions of the European Union (EU), universities, newspapers, etc., report on the development of the COVID-19 pandemic daily. Here below have been listed the most relevant and reliable. In particular, are highlighted those that provide updated information in the open-data repository with easy access on a regular basis. Some of the enumerated institutions make a great effort to provide consolidated data, describing the sources and limitations of the datasets provided in a rather exhaustive form.

Local information sources:

Open data from the Spanish Government [22]: An initiative launched in 2009 with the aim of promoting the opening of public information and the development of advanced data-based services. It is promoted by the Ministry of Economic Affairs and Digital Transformation and the Public Business Entity Red.es.

Open data from the Generalitat de Catalunya [23]: The Transparència Catalunya portal is managed by the Generalitat de Catalunya. It is a system of information that allows people easy and free access to information about local governments and other actors. These actors, due to their relationship with the Administration, are obliged to publish their data (entities, organizations, interest groups, etc.) so that citizens can identify them.

Open Data BCN [24]: Initiative promoted by public administrations with the main objective of making the most of available public resources, exposing the information generated or guarded by public bodies, allowing access and reuse for common good and for the benefit of interested persons and entities.

International information sources [2]:

World Health Organization [25]: WHO's primary role is to coordinate international health within the context of the United Nations system and to lead global health response partners. WHO offers constant updates on the latest situation around the world in the context of the Covid-19 pandemic.

Johns Hopkins University [26]: Johns Hopkins experts have been at the forefront of the international response to Covid-19 in global public health, infectious disease, and emergency preparedness since the beginning. This university offers a regular update on the global pandemic map. One of the most commonly used by researchers and journal media is the dataset supplied by Johns Hopkins University (JHU).

University of Oxford [27]: The Blavatnik School of Government is an Oxford University department that is focusing on the pandemic of Covid-19 and the policy responses that we see around the world. One of their Covid-19 research initiatives focuses on monitoring whether

governments around the world are responding to the pandemic and how they compare with others.

European Union: Access to open data released by EU agencies and bodies is provided by the European Data Portal (EDP) [28], which is an official open data portal of the European Union. The European Union has opened a specific data portal, called Covid-19 Data Portal [29], to support research on Covid-19.

Joint Research Centre [30]: The Joint Research Centre (JRC) is a science and information service of the European Commission, hiring researchers to provide independent scientific advice and support for EU policy.

Google: A visual map of Covid-19 has been created by the global technology company Google, where relevant information can also be found, both worldwide and by region [31]. Statistics on the number of cases confirmed, cases per million people (normalized data), the number of people rescued, and deaths are also presented. Google DataSet Search [32] is another relevant tool developed by Google, which can be used to collect data about Covid-19. You will find various data sets looking for the Covid-19 word. The application allows users to filter multiple fields in the datasets, such as last modified, download format, use rights, topic, and accessibility, etc.

MIDAS Network [33]: MIDAS is a global network of scientists and practitioners from academia, industry, government and non-governmental organizations who develop and use theoretical, statistical and mathematical models to enhance the understanding of the dynamics of infectious diseases in terms of pathogenesis, transmission, efficient control strategies and forecasting. They also developed a Covid-19 modelling platform, which offers an essential and reliable data resource catalogue, including datasets, webinars, and announcements of funding.

Covid-19 Data Hub [34]: The Covid-19 Data Center project has been sponsored by the IVADO Institute for Data Valorisation, Canada. The purpose of the project is to provide a single data centre for the research community by gathering worldwide fine-grained case data combined with demographics, air pollution, and other exogenous variables that help to better understand Covid-19. Furthermore, they include the R package to download datasets relevant to Covid-19.

In this section, a brief analysis of the COVID-19 data on the study area is carried out in order to characterise the territory.

4.2. Data source

Obtaining the COVID-19 dataset has been a great challenge due to the peculiarity of the zoning of the study area. Since the study differentiates between municipalities and districts, there is a great difficulty in finding data sources with this type of granularity as well as joining different datasets without double counting certain areas.

For this reason, a zoning corresponding to the basic health areas (ABS) has been used as a data source. This granularity size is small enough to assign a municipality or district to each basic health area. The Figure 20 shows the organisation of the basic health areas for each of Barcelona's 10 districts. All municipalities and districts have been related to one or more basic health areas except for Castellbisbal, Cervelló, El Papiol, Sant Climent de Llobregat, Santa Coloma de Cervelló, Tiana and Torrelles de Llobregat.

Figure 20. Example map of the Basic Health Areas (ABS) from the Barcelona districts. Source: [73]

The data was obtained from the Catalan open data portal [5]. It comes from different information systems of the Departament de Salut i del Servei Català de la Salut. In this way, a manual labelling of all the basic health areas with their corresponding municipality or district has been carried out to obtain an accurate allocation of COVID-19 positive cases for each area.

As it can be seen in Figure 21, the resulting dataset shows for each day, zone, gender and diagnostic procedure the number of cases identified as positive. Being the date of the case, the date of onset of symptoms, not the date of the diagnostic test. All of them are cases activated by the epidemiological surveillance services.

	date	zone	test_description	sex	num_cases
0	15/03/2020	01	Epidemiologic	M	1
1	19/03/2020	01	Epidemiologic	W	1
2	20/03/2020	01	Epidemiologic	W	1
3	21/03/2020	01	Epidemiologic	M	2
4	21/03/2020	01	Epidemiologic	W	2
...	...	...	...	...	...

Figure 21. Fragment of the processed dataset downloaded from the Catalan Open Data Portal source.

There are some areas where it has not been possible to link ABS and the municipality because some very small municipalities do not have health centres where COVID-19 positive cases can be counted. Another consideration about the dataset is that in cases of persons residing

in municipalities with a population of less than 200 inhabitants, the cases have not been considered (by the Catalan Heath Department).

4.3. COVID-19 data analysis related to mobility

Any individual on the move can be considered a danger and a possible vector of contagion during the outbreak of a pandemic. Therefore, in a variety of countries, lockdown policies that restrict daily movement, commuting and residential mobility have been implemented as the best available strategy to avoid (or slow) the spread of the disease, given the current shortage of available vaccines. Studies based on information provided by the major telecommunications companies from mobile devices have shown the efficacy of lockdown policies in decreasing the population's everyday mobility. The current literature on the impacts of lockdowns on residence transitions, however, is scarce [35].

4.3.1. Time series analysis

Looking at the temporal evolution of the number of COVID-19 infections, it can be seen how the data obtained and processed from the open data source corresponds to the known public data and is correlated to the mobility restrictions. It is possible to identify peaks of infections corresponding to the first and second waves.

In order to observe trends and patterns clearly, a 7-day moving average has been applied [7]. The value for 7 March is in fact the average of the journeys from 1 to 7 March and so on for all other days.

According to Figure 22, it is worth bearing in mind that the first peak of new infections per day was reached in April, which progressively decreased until the summer, when the same figures were reached again. From then on, it stagnated until the end of October, where there was a second peak in which the number of new infections doubled the first peak.

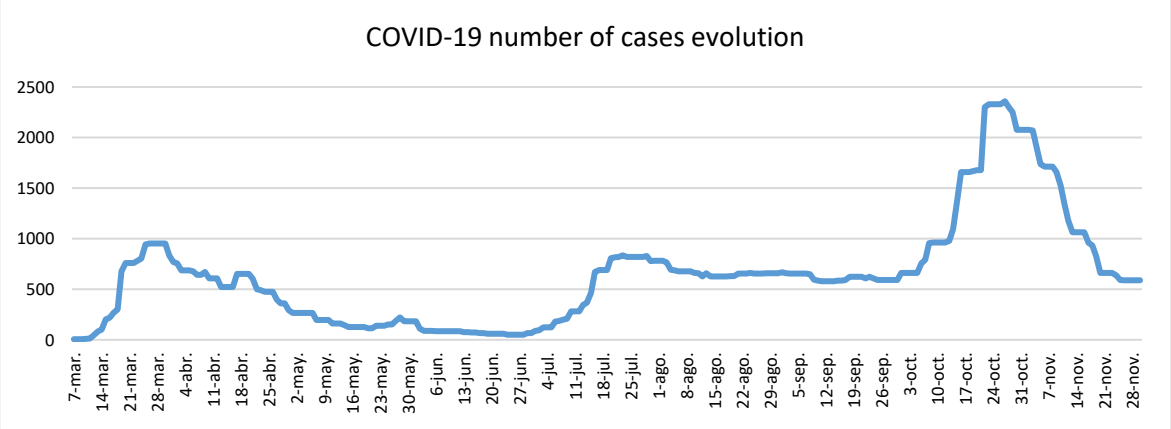

Figure 22. COVID-19 number of new cases evolution per day.

4.3.2. Spread by gender and zones

Comparing the COVID-19 spread by gender, it can be determined that among the daily cases of new infections, there is a 10% difference between women and men, having women the highest number of infections (as it can be seen in Figure 23). However, it should be stated that the population of Barcelona has 5% more women than men [36]. therefore, the absolute number of infections among women would be higher than among men in case that the probability was the same.

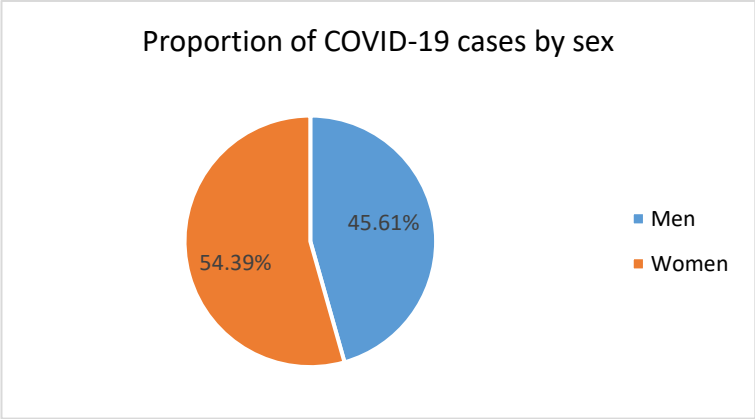

Figure 23. Proportion of COVID-19 new daily cases by gender.

With regard to gender-specific infections in the study areas, Figure 24 show a very uniform distribution in the proportion previously mentioned, corresponding to a higher infection rate among women than among men. It will therefore be necessary to wait for further correlation and causality analysis to confirm that there is a direct relationship between gender and COVID-19 infection.

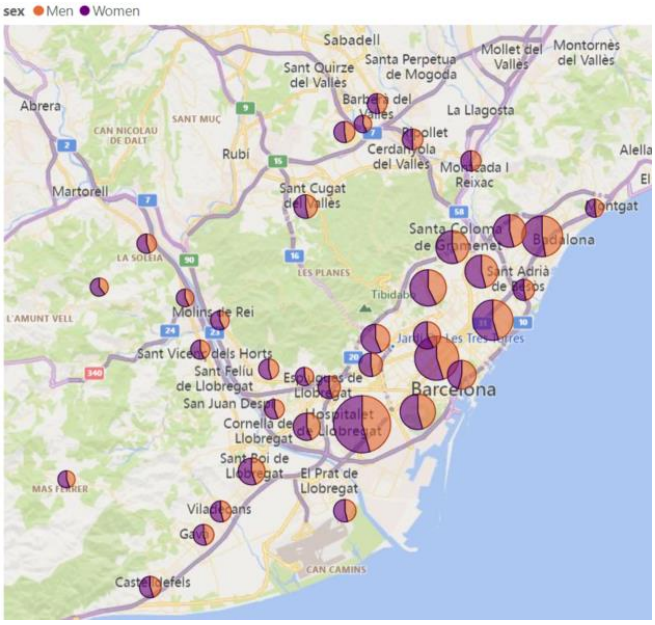

Figure 24. COVID-19 new daily cases for each zone by gender.

4.3.3. Spread by clusters

Looking at the proportion of new COVID-19 infections among the clusters, a close relationship with population density can be intuited. According to Figure 25, the cluster with the highest number of new infections is the one corresponding to the 10 districts of Barcelona, with almost 50%. The second cluster is Baix Llobregat with 31% of the total number of new daily infections, containing l'Hospitalet de Llobregat which is the most densely populated city in Catalonia. It is interesting to note that Figure 24 shows that l'Hospitalet de Llobregat is the area with the highest number of infections in the study area of the first ring of Barcelona. In third place is the Maresme cluster with 13.6% and finally Vallès Occidental with 7%. Despite having much lower numbers than the first two, their proportion is coherent, as the Maresme cluster has a higher population density than the Vallès Occidental cluster because it contains the areas of Badalona and Santa Coloma de Gramenet.

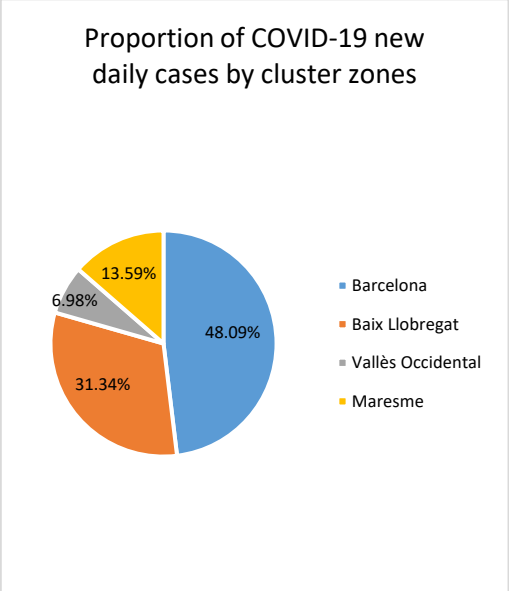

Figure 25. Proportion of COVID-19 new daily cases by cluster zones.

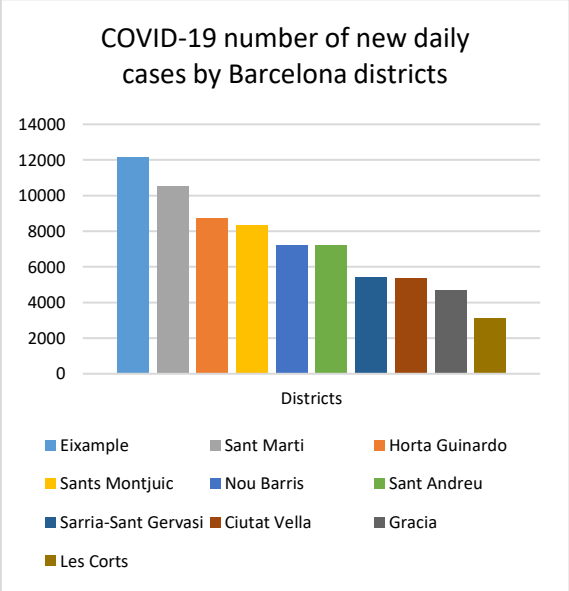

Figure 26. COVID-19 number of new daily cases by Barcelona districts.

Figure 26 shows, in descending order, Barcelona districts with the highest number of new COVID-19 infections per day for the whole pandemic until the November 30. The first is the most densely populated district, l'Eixample. Following in order, it is conspicuous that the next most contagious districts are those with a lower socio-economic level. This may be because of a multitude of factors such as the lack of resources, the kind of job profiles, the lack of space at home, etc. Therefore, the data of this study confirm a study by the Institut de l'Hospital del Mar which indicates that COVID-19 affects more the poorest areas of the city of Barcelona [37]. Finally, there are different districts with a higher socio-economic level and a variety of population densities.

5. Feature correlation and causality

The term "correlation" is commonly used to describe a relationship between two or more things (ideas, variables, etc.). In statistics, it refers to the relationship between two variables. It should be borne in mind that correlation does not imply causality, this being the relationship established between an event and its influence on the cause of another event. The goal of the section is to be able to quantify these relationships, assess its intensity and draw conclusions about the causality of the relationships [55].

However, first it is necessary to describe the data to be correlated in order to understand them and find their causalities.

5.1. Feature description

Three datasets have been combined for clustering: the mobility dataset provided by MITMA, the dataset of positive cases from COVID-19 and a dataset created from different sources with demographic data and characteristics of the territory.

Firstly, in order to be able to analyse the evolution during the year, the days have been grouped by periods according to the mobility restrictions imposed by the government. This is the most meaningful grouping, as it is the time characteristic that will show the most clearly marked patterns, as it has a large impact on both mobility and the number of COVID-19 positive cases. The study time span from 14 February to 30 November has been divided into the following nine periods:

	Period	Retail	Bars and restaurants	Mobility restrictions	Work activity
Pre-pandemic (None)	14 February to 14 March	No restrictions	No restrictions	No restrictions	No restrictions
Household confinement (HC) [38]	14 March to 28 March and 14 April to 17 May	Suspended the opening to the public with the exception explained below *	Closed	No meetings allowed	Teleworking is strongly recommended and encouraged
Household confinement with only essential services (OE)	29 March to 13 April	Only essential businesses may open	Closed	No meetings allowed	All workers in non-essential services are forced to stay at home
Phase 1 (Fase1) [39]	18 May to 7 June	30% of capacity	Terraces with 50% of capacity	Up to 10 persons, CCAA confinement and time slots for walking	Restart of the specified activities **
Phase 2 (Fase2) [40]	8 June to 17 June	Local trade at 40% of capacity and shopping centres at 30% of capacity	40% capacity and indoor table service with segregation measures	Up to 15 persons, CCAA confinement and time slots for walking	Restart of the specified activities ***
Phase 3 (Fase3) [41]	18 June	50% of capacity	50% indoor capacity and 75% terrace capacity	Up to 20 persons and CCAA confinement	Restart of the specified activities ****
New normality (NN) [42]	18 June to 24 August	Use of a mask is mandatory in public spaces	Use of a mask is mandatory in public spaces	No restrictions (although the use of a mask is mandatory)	Enhancing teleworking and increased contact traceability
Meeting restrictions to 10 persons (MR_10p) [43]	25 August to 24 October	Use of a mask is mandatory in public spaces	Use of a mask is mandatory in public spaces	Up to 10 persons	Use of a mask is mandatory in public spaces
Curfew (CFW) [44]	25 October to 20 November	Use of a mask is mandatory in public spaces	Opening hours restrictions and indoor consumption prohibited	Up to 6 persons, 10pm to 6am curfew and CCAA confinement	Use of a mask is mandatory in public spaces

Table 3. Mobility restriction periods by year period and its restrictions.

* Retail establishments selling food, beverages, essential goods and products, pharmaceutical and medical establishments, hairdressers, opticians and orthopaedic products, hygiene products, press and stationery, automotive fuel, tobacconists, technological and telecommunications equipment, pet food, internet, telephone and mail order shops, dry cleaners, and laundries.

** Opening by appointment of car dealerships, MOT stations and garden centres, High Performance Centres and other sports centres, public libraries for lending, return, reading and consultation, university laboratories and scientific research facilities, open-air markets with a capacity limited to one third, activities in the agri-food and fishing sectors, active and nature tourism, wakes and funerals, and filming of audio-visual works.

*** Individual attendance by appointment is allowed in senior citizens' centres for exceptional cases (centres without COVID-19 cases), the resumption of professional sports leagues, always behind closed doors and without public, wedding celebrations, beaches with reduced seating capacity and congresses and similar events with up to 50 attendees. The following openings at one-third capacity; cinemas and theatres with online ticket sales and capacity, common areas of hotels with a maximum capacity of one-third, exhibition halls, monuments and other cultural facilities with a maximum capacity of one-third, sports facilities and swimming pools by appointment.

**** Capacity increase to 75% in places of worship and wedding celebrations. Capacity increased to 50% in exhibition halls, monuments and other cultural facilities and common areas of hotels, indoor study and cultural events in libraries, tourist recreation centres, zoos and aquariums, casinos and betting shops, and children's and young people's activities. Reduced restrictions on funerals and wakes, active and nature tourism, and congresses and similar events.

5.1.1. Mobility dataset

In the mobility dataset, the origin-destination pairs have been eliminated to create the mobility indicator according to internal (do not leave the zone), inwards (arrive to the zone) and outwards (leave the zone) trips [8].

According to the orthodromic distance between origin and destination, distinguishing 4 distance ranges (originally 6 but there are no trips longer than 50 km within the first ring):

- 0,5 - 2 km
- 2 - 5 km
- 5 - 10 km
- 10 - 50 km

In addition, the records have been grouped separately for each hour of the day (0h to 23h) according to the following four determined time slots [45]:

- Morning rush: Period from 7.00h to 9.59h corresponding to the morning mobility peak hours.
- Afternoon rush: Period from 16.00h to 19.59h corresponding to the evening mobility peak hours.
- Night: Period from 22.00h to 5.59h in the evening. It is interesting to take this into account, as night-time mobility restrictions have been applied at certain times of the year.

- Off-peak: Period from 6.00h to 6.59h, from 10.00h to 15.59h, from 20.00h to 21.59h. In short, it includes the remaining hours not included in the three previous bands and which are off-peak hours with fewer journeys.

In this way, a labelling linked to the most significant hourly evolution during the day has been obtained, which will show more marked patterns for clustering.

Through this processing, the data structure of 16403 rows × 6 columns that can be seen in Figure 27 is obtained, where each row represents the number of trips per hour for the 5 corresponding characteristics.

	zone	covid_restrict	mobility_indicator	distance	timezone	tripshour
0	01	CFW	internal	0005-002	Afternoon_rush	14359357.0
1	01	CFW	internal	0005-002	Morning_rush	10862796.0
2	01	CFW	internal	0005-002	Night	4619638.0
3	01	CFW	internal	0005-002	Off-peak	12655613.0
4	01	CFW	internal	002-005	Afternoon_rush	2541045.0
...	...	...	...	...	...	...

Figure 27. Fragment example of the processed dataset downloaded from the MITMA open data source.

5.1.2. COVID-19 dataset

In the dataset of COVID-19 positive cases, as mentioned above, there have been major limitations because of the granularity of municipalities and districts in the study. This fact has prevented us from being able to obtain more variables for the clustering that represent the evolution of COVID-19. Therefore, the COVID-19 positive cases are segmented according to the day, the zone (resulting from manual labelling of ABS), the gender and the diagnostic procedure. Through this processing, the data is structured by 2304 rows × 5 columns, which can be seen in Figure 28, where each row represents the number of positive COVID-19 cases per day for the 4 corresponding characteristics [46].

	zone	covid_restrict	test_description	sex	num_cases
0	01	CFW	Epidemiològic	Home	0.027027
1	01	CFW	PCR probable	Dona	0.189189
2	01	CFW	PCR probable	Home	0.162162
3	01	CFW	Positiu PCR	Dona	27.189189
4	01	CFW	Positiu PCR	Home	25.405405
...	...	...	...	...	...

Figure 28. Fragment of the processed dataset downloaded from the Catalan Open Data Portal source.

5.1.3. Demographic and territory dataset

In the dataset with demographic and territory data, the same problem has arisen in relation to the zoning of the study. For one variable, it was necessary to search for the municipalities, then the districts and then join them together by processing each one, as different labels are often used.

Since there are common areas which can represent potential points of contagion, the following data sources have been processed:

- Facilities in Catalonia [47]
- Sports facilities in the city of Barcelona [48]
- Information from the stations of the new Bicing system in the city of Barcelona [49]
- Residences for the elderly in the city of Barcelona [50]
- Sports facilities in the city of Barcelona [48]
- Authorisations for terraces in public spaces in the city of Barcelona [51]

However, because of the peculiar zoning of the study it has only been possible to find records for all zones on facilities in Catalonia. Therefore, only this category has been taken from the previous list for the data on the clustering of the zones. This has a count of different types of labels that have been categorised for the study according to the following types of installations [52]:

- Work
- Culture
- Health
- Justice
- Retail
- Housing
- Economy
- Industries
- Sports
- Tourism
- Emergencies
- Innovation
- Communication
- Transportation
- Environment
- Universities
- Research
- Education
- Citizenship
- Primary sector
- Public administration
- Basic social services
- Especial social services

On the other hand, it is necessary to minimally characterise the study areas at the demographic level. For this purpose, the following data have been taken for the municipalities [53] and the districts [54]:

- Total population
- % of male population
- % of female population
- % of population aged 0 to 15
- % of population aged 15 to 64
- % of population aged 65 and over

The surface area and population density of each area (municipalities [13] and districts [55]) have also been considered.

5.2. Correlation techniques

For its calculation, the data has been processed by generating a matrix that groups the 3 datasets of the study. Firstly, the mobility data has been merged with the COVID-19 data using the area and the period of mobility restrictions as common variables. Subsequently, the territorial data (demographics and facilities) have been added using the zones as a common variable. When all the variables have been put together, the matrix is divided into qualitative and quantitative variables, as the methods of calculating correlations change for each of them.

Table 2 below shows the most common correlation calculation methods according to the type of variables under study:

Variable Y\X	**Quantitative X**	**Qualitative X**
Quantitative Y	Pearson *r*	Point Biserial r_{pb}
Qualitative Y	Point Biserial r_{pb}	Phi, L, C, Lambda

Table 4. Comparison of correlation calculation methods according to the type of variable.

For the correlation study, the Pearson coefficient has been chosen for quantitative variables and the Point Biserial for qualitative variables.

The features have been standardised beforehand by removing the mean and scaling to unit variance. In this way, COVID-19 case data can be compared with mobility data even if they are orders of magnitude exaggeratedly different. For a sample x, with a mean μ and a standard deviation s, the new scaled variable z is obtained:

$$z = \frac{x-\mu}{s} \qquad (Eq.\ 6)$$

5.2.1. Pearson correlation

The Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations as it can be seen in Equation 7 [56]:

$$r_{xy} = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y} \qquad (Eq.\ 7)$$

where:

cov(X, Y) is the covariance

σ_x is the standard deviation of X

σ_y is the standard deviation of Y

When applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample *correlation coefficient* or the sample *Pearson correlation coefficient* [56]. Given paired data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of n pairs, r_{xy} is defined as in Equation 8:

$$r_{xy} = \frac{\sum_{i=1}^{n}[(x_i-\mu_x)\cdot(y_i-\mu_y)]}{\sqrt{\sum_{i=1}^{n}(x_i-\mu_x)^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\mu_y)^2}} \qquad \textit{(Eq. 8)}$$

where:

n is sample size

x_i, y_i are the individual sample points indexed with i

$\mu_x = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i$ is the sample mean; and analogously for μ_y

The correlation values shown in Figures 29 and 30 are the result of taking trips per hour as the Y variable and each variable on the horizontal axis as the X variable. By applying Python functions based on the formulas shown, which can be found in the code of Annex, the coefficients have been obtained. The figures show the coefficients in decreasing order with their value on the vertical axis. The correlations in Figure 29 are of the variable hourly trips with respect to the corresponding one on the horizontal axis. The correlations in Figure 30 are of the variable number of positive COVID-19 cases with respect to the corresponding one on the horizontal axis.

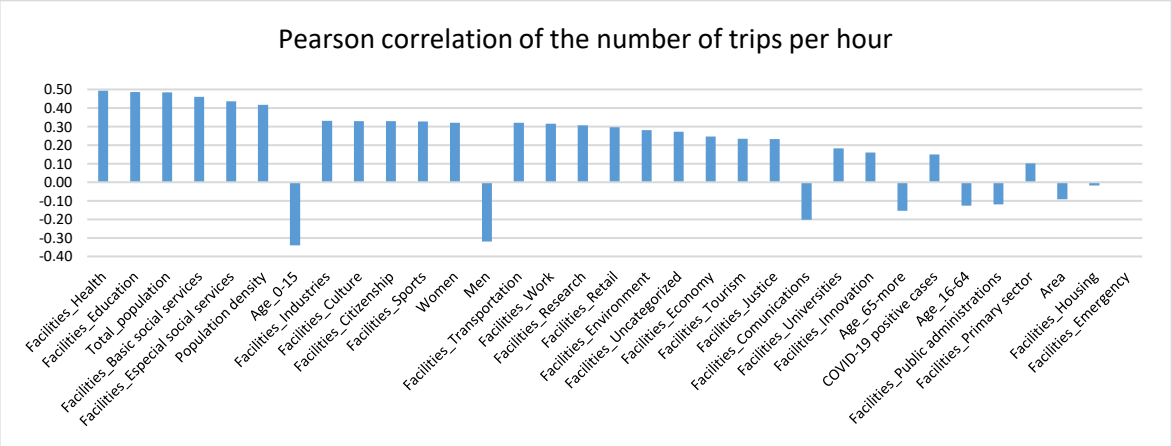

Figure 29. Pearson correlation of the number of trips per hour with all the quantitative features.

Observing the value of the Pearson coefficients in Figures 29 and 30, medium (0,3 – 0,5) and small (less than 0,3) correlations are found. All the quantitative variables correspond to the demographic and territorial dataset, and observing the values obtained, it can be determined

that there is a correlation between the zones of the first crown of the metropolitan area of Barcelona with their mobility and the number of cases of COVID-19 [57].

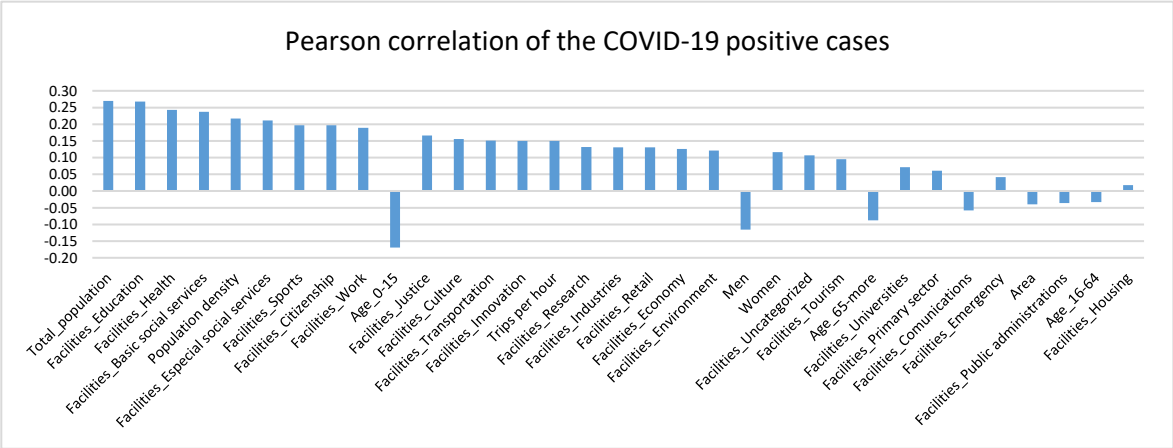

Figure 30. Pearson correlation of the COVID-19 positive cases with all the quantitative features.

In addition to correlation, causality can also be established. It is very consistent that the greater the population, the greater the population density, and the more facilities of all kinds, the more mobility and more contagions occur because there is more people moving around in less space and there is a greater existence of common physical spaces.

In particular, as can be seen in Figures 29 and 30 there is a strong direct correlation of the variables of total number and density of population, health facilities, education facilities and social services (basic and specialized) with trips per hour and COVID-19 cases.

On the other hand, an unexpected inverse correlation is observed between the % of population aged 0-15 and mobility and COVID-19 cases. One reason for this correlation may lie in the fact that many people move to the suburbs or to areas farther away from the centre in search of lower prices, more space and more tranquillity. In turn, these areas correspond to those with the lowest mobility and COVID-19 cases, which is why it is consistent that there is an inverse correlation between both variables and the % of the population aged 0-15.

Another unexpected correlation is with respect to gender, where females have a low but significant direct correlation with the number of COVID-19 cases. This correlation could be due to the fact that women are more exposed to COVID-19 because they occupy front-line jobs, such as caregiving, as indicated by a study in the autonomous community of Madrid [58].

5.2.2. Point biserial correlation

When one variable is dichotomous (takes one of only two possible values), the point biserial correlation coefficient (r_{pb}) can be used. The variable can be "naturally" dichotomous, such as whether a coin lands heads or tails, or artificially dichotomized (which is the case). An artificially

dichotomized variable is one that originally had more than two qualitative values and has been split into as many binary variables as qualitative values can take. When a new variable is artificially dichotomized, it is possible that the new dichotomous group of variables introduce a new underlying structure [59].

To calculate r_{pb}, assume that the dichotomous variable has the two values 0 and 1. If we divide the data set into two groups, group 1 which received the value "1" and group 2 which received the value "0", then the point-biserial correlation coefficient is calculated as follows in equation 9 [59]:

$$r_{pb} = \frac{\mu_1 - \mu_0}{s_n} \cdot \sqrt{\frac{n_1 \cdot n_0}{n^2}} \qquad \textit{(Eq. 9)}$$

μ_1 being the mean value on the continuous variable Y for all data points in group 1, and μ_0 the mean value on the continuous variable Y for all data points in group 2. Further, n_1 is the number of data points in group 1, n_0 is the number of data points in group 2 and n is the total sample size.

Where s_n is the standard deviation, whose formula can be seen in equation 10:

$$s_n = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu_x)^2} \qquad \textit{(Eq. 10)}$$

x_i being the individual values of the sample, μ_x the mean of these values and n the total sample size.

By applying these calculations to the code in Annex, the coefficients shown in Figures 31 and 32 have been obtained.

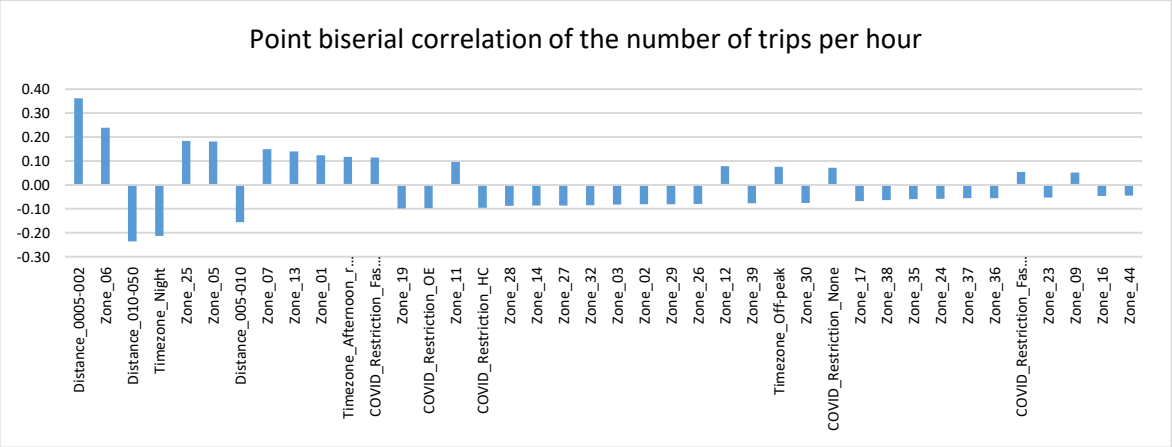

Figure 31. Point biserial correlation of the number of trips per hour with all the qualitative features.

Looking at the Point biserial correlation values in Figures 31, four medium correlations and

several low correlations are observed.

The trips with distance 0.5 – 2 km (which a big part is intrinsically internal trips) are directly correlated with the total number of trips. Therefore, as seen in the preliminary analysis of mobility data, this range of distance is found to represent a large part of mobility. On the other hand, the distance range from 10 km to 50 km presents medium values of inverse correlation, representing a very small part of the mobility.

Zone 6 corresponding to l'Eixample presents a direct median correlation, so that a very representative part of the total trips is involved. On the other hand, the coefficient for the night-time slot shows an inverse correlation, so that, as expected, the lowest number of trips per hour of the day occurs between 22h and 6h.

The other correlations are not particularly strong or unexpected. They are simply useful to see in order which zones, restriction periods, distances and time slots have the greatest influence (direct or inverse) on the number of trips.

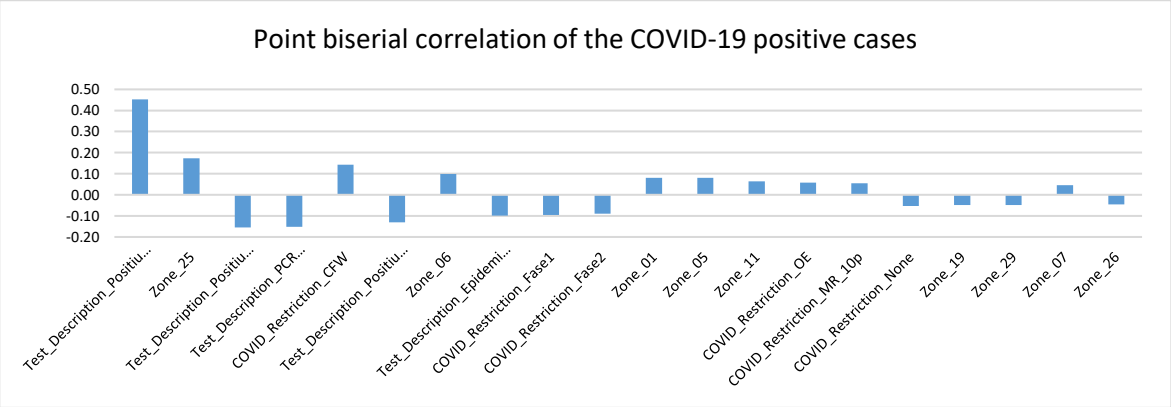

Figure 32. Point biserial correlation of the COVID-19 positive cases with all the qualitative features.

Figure 32 shows Point biserial values in relation to COVID-19 positive cases, all of them in order of importance and their values. It is readily apparent that only the number of positive PCR tests have a direct correlation, as expected, while the rest are predictable weak correlations.

When finding correlations there is a great limitation linked to the lack of a common public data source. For the calculation of all the above coefficients, the three large datasets explained in the feature description section have been put together through the maximum common variables. For this purpose, the mobility data and the positive cases of COVID-19 data have been joined using the zones and the periods of mobility restrictions as common variables. Subsequently, the dataset with demographic and territory data has been added using the zones as a link variable.

In consequence, is exceedingly difficult to mathematically correlate COVID-19 contagion patterns with mobility by daily time zones or according to distances, since the nexus between variables are just the zones of the territory (municipalities and districts) and the restriction periods. It would only be possible to correlate them if there were radically different patterns for each zone and period, which is unlikely in the case of the study area, being so small geographically and with so much mobility flow. In this case the correlation would not be with the characteristics but with the zones and periods.

Quantitative variables are able to show strong correlations because all records have representation of the variables. On the other hand, this is not the case for qualitative variables, as artificially dichotomized variables are generated for each qualitative value that any variable may take. In this way, it has been adopted as many variables as qualitative values there may be. In an attempt to measure quality, p-values have been observed [60] but there is little point in doing so since it corresponds to a structure resulting from a count and not from individually recorded samples.

As mentioned, due to the difficulty of finding correlations mathematically by means of coefficients, a study by zones is necessary to observe how they are characterized. Since there are so many zones (municipalities and districts), it is necessary to carry out the study by groups. Previously, a preliminary observation of the data has been made according to four clusters chosen arbitrarily based on the existing territorial units, taking the 10 districts of Barcelona and the rest of the municipalities grouped according to their proximity to the comarcas. However, it is much more effective to make a grouping of the areas using machine learning techniques that allow us to form groups according to their similarity. In this way, when they are analysed, they present more noticeable patterns due to the similarity between them. For this reason, in the following section unsupervised learning algorithms will be analysed to perform these groupings.

6. Pattern recognition through machine learning techniques

Machine learning (ML) is the study of computer algorithms that automatically adjust their performance from exposure to information encoded in data. This learning is achieved via a parameterized model with tuneable parameters automatically adjusted according to a performance criterion [61].

There are three major classes of machine learning [62]:

- Supervised learning: Algorithms which learn from a training set of labelled examples (exemplars) to generalize to the set of all possible inputs. Examples of techniques in supervised learning include regression and support vector machines.

- Unsupervised learning: Algorithms which learn from a training set of unlabelled examples, using the features of the inputs to categorize inputs together according to some statistical criteria. Examples of unsupervised learning include k-means clustering and kernel density estimation.

- Reinforcement learning: Algorithms that learn via reinforcement from a critic that provides information on the quality of a solution, but not on how to improve it. Improved solutions are achieved by iteratively exploring the solution space. We will not cover RL in this course.

In the case of this master's thesis, the aim is to group the zones according to the similarities of their data through unknown patterns using data clustering. For this reason, it is necessary to focus on unsupervised learning algorithms since the data to be analysed are not labelled according to a predefined criterion. The aim is precisely to discover unknown patterns.

6.1. Unsupervised learning algorithms

The problem of finding a hidden structure in unlabelled data is known as unsupervised learning. As the examples given to the learner are unlabelled, there is no error or reward signal to test a potential response. Based on the possible needs of the study in question, the following unsupervised learning techniques could be used [63]:

- Clustering: Unsupervised learning technique which divides samples into groups with no pre-defined categories/classes available.
- Dimensionality reduction: Principal component analysis, independent component analysis, non-negative matrix factorization.

- Detecting outliers (statistically atypical values): Find unusual events (e.g., malfunction).
- Novelty detection: Find changes in data.

The major data sources of the master's thesis have been previously treated, and the rest has been easily cleaned, so it is unnecessary to use machine learning to detect outliers for data cleaning. It is also not interesting to apply novelty detection as it is a fixed data source that can be analysed in one go.

Therefore, in order to recognise patterns by analysing groups of similar zones, clustering techniques have been applied to group the 44 zones of the study area. These clusters have been analysed to characterise and understand the behaviour of the municipalities of the first crown of the metropolitan area of Barcelona and the 10 districts of the city. In addition, dimensionality reduction techniques have also been used to deal with a large number of variables.

6.1.1. Clustering techniques

There are two big families of clustering techniques:

Partitional algorithms:

Start with a random partition and redefine it iteratively. Partitional algorithms can be divided in two branches:

- *Hard partition algorithms*, such as K-mean, assign a unique cluster value to each element in the feature space.
- *Soft partition algorithms*, such as Mixture of Gaussians, can be viewed as density estimators and assign a confidence or probability to each point in the space.

Hierarchical algorithms:

They create nested clusters by successively merging or splitting them. This cluster hierarchy is depicted as a tree (or dendrogram). The unique cluster at the tree's root collects all the samples, while the leaves are clusters with only one sample. It is a useful tool because of its interpretability, the technique yields a tree that depicts the degree of similarity between the samples. The partitioning is calculated by selecting a cut at a specific level on that tree, which is known as a dendrogram. They are divided into the following categories [64]:

- *Bottom-Up agglomerative*: Starts with each sample data in a separate cluster. Then, repeatedly joins the closest pair of clusters until there is only one cluster. The history of merging forms in a binary tree or hierarchy is obtained.

- *Top-Down divisive*: Starts with all the data in a single cluster. It considers every possible way to divide the cluster into two and chooses the best division. The history of divisions forms in a binary tree or hierarchy is obtained.

Since the result is intended to be specific clusters of the zones, soft partition algorithms are discarded because they offer unnecessary information, and the clusters definition is more diffuse. So, they are less suitable than the others for the purpose of the study.

In Table 5 it can be seen a comparison with the clustering techniques that have been considered for the study:

Method name	Parameters	Scalability	Use	Metric	Input
k-means clustering	Number of clusters	Very large number of samples and medium number of clusters	General purpose, even cluster size, flat geometry and not too many clusters	Distances between points	Data samples
Spectral clustering	Number of clusters	Medium number of samples and small number of clusters	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbour graph)	Affinity/ distance matrix
Hierarchical clustering	Number of clusters / Threshold	Large number of samples and large number clusters	Many clusters, possibly connectivity constraints	Distances between points	Data samples

Table 5. Comparison table with the clustering techniques considered.

Looking at the comparison in Table 3, spectral clustering has been discarded because by creating the starting graph you get a flat geometry where practically all nodes are connected [65]. Between k-means and hierarchical, the hierarchical algorithm could be used but it is considered more convenient to implement the k-means algorithm because of its simplicity, its high sensitivity to differences in the variance of the characteristics, and its resistance to irrelevant dimensions (as it does not introduce biases) [66].

6.2. K-means clustering

The K-means algorithm clusters data by trying to separate samples in K groups of equal variances. In other words, the K-means algorithm divides a set of N samples X_i into K disjoint clusters, each described by the mean of the samples in the cluster. The means are commonly called the cluster “centroids” [67].

First, starting from the 3 datasets explained in the previous section, their qualitative variables have been pivoted in such a way that a column has been created for each combination of one of them, taking the zones as rows. In other words, 3 pivoted matrices have been created for

each of the 3 datasets, in which the columns are all the possible combinations of qualitative variables, each row represents the values of an area and the cells represent the quantitative values of each matrix (number of trips, COVID-19 cases, population density, number of facilities, etc.). Subsequently, the row matrices have been concatenated. This results in a matrix of 44 rows (zones) x 554 columns (combined qualitative features) which can be seen in Figure 33 it brings together all the data. This type of structure has been chosen to ensure that the clustering algorithm will not assign different clusters for the same zone, as it will perform row groupings.

When the matrices are pivoted, boxes with no value appear because there are physically impossible combinations. For example, if the maximum distance of an end-to-end zone is less than 5 km, it will be impossible for this zone to have internal trips with a distance greater than 5 km. For this reason, all these non-existent combinations that have Nan as a value are changed to 0 to give them a physical meaning. If such trips cannot really exist, it means that there will be 0 such trips (the same for COVID-19 cases and territory data).

	zone	covid_restrict	mobility_indicator	distance	timezone	tripshour
0	01	CFW	internal	0005-002	Afternoon_rush	14359357.0
1	01	CFW	internal	0005-002	Morning_rush	10862796.0
2	01	CFW	internal	0005-002	Night	4619638.0
3	01	CFW	internal	0005-002	Off-peak	12655613.0
4	01	CFW	internal	002-005	Afternoon_rush	2541045.0
...	...	...	...	...	...	...

zone	('tripshour', 'CFW', 'internal', '0005-002', 'Afternoon_rush')	('tripshour', 'CFW', 'internal', '0005-002', 'Morning_rush')	('tripshour', 'CFW', 'internal', '0005-002', 'Night')	('tripshour', 'CFW', 'internal', '0005-002', 'Off-peak')	('tripshour', 'CFW', 'internal', '002-005', 'Afternoon_rush')	...
1	1.435936e+07	1.086280e+07	4619638.297	1.265561e+07	2541045.297	
2	9.579973e+04	6.540478e+04	45829.081	8.577860e+04	0.000	
3	1.196805e+06	8.463498e+05	399483.108	1.055706e+06	70842.081	...
4	3.084486e+06	2.065257e+06	805414.973	2.689856e+06	33412.324	
5	1.082831e+07	7.975523e+06	3262519.243	9.405291e+06	943233.324	
...			...			

Figure 33. Evolution from one of the three datasets structure to the pivoted and concatenated dataset with the zones as rows and the combined qualitative features as columns.

On the other hand, in the areas of Castellbisbal, Cervelló, El Papiol, Sant Climent de Llobregat, Santa Coloma de Cervelló, Tiana and Torrelles de Llobregat it has not been possible to compute the number of COVID-19 infections, as they do not have their own basic health area code. Therefore, for the grouping, the neighbouring area most similar to it has been included in its travel and territorial characteristics, taking the sum of the two as the new area. Zones

have been grouped as shown in Table 6. With this aggregation a data matrix of 37 rows (zones) by 554 columns (combined qualitative features) is obtained.

Zones with no COVID-19 data	Most similar neighbouring area
Castellbisbal	Sant Andreu de la Barca
Cervelló	Pallejà
El Papiol	Molins de Rei
Sant Climent de Llobregat	Viladecans
Santa Coloma de Cervelló	Sant Vicenç dels Horts
Tiana	Montgat
Torrelles de Llobregat	Begues

Table 6. Assignment list of zones without COVID-19 data.

Finally, the features have been standardised by removing the mean and scaling to unit variance. In this way, COVID-19 case data can be compared with mobility data even if they are orders of magnitude exaggeratedly different. For a sample x, with a mean μ and a standard deviation s, the new scaled variable z is obtained:

$$z = \frac{x-\mu}{s} \qquad \textit{(Eq. 8)}$$

Partitional clustering aims to partition a given data set into disjoint subsets (clusters) in order to optimize specific clustering criteria. The clustering error criterion computes each point's squared distance from the corresponding cluster centre (also called inertia) and then adds these distances for all points in the data set. The k-means algorithm is a popular clustering method that minimizes clustering error.

The K-means algorithm aims to choose centroids minimizing the inertia or within-cluster sum-of-squares criterion:

$$Clustering\ Error\ Criterion\ (Inertia) = \sum_{i=0}^{n} \min\left(\left\|x_{i,j} - \mu_j\right\|^2\right) \qquad \textit{(Eq. 9)}$$

Inertia, or the within-cluster sum of squares criterion, can be recognized as a measure of how internally coherent clusters are.

Another important coefficient to evaluate the goodness of the clustering, the silhouette coefficient is used. It measures the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample [68] represented in Figure 34.

$$Silhouette\ score = \frac{\text{mean nearest cluster distance} - \text{mean intracluster distance}}{\max\left(mean\ nearest\ cluster\ distance, mean\ intracluster\ distance\right)} \qquad (\textit{Eq.10})$$

Its value ranges from -1 to 1. The closer it is to 1, the means clusters are well apart from each other and clearly distinguished. 0 value means that the clusters are indifferent, or we can say that the distance between clusters is not significant. If it is closer to -1, means that the clusters are assigned in the wrong way.

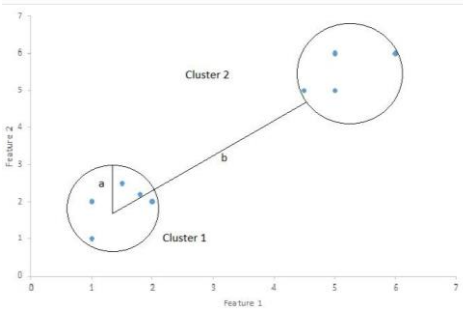

Figure 34. Graphical representation of the silhouette coefficient.

The silhouette coefficient is the mean of all the silhouette score values for all the clusters.

Based on the above criteria, the algorithm follows the following steps at the conceptual level:

1) Initialize the value K of desirable clusters.
2) Initialize the K cluster centres, e.g. randomly.
3) Decide the class memberships of the N data samples by assigning them to the nearest cluster centroids (e.g. the centre of gravity or mean).
4) Re-estimate the K cluster centres, by assuming the memberships found above are correct.
5) If none of the N objects changed membership in the last iteration, exit. Otherwise go to step 3.

Several issues should be considered:

- It is non-deterministic - depends on the initialization.
- Inertia makes the assumption that clusters are convex and isotropic, which is not always the case. It responds poorly to elongated clusters, or manifolds with irregular shapes which is not the case of the data of the study.
- Given enough time, K-means will always converge to the global minimum.
- The algorithm requires the number of clusters to be specified (inconvenient solved with the elbow technique).
- It scales well to large number of samples and has been used across a large range of application areas.

As it has been stated, since it is a non-deterministic algorithm, is a local search procedure with the well-known drawback that its performance is highly dependent on the initial starting conditions. For this reason, it is necessary to consider that the computation has been done several times, with different initializations of the centroids. The method used in the Annex helps to address this issue. The k-means++ initialization scheme, which has been implemented in scikit-learn (use the init='kmeans++' parameter), initializes the centroids to be (generally) distant from each other, leading to provably better results than random initialization [69].

6.2.1. The elbow technique

Subsequently, a heuristic technique is used to compute the compactness of different clustering with different number of clusters to determine the optimal number of clusters, it is called the elbow technique. The method consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of clusters to use.

Clusters must be compact according to the definition of clustering. The distance between the cluster's members and its centroid can be used to determine compactness. The average distance to the cluster's centroid is a crude indicator of the cluster's overall quality. In the process of comparing this value to the number of clusters, the elbow technique distinguishes two phases. In the first phase the average will drop dramatically. Then it will gradually stabilize in the second phase. The elbow technique entails deciding on the value at which the transition occurs [70].

Using the code of the Annex, compute the compactness values of the scaled data for 2 to 6 clusters. It has not been considered taking more than 6 clusters, as it is too impractical to analyse and the usefulness of using clusters is lost.

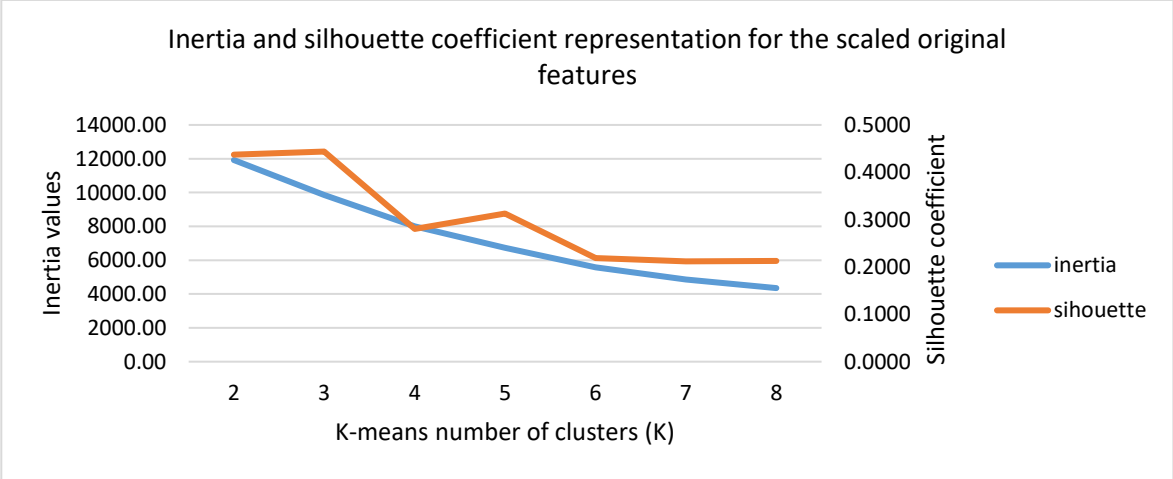

Figure 35. Compactness representation of the elbow technique with the original scaled features.

Observing Figure 35, it is not especially clear when the change between the 2 phases occurs. Therefore, a decision cannot be made solely on the basis of the quality of the clusters, it must also be considered the expected use and usefulness of the clusters. Performing 2 clusters is discarded because performing 3 clusters gives a better silhouette coefficient value, reducing greatly the inertia of the clusters (more compact clusters). The decision is between 3, 4 and 5. Taking 4 gives a worse silhouette coefficient but with much lower inertia. By taking 5 clusters, the silhouette coefficient improves, but in the change from 4 to 5 clusters the inertia is not reduced as much as in the change from 3 to 4 clusters. Therefore, it is very likely that the elbow of the graph is found for 4 clusters. For the remaining clusters, the reduction in inertia becomes

smaller and smaller as more clusters are used. Therefore, they will not be taken into account for further analysis. In short, considering the values with 3, 4 and 5 clusters, the best solution is to compute the assignment for both numbers of clusters and make a decision based on the results obtained.

6.2.2. Principal Component Analysis (PCA)

In addition, the algorithm has also been computed by previously decomposing the data by performing a Principal Component Analysis (PCA). PCA is a statistical procedure that decomposes a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance. It uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

This transformation is defined in such a way that the first principal component has the largest possible variance (accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The principal components are obtained as the eigenvectors of the covariance matrix, hence are orthogonal.

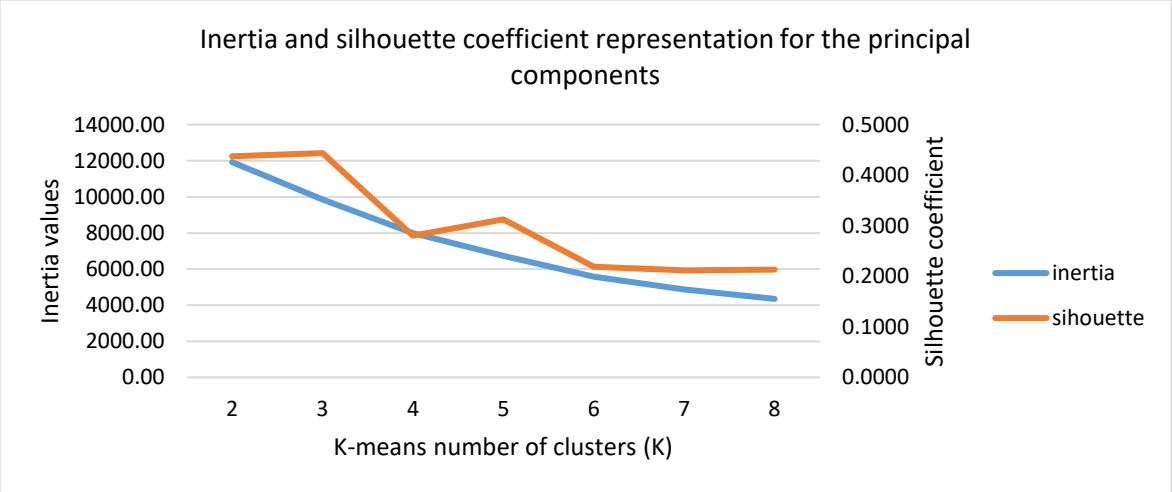

Figure 36. Compactness representation of the elbow technique with the principal components.

Applying the k-means clustering algorithm with the principal components instead of the original features shows a small improvement of the silhouette coefficient and the inertia value. Looking at the Figure 36, no significant trend change is observed with respect to Figure 35 that would change the number of clusters to be studied.

6.2.3. Interpreting the k-means clustering results

Comparing the exact results of the values obtained in Table 7 with the original features and with the principal components, it can be seen how the values improve but there is no significant trend change. Moreover, for both clusterings (original scaled features and principal components) the same zone groupings are taken (as expected).

Number of clusters	Original scaled features		Principal components	
	Inertia	Silhouette	Inertia	Silhouette
K = 3	9857,19	0,4438	8889,63	0,4732
K = 4	7995,85	0,2806	7047,15	0,3456
K = 5	6727,78	0,3126	5781,22	0,3479

Table 7. Comparison between the number of clusters and the use of features (original scaled or the principal components).

Comparing the results for 3 and 4 clusters, the assignment of zones changes a lot, since a new cluster is formed with 5 zones (Badalona, Sant Martí, Eixample, Sants-Montjuic, Hospitalet de Llobregat) which, as seen in previous sections, represent an important part of the mobility of the first crown of the metropolitan area of Barcelona. Therefore, considering that using 4 clusters greatly reduces inertia, and that it is not a large number of groups to analyse, the use of 3 clusters is discarded.

Comparing the results for 4 and 5 clusters, it can be seen in Figure 37 how the algorithm takes exactly the same groupings of zones for both numbers of clusters except for the fifth cluster which is only formed by Sants-Montjuic. Although taking 5 clusters results in a better silhouette coefficient, the inertia is not reduced as much as in the step from 3 to 4. Moreover, knowing that this is a new single-zone cluster, its separate study is not particularly enriching if the reduction in inertia does not compensate for it.

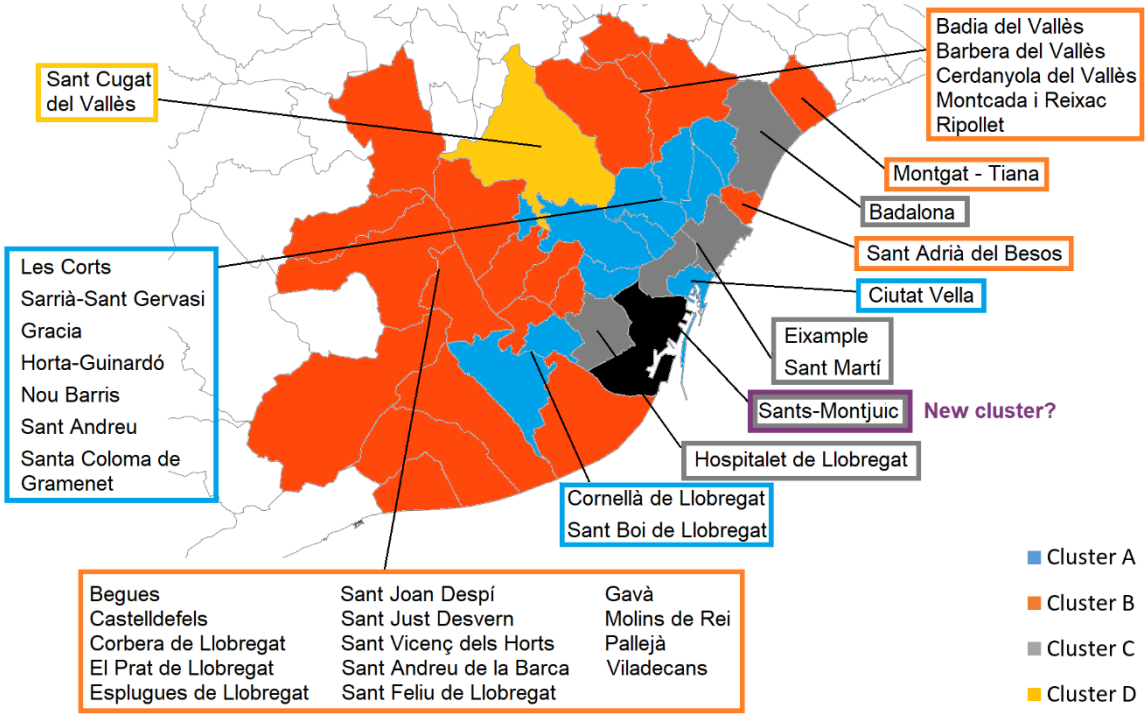

Figure 37. Assignation map of the clusters grouped by the k-means algorithm for k=4 and k=5.

This has been a difficult decision, as the indicators point to a higher quality clustering using 5 groups, but analysing the resulting groupings, it is more convenient to take 4 clusters. Therefore, as they are groupings that should be useful later on to characterise the territory and to have a better understanding of its behaviour by similar groups, it has been decided to use 4 clusters. It has also been considered counterproductive to add the Sants-Montjuic cluster, as it represents an excessively small area for the little improvement it introduces in the clustering.

6.3. Data driven clusters characterisation

Figure 38 shows the final allocation resulting from applying the clustering algorithm for 4 clusters. Broadly speaking, they are distributed in the territory as follows:

- Cluster A: Upper half of the districts of Barcelona and adjoining municipalities.
- Cluster B: Municipalities of the first crown far from the city.
- Cluster C: Lower half of the first crown and neighbouring municipalities.
- Cluster D: Formed only by Sant Cugat del Vallès.

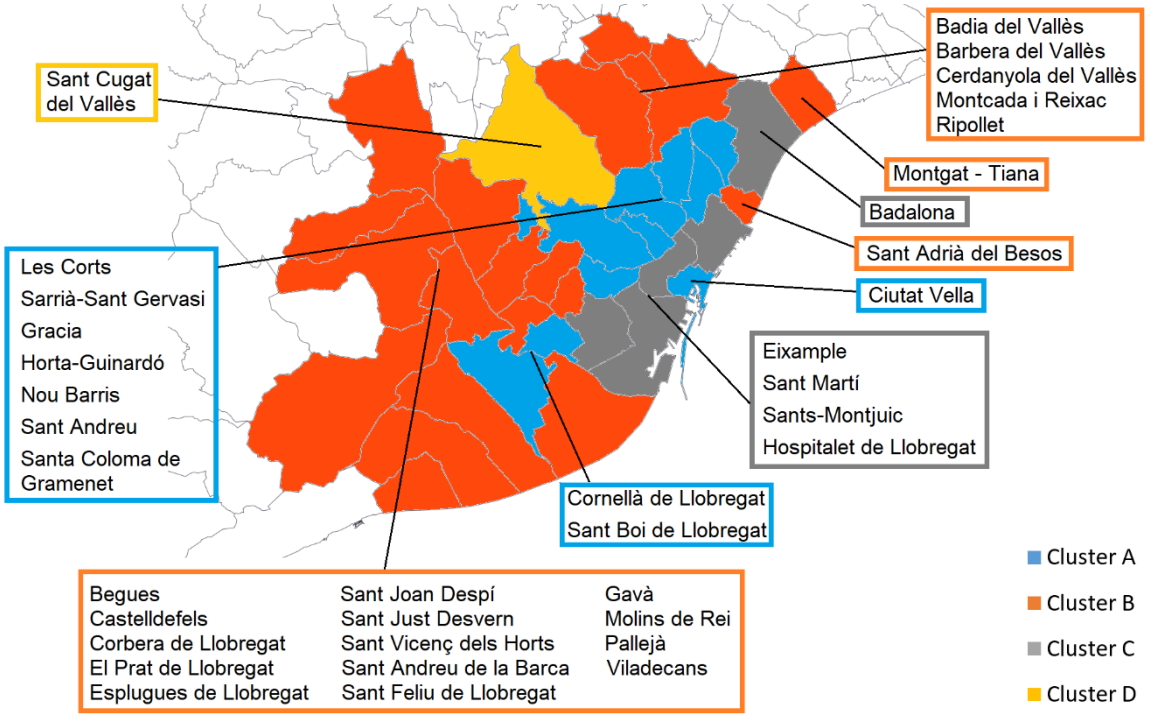

Figure 38. Assignation map of the clustering k-means algorithm by 4 clusters.

For the analysis and characterisation of each cluster, the average values of the areas forming the clusters are shown (not the sum of all the zones of the cluster). The final aim of the analysis is to characterise the zones by clusters, getting a definition that describes each one as a group of zones.

The 4 clusters have very different daily trip volumes that clearly differentiate them, as it can be seen in Figure 39. Complementing with the Figure 38, it can be seen how clusters A and C, which include the Barcelona city districts, are those with higher trip volumes compared to others.

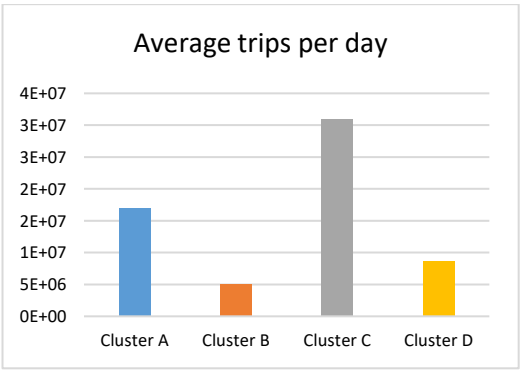

Figure 39. Average trips per day for each cluster.

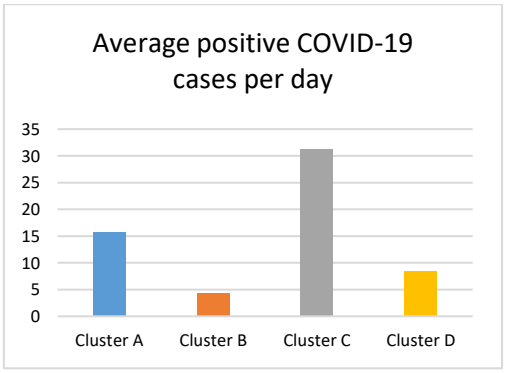

Figure 40. Average COVID-19 cases per day for each cluster.

In Figure 40, it can be seen how the order of the clusters by number of COVID-19 cases equals the order of the clusters by number of trips. Therefore, in the same way the clusters covering the city areas have the highest number of COVID-19 infected cases. This is influenced by the total population residing in the cluster areas, which follows the same pattern according to Figure 41. Consequently, it is expected that the more population there is, the more mobility and COVID-19 cases there will be in absolute numbers. On the other hand, one would also think that the higher the population density is, the higher the number of COVID-19 cases there will be. However, according to Figure 42, Sant Cugat del Vallés has a higher number of infections and lower population density than Cluster B. It can be deduced that contagion does not only depend on the mobility and the number of people grouped in the same area, but also on the idiosyncrasies of each area.

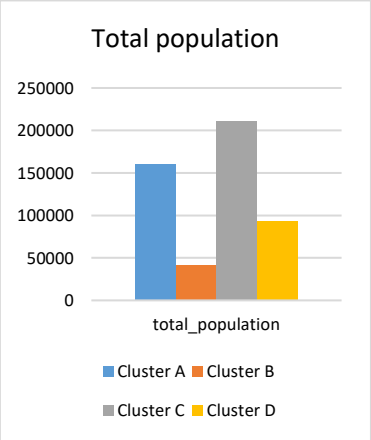

Figure 41. Total population for each cluster.

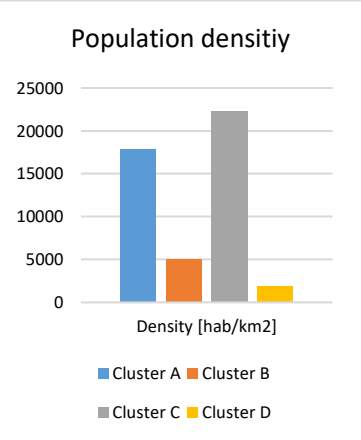

Figure 42. Population density for each cluster.

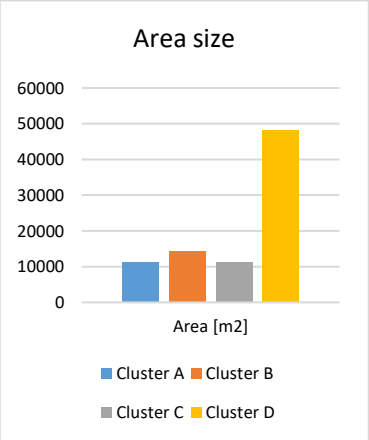

Figure 43. Average area size for each cluster.

It is worth mentioning that although cluster D is formed by only Sant Cugat del Vallès area, Figure 43 shows that it is a large area compared to the average size of the other clusters. In fact, it is not the largest area, however, the average size of the other clusters is reduced by other smaller areas inside the same cluster.

Before proceeding with the analysis, it is important to check the distance between the values of the zones that form each cluster. Cluster D will be ignored since it only comprises one specific zone.

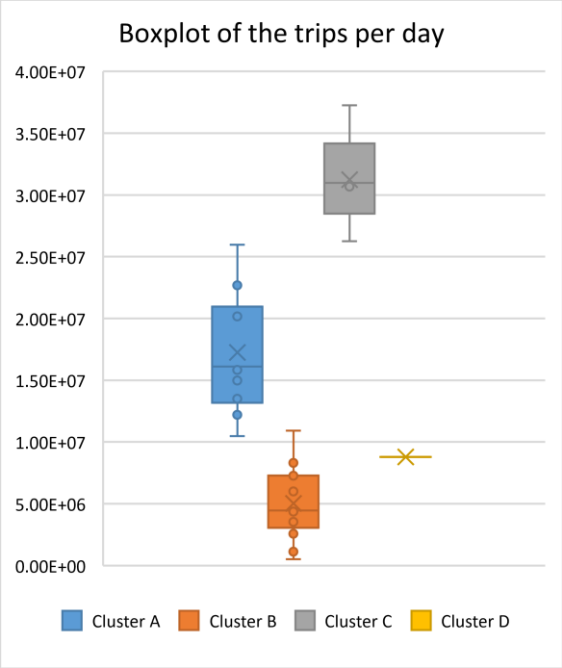

Figure 44. Boxplot of the trips per day for each cluster.

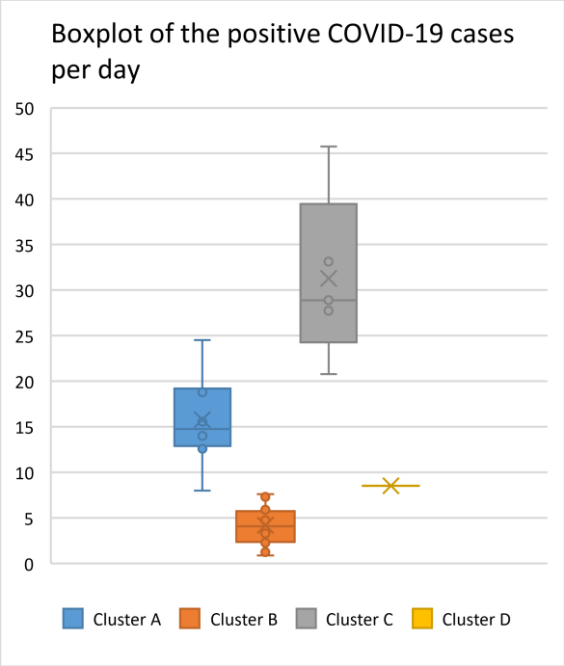

Figure 45. Boxplot of the COVID-19 cases per day for each cluster.

Considering Figures 44 and 45, it can be said that the data of all clusters are sparsely dispersed with the exception of the number of COVID-19 cases in cluster C (its interquartile is large). This was very predictable, since the clustering algorithm also proposed a grouping with only Sants-Montjuic (an area included in this cluster). The values in general are not symmetrically distributed, as it can be seen how the mean is displaced from the centre in all cases. The absence of outliers for all clusters are good news. These boxplots give an idea of the distribution of the total values but not of the values for each combination of features. It is interesting to know them, but we should not expect them to be equal. They have only been checked to ensure that no value is too far away from its cluster (outlier).

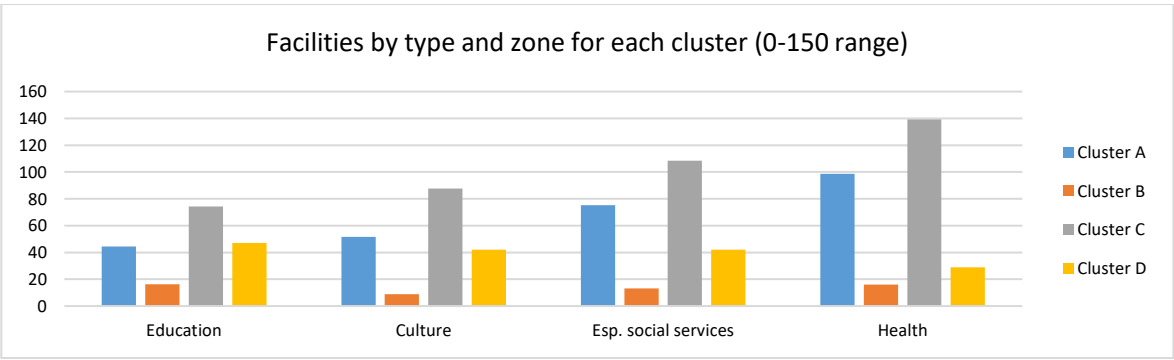

Figure 46. Facilities by type and zone for each cluster (0-150 range).

In relation to facilities, as noted in previous sections, there are quite strong correlations between the number of educational, cultural, and social service facilities with the amount of

mobility and COVID-19 positive cases. Firstly, it is linked to the total population and population density, since the more people that live in an area, the more common it is to have more facilities. However, in turn, these areas are common points of contagion and mobility hotspots that attract the population to carry out the management that corresponds to the facility. According to the average number of facilities in each cluster represented by Figure 46, the same pattern as in Figures 39 and 40 (average number of trips per day and average number of COVID-19 positive cases) is maintained.

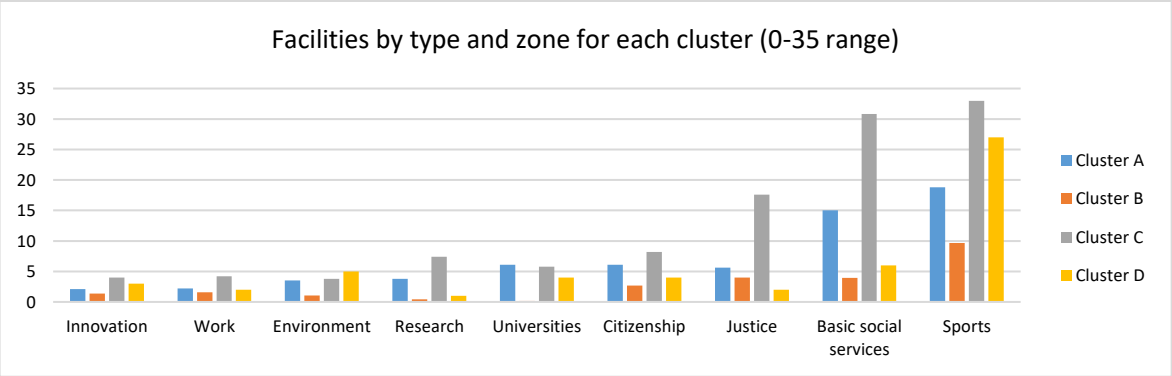

Figure 47. Facilities by type and zone for each cluster (0-35 range).

Figure 47 shows how the type of facilities with fewer number, do not follow the pattern of Figures 39 and 40. Looking at Figures 46 and 47, the correlations obtained in previous sections are verified, where only the installations in Figure 46 have a moderately strong correlation with the average number of trips per day and average number of COVID-19 positive cases.

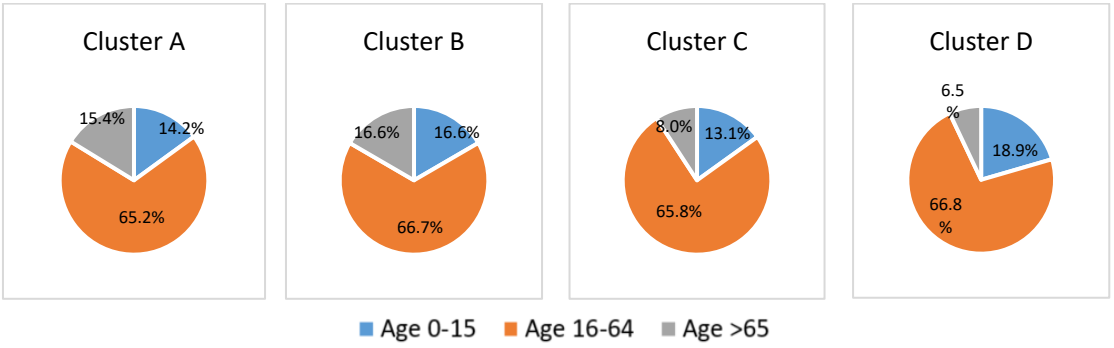

Figure 48. Average proportion range age for each cluster.

Continuing with the demographic characterisation of the territory, different proportions of mean age ranges can be observed for each cluster in Figure 48. It is interesting how the correlation drawn in previous sections between low mobility and higher proportion of population between 0 and 15 years old is fulfilled with a marked pattern among the clusters. Zones B and D with much lower mobility than the rest have up to 4% more population in this age range. In contrast, there is no pattern for the other age ranges (as stated in the correlation and causality section).

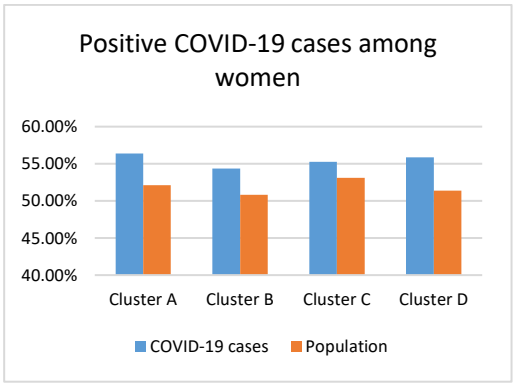

Figure 49. Average proportion of positive COVID-19 cases and population for women.

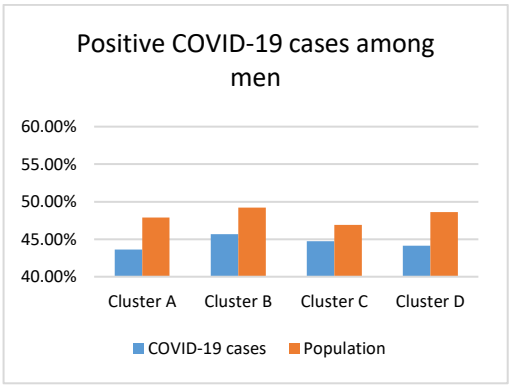

Figure 50. Average proportion of positive COVID-19 cases and population for men.

Analysing by gender, women test positive 10% more than men (55%-45% ratio). In this aspect, no pattern related to gender can be observed to differentiate between clusters. Figures 49 and 50 also show the pattern justified by the correlation found in previous sections that shows how for all clusters the proportion of infected women is higher than the proportion of the female population (and lower for men in an analogous way). As discussed, the causality is probably that women are more exposed to COVID-19 because they occupy front-line jobs, such as caregiving, as indicated by a study in the autonomous community of Madrid [58].

In order to analyse the evolution over mobility restriction periods, the average number of trips per day of the pre-pandemic period has been taken as a reference, according to the following formulation:

$$\%\ Y\ trips/day = \left(\frac{X\ period\ trips/day}{Pre\text{-}pandemic\ period\ trips/day}\right) \cdot 100 \qquad \textit{(Eq. 14)}$$

According to Figure 51, all the clusters follow the same pattern but varying on the percentage of trip reduction for each cluster.

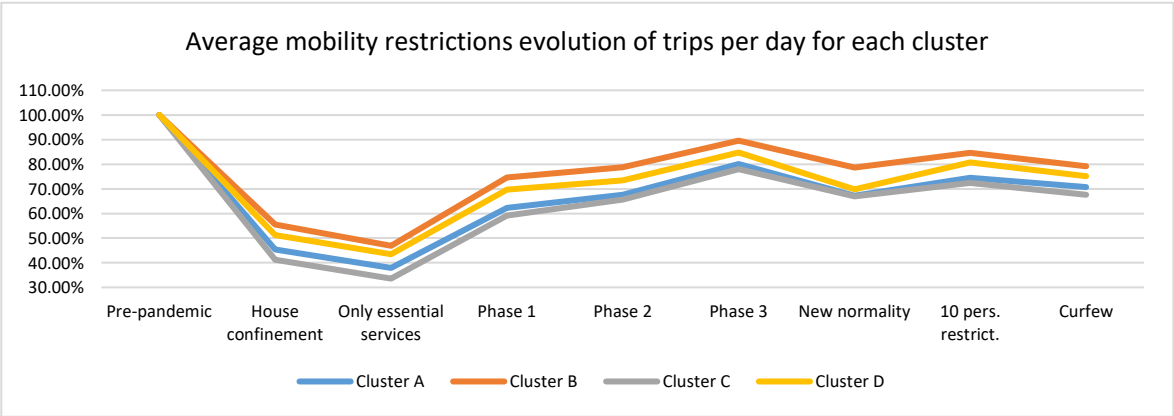

Figure 51. Average number of trips evolution through the mobility restriction periods.

Figure 51 also shows that the higher is the volume of trips in the clusters, the higher is the percentage reduction in trips through the pandemic stages. In other words, the greater number of trips, the more percentage of its reduction. Figure 51 shows that mobility restrictions have been respected in a similar way in all zones. If any zone presented an evolutionary pattern very different from the rest, it would have been included in a different cluster on its own, as in the case of Sant Cugat, which, as it can be seen, shows some discrepancy with the common pattern. For example, in the New normality period it has a much greater reduction of trips than the rest of the clusters.

As it can be seen in Figure 52, during confinement there was an equal pattern with different proportions of increase and subsequent decrease of cases for all clusters. However, from Phase 3 onwards there is a break in the pattern between the clusters close to the city (A and C) and those further away (B and D). This break in the pattern means that A and C continue with the increasing trend originating from Phase 2, and B and D decrease their cases to almost minimum numbers. This suggests that distancing and protective measures against COVID-19 may be almost nullifying the spread of the virus in sparsely populated and low-density territories. In highly populated areas, on the other hand, they help to mitigate the spread of the virus but are insufficient to correct the trend.

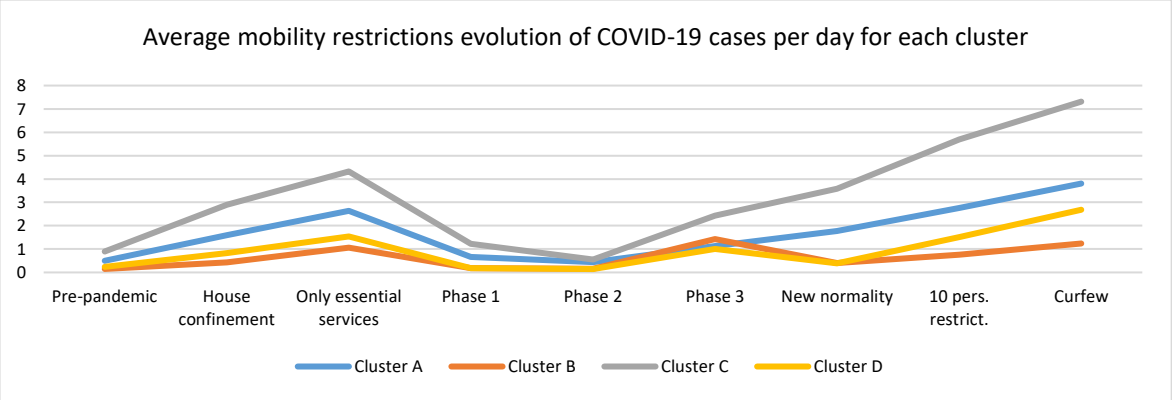

Figure 52. Mobility restrictions evolution of COVID-19 cases per day and zone for each cluster.

These plots observed in Figure 52 are very significant, as together with the correlations obtained in previous sections between positive cases with population number and density, demonstrate how enforcing more mobility restrictions on the city zones (clusters A and C), such as perimeter confinements, could improve the mitigation of the spread of the pandemic by restricting according to a choice of grouped zones based on updated data. Subsequent to the New normality, after two months of free movement through the territory, the increasing pattern of infections for the 4 clusters has been recovered.

In Figure 53, where it is analysed the evolution of trips according to their distance, all clusters follow similar patterns in relation to slope, but the y-intercept varies. During lockdown, trips

from 0.5 km to 2 km in the clusters of the city (A and C) were significantly less in percentage terms. Considering that these are areas with a higher density of stores, people bought in the nearest shops. In contrast, in the municipalities of the first ring, it is quite common to live in an area with no commerce within 500 metres. This same pattern of behaviour can also be observed for journeys of 2 km to 5 km in an even more accentuated way. In districts A and C (which comprise the city districts) there is up to almost twice the percentage reduction in trips as in the surrounding clusters.

Figure 53. Mobility restrictions evolution of the trips by its distance per zone for each cluster.

Regarding trips of 5 km to 10 km, there is a change in the pattern, as it can be observed that in all clusters there is a similar percentage decrease in mobility. However, there are still higher reduction numbers in the clusters that cover the city.

Finally, the trips from 10 km to 50 km during the lockdown are reduced to almost equal percentage numbers for all clusters. However, as the 3 phases of the lockdown progress until the New normality was reached, there was a substantially lower recovery in cluster D, which corresponds to the area of Sant Cugat del Vallès. This shows that a large part of the mobility of more than 10 km in the area is leisure mobility or due to skilled jobs that can be carried out

teleworking. Another trend breaker compared to the rest is cluster B in the New normality period. It keeps its volume of trips constant in contrast to the rest, which decreases by a little less than 10%. This is because, as it has been mentioned in previous sections, the cluster B areas have a much smaller number of facilities of all kinds and the population is forced to make trips of more than 10 km for necessary non-leisure related tasks.

Analysing the type of trips in each zone, Figure 54 shows that the inwards and outwards are exactly the same, so almost all trips leaving a cluster re-enter the cluster in each period. This means that there were no significant changes of address within the first ring of Barcelona (e.g. young people returning to the parental residence in the first crown), unlike what is known to have happened in second homes or family residences further away from the urban centres. It is also worth bearing in mind that cluster D (the Sant Cugat area) is once again the area with the greatest drop in the volume of long-distance trips (as inwards and outwards trips leave the municipality/district). The rest of the districts follow the expected patterns, with a reduction in trips proportional to the absolute trip volume as discussed in Figure 54. About the internal trips, it can be seen again that for the shortest distance trips, the reduction in trips is proportional to the absolute number of trips. For the most part it follows a pattern with the same changes in slope except when the new normality is reached. In this period, it can be seen how the internal trips of cluster B exceed those of cluster D.

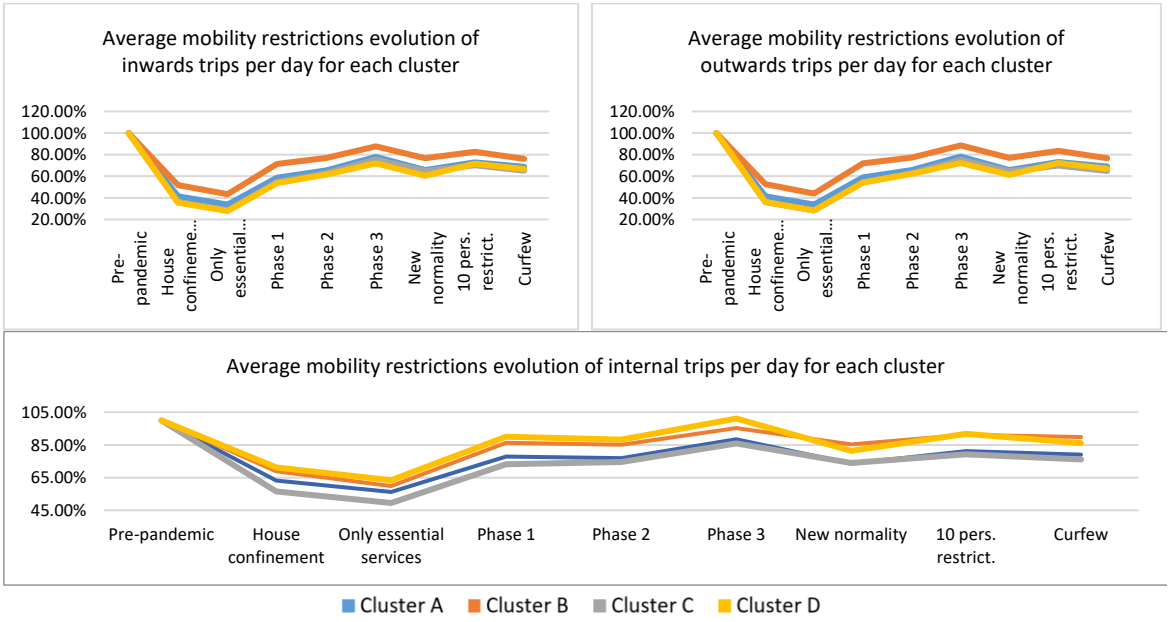

Figure 54. Mobility restrictions evolution of trips by the mobility indicator typology per zone for each cluster.

Analysing the evolution of rush hour trips in Figure 55, the morning and afternoon follow the same pattern during the different stages of mobility restrictions. During home confinement, there was a very sharp drop which recovered as the new normality was reached. During the new normality it dropped again because of summer holidays and then recovered again. The

off-peak mobility follows a very similar pattern to the one of the rush hours. On the other hand, overnight trips show a different evolutionary pattern. During the confinement, there was a more modest drop, so it can be concluded that proportionally, night-time mobility is characterised by work-related motivations that cannot be substituted by teleworking or motivations unrelated to leisure. Finally, in the curfew period there was a significant drop in travel to numbers similar to those of confinement, because of night-time mobility restrictions. In summary, except for night-time, the rest of the time slots follow the same evolutionary pattern during the periods of restriction, according to the reductions in trip volumes shown in Figure 55.

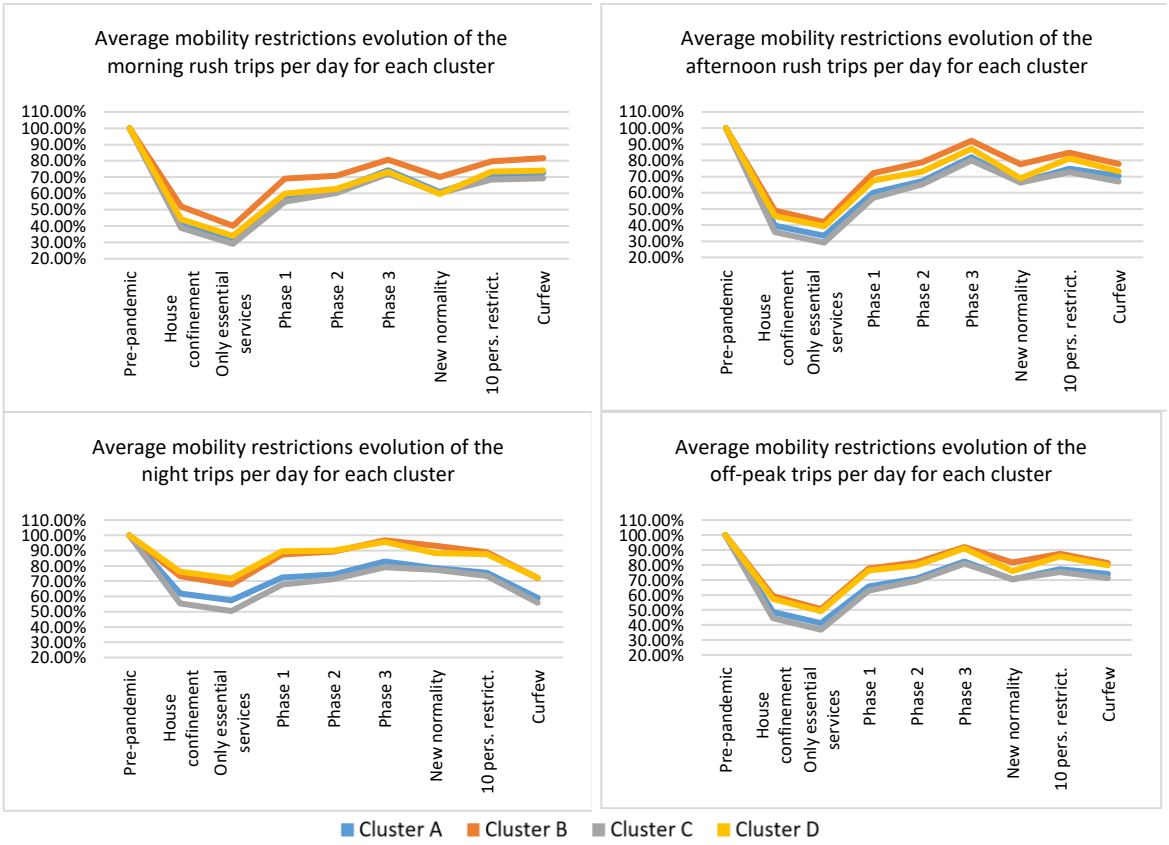

Figure 55. Mobility restrictions evolution of the trips by the daily time zones per zone for each cluster.

7. Budget

The work has lasted 26 weeks during which the six stages of the project have been developed. Figure 56 shows the stages and their sub-stages with their time duration in a Gantt chart.

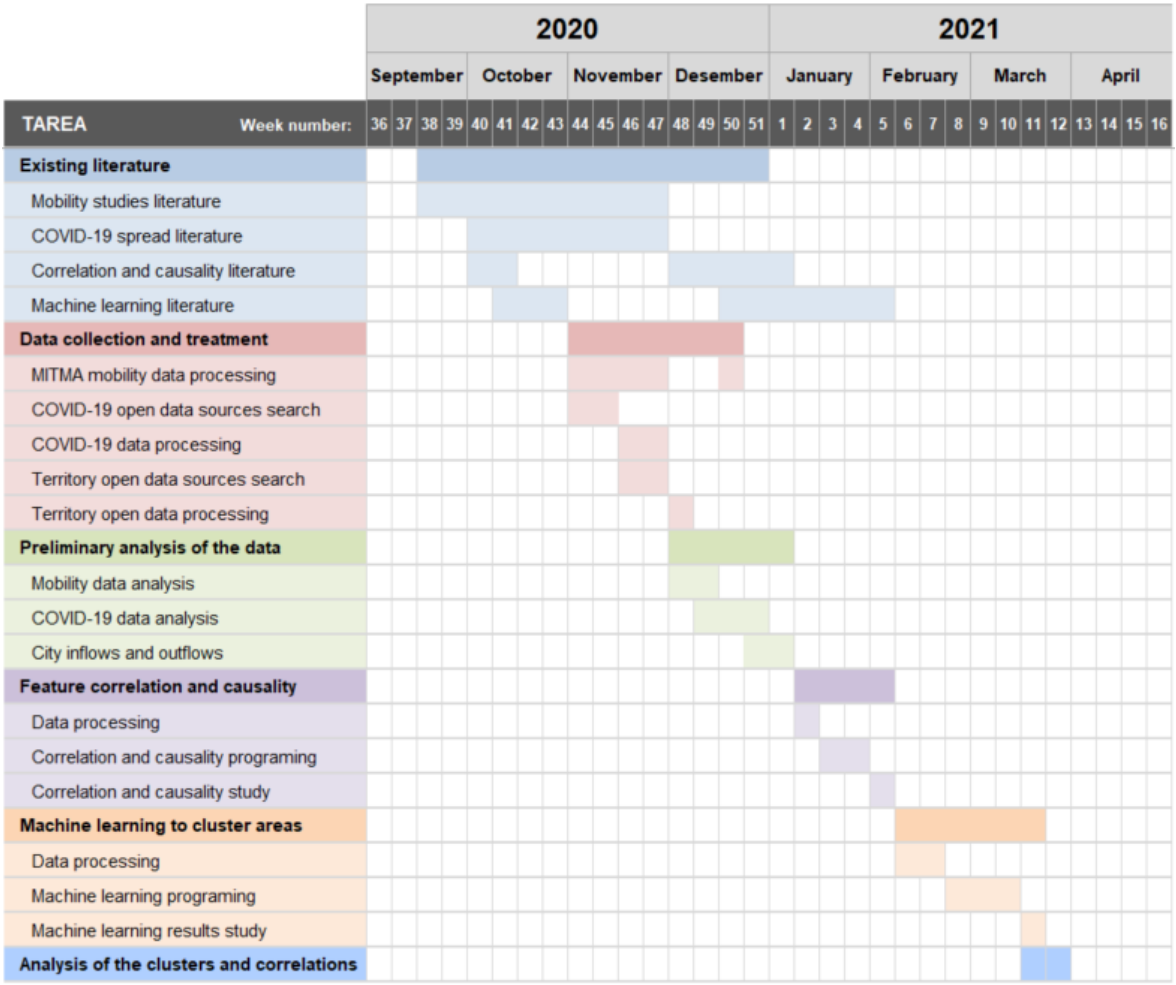

Figure 56. Time planning of the methodology structured in a Gantt chart.*

**It is considered interesting to take into consideration for the time planning that the author of the master's thesis did not have access to the mobility database until the last week of October due to external reasons.*

Since the study was carried out by a single person and its supervisor, and required the development of software using open-source resources for data processing, the breakdown of costs is relatively simple.

To establish the cost of human resources, a 25 €/hour salary for a junior engineer is taken. Considering the 26 weeks duration of the project working 5 days a week half-time (520 h), the cost of the junior engineer is 13000 €.

In addition, this study required the supervision of a senior engineer to guide the junior during

the project to help shaping the final study concept. A salary of 60 €/hour has been taken. Considering the 26 weeks duration of the project taking an average of 2.5 hours per week (65 h) of dedication between follow-up meetings and support, the cost of the senior engineer is 3900 €.

In order to account for the cost of materials, the project required the use of a computer costing 900 euros. For the calculation of its depreciation, a useful life of 4 years has been taken with a linear depreciation coefficient of 25% per year, in accordance with the limits established by the Tax Agency [71]. Therefore, if the duration of the project is 26 weeks (6.5 months), eq. 15 shows that the cost derived from the use of the computer is 122 euros.

$$6{,}5\ months \cdot \frac{1\ year}{12\ months} \cdot \frac{0.25}{1\ year} \cdot 900\ euros = 122\ euros \qquad \textit{(Eq. 15)}$$

Concept	Cost per unit	Number of units	Associated time	Total
Human resources				**16900 €**
Junior Engineer	25 €/hour	1 unit	520 hours	13000 €
Senior Engineer	60 €/hour	1 unit	65 hours	3900 €
Material costs				**122 €**
Computer	225 €/ 12 months	1 unit	6.5 months	122 €
			Total before taxes:	17022 €
			Value Added Tax (21% IVA):	**3575 €**
			TOTAL:	**20597 €**

Table 8. Summary table with the budget of the study.

Therefore, considering the cost of the human resources (16900 €), the cost of the material (122 €), and the taxes (3575 €), the total cost of the project amounts to 20597 €.

8. Environmental study

From an environmental point of view, the development of this study has not had any impact beyond the energy used by the computer during the different stages (preparation and research, software programming and development of the study) and the waste generated by the computer at the end of its useful life (although the computer has not been used exclusively for this study).

However, apart from these negative components, it should be considered that the study can potentially contribute to generate changes in the mobility of the territory in a more sustainable direction. Figure 57 shows that transport sector has been for a long time one of the main actors in CO_2 emissions in the metropolitan area of Barcelona. Concerns about climate change and high pollution in urban areas have led to mobility restrictions measures for the most polluting vehicles with combustion engines. However, this is not the only way to mitigate the problem.

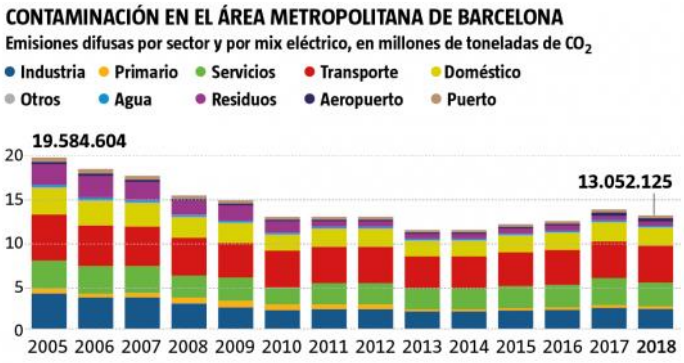

Figure 57. Temporal and sectoral evolution of pollution in the metropolitan area of Barcelona. Source: [72]

By understanding and characterising the mobility patterns of a territory it is possible to move around better. The use of this knowledge can lead to improvements such as freight companies delivering during off-peak periods of the day to reduce high fuel consumption caused by traffic jams; modifying the hourly frequency or the number/size of wagons to reduce capacity depending on the stage of mobility restriction to save energy; and so forth.

On the other hand, the data processing and the use of machine learning tools can contribute to smart mobility projects that bring the city of Barcelona closer to a more sustainable mobility.

Although it is a study without direct impact, it will be helpful for the institution of the Metropolitan Area of Barcelona to gain more knowledge about the territory, as well as for people or entities with projects related to smart mobility that can find in this master's thesis some idea or concept that will help them to the extent possible.

Conclusions

Starting from different official databases, patterns have been searched and the mobility and spread of COVID-19 have been characterized in 44 areas between municipalities and districts from the first crown of the metropolitan area of Barcelona. The main aim has been achieved despite the existing difficulty in finding very large and complete open databases, and despite the chosen zoning that implies joining different databases for the same types of data.

First of all, the temporal evolution of mobility and COVID-19 infections from February 14 to November 30 was observed. This brief analysis has allowed to observe the changes through the different stages of mobility restrictions, showing to what extent they have reduced mobility and mitigated the spread of the virus. Subsequently using correlations and causalities analysis, the first interesting conclusions have been drawn.

There is a strong medium correlation throughout the territory between having more infections and a higher percentage of female population. This observation, together with the higher proportion of positive cases in women, leads to the conclusion that women get more infected than men in the study area. The causality is found in a study carried out in a population with similar characteristics (Madrid) where it is concluded that women get more infected than men because they occupy more front-line jobs, such as caregiving.

An inverse correlation was found between the percentage of the population aged 0 to 15 years and the number of infections and the volume of mobility of the areas. It was concluded that the areas of the first crown with a lower flow of movement and concentration of people are preferred by the population to raise children these ages.

Regarding the number of facilities, a strong direct correlation has been found with mobility and the number of COVID-19 positive cases. Especially with health, education, and social services facilities. The causality lies in the fact that these facilities are meeting points between people and are potential common areas of contagion. However, this conclusion conflicts with the controversial message that has been launched from the institutions, which states that educational centres are safe environments and are not potential points of contagion.

In order to characterize the territory through direct observation of the behaviour of the areas, machine learning techniques have been used to group them. The technique used generates clusters of zones by direct resemblance of the numbers of the features of each zone, measuring this similarity computing each point's squared distance from the corresponding cluster centre and adding these distances for all points in the data set. It is interesting to perform this clustering because it greatly simplifies visual pattern recognition. In the relevant section, the zones have been extensively characterized according to all the available features,

observing that there are marked patterns in relation to mobility and COVID-19 that clearly differentiate the 4 chosen clusters. Therefore, there are major differences between the zones of the first crown of the metropolitan area of Barcelona. They can be grouped according to:

- Urban areas with moderate number of movements and COVID-19 positive cases.
- Semi-urban areas of small area and few facilities with low number of movements and COVID-19 positive cases.
- Urban area with high number of movements and COVID-19 positive cases.
- Semi-urban area of large well-equipped area with medium number of movements and COVID-19 positive cases.

Observing the results obtained as a whole, the initial motivation from which this master's thesis started is satisfied. It has been possible to develop a study using big data on mobility, COVID-19, and the territory of the first crown of Barcelona not exploited by the corresponding public administration. This is an important study since the Metropolitan Area of Barcelona (AMB) is interested in the results of this analysis.

As a continuation and future work, it would be interesting to extend the study with the data that MITMA continues to share to this day, since the software developed for its processing is generic for any date. In this way, including the new files in the local database, the analysis could be extended to the relevant date using the software developed for the filtering of the study areas (or easily extended to more areas) and the processing of trip characteristics.

Acknowledgements

In this section I would like to thank a few people for their help during the work.

First, I would like to thank the director of this master's thesis, Imma Ribas Vila, for her support in guiding me during these months of work and for her help in shaping this study.

I would like to thank my family for their unconditional support, especially my sister Maria for her constant feedback on the study.

Finally, I would like to thank my colleagues at CARNET Barcelona for their support throughout these months, especially Laia Pagès Giralt for offering me the opportunity to carry out this interesting master's thesis.

Bibliography

Examples of books, articles, catalogues, computer material and material obtained online:

[1] A. Aloi *et al.*, “Effects of the COVID-19 lockdown on urban mobility: Empirical evidence from the city of Santander (Spain),” *Sustain.*, vol. 12, no. 9, 2020, doi: 10.3390/su12093870.

[2] T. Alamo, D. G. Reina, M. Mammarella, and A. Abella, “COVID-19: Open-data resources for monitoring, modeling, and forecasting the epidemic,” *Electron.*, vol. 9, no. 5, pp. 1–30, 2020, doi: 10.3390/electronics9050827.

[3] “The Best Global Responses to COVID-19 Pandemic | Time.” https://time.com/5851633/best-global-responses-covid-19/ (accessed Feb. 01, 2021).

[4] CNMC, “Publicaciones periódicas - CNMC.” http://data.cnmc.es/datagraph/jsp/graph/grafico-cuota-mercado.jsp (accessed Jan. 31, 2021).

[5] “Registre de casos de COVID-19 realitzats a Catalunya. Segregació per sexe i àrea bàsica de salut (ABS) | Dades obertes de Catalunya.” https://analisi.transparenciacatalunya.cat/Salut/Registre-de-casos-de-COVID-19-realitzats-a-Catalun/xuwf-dxjd (accessed Apr. 14, 2021).

[6] “Pattern Recognition and Machine Learning | Christopher Bishop | Springer.” https://www.springer.com/gp/book/9780387310732 (accessed Apr. 18, 2021).

[7] “SciPy and NumPy [Book].” https://www.oreilly.com/library/view/scipy-and-numpy/9781449361600/ (accessed Apr. 18, 2021).

[8] C. Santamaria *et al.*, “Measuring the impact of COVID-19 confinement measures on human mobility using mobile positioning data. A European regional analysis,” *Saf. Sci.*, vol. 132, no. July, p. 104925, 2020, doi: 10.1016/j.ssci.2020.104925.

[9] B. Jenny *et al.*, “Design principles for origin-destination flow maps,” *Cartogr. Geogr. Inf. Sci.*, vol. 45, no. 1, pp. 62–75, 2018, doi: 10.1080/15230406.2016.1262280.

[10] MITMA, “Informe metodológico: Análisis de la movilidad en España con tecnología Big Data durante el estado de alarma para la gestión de la crisis del COVID-19 Informe metodológico,” 2020.

[11] “Autoridad del Transporte Metropolitano - Wikipedia, la enciclopedia libre.” https://es.wikipedia.org/wiki/Autoridad_del_Transporte_Metropolitano (accessed Feb. 24, 2021).

[12] “Idescat. Producto interior bruto territorial. Valor afegit brut. Base 2000. Serveis per branques d’activitat. Comarques i Aran.” http://www.idescat.cat/pub/?id=pibc&n=4253&lang=es&by=com (accessed Mar. 24, 2021).

[13] “Idescat. El municipio en cifras.” https://www.idescat.cat/emex/?lang=es (accessed Mar. 13, 2021).

[14] “Idescat. Producto interior bruto territorial. Valor afegit brut. Base 2000. Per sectors. Comarques i Aran.” http://www.idescat.cat/pub/?id=pibc&n=4251&lang=es&by=com (accessed Mar. 25, 2021).

[15] “Idescat. Producto interior bruto territorial. Valor afegit brut. Base 2000. Indústria per branques d’activitat. Comarques i Aran.” http://www.idescat.cat/pub/?id=pibc&n=4252&lang=es&by=com (accessed Mar. 25, 2021).

[16] “Idescat. Censo de población y viviendas. Habitatges en edificis destinats principalment a habitatge. Per tipus. Comarques i Aran.” https://www.idescat.cat/pub/?id=censph&n=30&lang=es&by=com (accessed Mar. 25, 2021).

[17] “Idescat. Producte interior brut territorial. Valor afegit brut dels serveis. Per branques d’activitat. Municipis.” https://www.idescat.cat/pub/?id=pibc&n=8281&by=mun (accessed Mar. 25, 2021).

[18] “Idescat. Producto interior bruto territorial. Valor afegit brut de la indústria. Per branques d’activitat. Municipis.” https://www.idescat.cat/pub/?id=pibc&n=8279&lang=es&by=mun (accessed Mar. 25, 2021).

[19] “Idescat. Censo de población y viviendas. Habitatges principals segons el nombre d’habitatges de l’edifici. Municipis.” https://www.idescat.cat/pub/?id=censph&n=308&lang=es&by=mun (accessed Mar. 25, 2021).

[20] “COVID-19 Singapore Dashboard | UCA.” https://againstcovid19.com/singapore/dashboard (accessed Feb. 15, 2021).

[21] E. Steiger, T. Mußgnug, and L. E. Kroll, “Causal analysis of COVID-19 observational data in German districts reveals effects of mobility, awareness, and temperature,” *medRxiv*, p. 2020.07.15.20154476, 2020, [Online]. Available: https://doi.org/10.1101/2020.07.15.20154476.

[22] Gobierno de España, “Iniciativa de datos abiertos del Gobierno de España.” https://datos.gob.es/ (accessed Feb. 15, 2021).

[23] G. de Catalunya, “Dades obertes de Catalunya.” https://analisi.transparenciacatalunya.cat/en/ (accessed Feb. 15, 2021).

[24] A. de Barcelona, “Open Data BCN | Ajuntament de Barcelona’s open data service.” https://opendata-ajuntament.barcelona.cat/en (accessed Feb. 15, 2021).

[25] “WHO Western Pacific | World Health Organization.” https://www.who.int/westernpacific/emergencies/covid-19 (accessed Feb. 15, 2021).

[26] “COVID-19 Map - Johns Hopkins Coronavirus Resource Center.”

https://coronavirus.jhu.edu/map.html (accessed Feb. 15, 2021).

[27] “Coronavirus research at the Blavatnik School | Blavatnik School of Government.” https://www.bsg.ox.ac.uk/news/coronavirus-research-blavatnik-school (accessed Feb. 15, 2021).

[28] “Welcome | European Union Open Data Portal.” https://data.europa.eu/euodp/en/data (accessed Feb. 15, 2021).

[29] “COVID-19 Data Portal - accelerating scientific research through data.” https://www.covid19dataportal.org/ (accessed Feb. 15, 2021).

[30] “JRC - Joint Research Centre | Knowledge for policy.” https://knowledge4policy.ec.europa.eu/organisation/jrc-joint-research-centre_en (accessed Feb. 15, 2021).

[31] “Coronavirus (COVID-19) - Google Noticias.” https://news.google.com/covid19/map?hl=es (accessed Feb. 15, 2021).

[32] “Dataset Search.” https://datasetsearch.research.google.com/ (accessed Feb. 15, 2021).

[33] “COVID-19 – MIDAS.” https://midasnetwork.us/covid-19/ (accessed Feb. 15, 2021).

[34] “COVID-19 Data Hub • COVID-19 Data Hub.” https://covid19datahub.io/ (accessed Feb. 15, 2021).

[35] R. Duque-Calvache, J. M. Torrado, and Á. Mesa-Pedrazas, “Lockdown and adaptation: residential mobility in Spain during the COVID-19 crisis,” *Eur. Soc.*, vol. 0, no. 0, pp. 1–18, 2020, doi: 10.1080/14616696.2020.1836386.

[36] “Idescat. El municipio en cifras. Barcelona.” https://www.idescat.cat/emex/?id=080193&lang=es#h2 (accessed Mar. 27, 2021).

[37] “La COVID-19 afecta más a las zonas más pobres de la ciudad de Barcelona - Noticias - IMIM Institut Hospital del Mar d’Investigacions Mèdiques.” https://www.imim.es/noticias/773/la-covid-19-afecta-mas-a-las-zonas-mas-pobres-de-la-ciudad-de-barcelona (accessed Apr. 14, 2021).

[38] M. DE La Presidencia and R. Y. Con Las Cortes Memoria Democrática, “I. DISPOSICIONES GENERALES MINISTERIO DE LA PRESIDENCIA, RELACIONES CON LAS CORTES Y MEMORIA DEMOCRÁTICA,” 2020. Accessed: Apr. 10, 2021. [Online]. Available: https://www.boe.es.

[39] “BOE.es - BOE-A-2020-4911 Orden SND/399/2020, de 9 de mayo, para la flexibilización de determinadas restricciones de ámbito nacional, establecidas tras la declaración del estado de alarma en aplicación de la fase 1 del Plan para la transición hacia una nueva normalidad.” https://www.boe.es/eli/es/o/2020/05/09/snd399 (accessed Apr. 10, 2021).

[40] “BOE.es - BOE-A-2020-5088 Orden SND/414/2020, de 16 de mayo, para la flexibilización de determinadas restricciones de ámbito nacional establecidas tras la

declaración del estado de alarma en aplicación de la fase 2 del Plan para la transición hacia una nueva normalidad." https://boe.es/diario_boe/txt.php?id=BOE-A-2020-5088 (accessed Apr. 10, 2021).

[41] "BOE.es - BOE-A-2020-5469 Orden SND/458/2020, de 30 de mayo, para la flexibilización de determinadas restricciones de ámbito nacional establecidas tras la declaración del estado de alarma en aplicación de la fase 3 del Plan para la transición hacia una nueva normalidad." https://boe.es/diario_boe/txt.php?id=BOE-A-2020-5469 (accessed Apr. 10, 2021).

[42] "Guía de la 'nueva normalidad', comunidad por comunidad." https://www.eldiario.es/sociedad/nueva-normalidad-comunidad_1_6067038.html (accessed Apr. 10, 2021).

[43] "Coronavirus: Cataluña prohíbe las reuniones sociales de más de 10 personas y Murcia las limita a seis | Cataluña | EL PAÍS." https://elpais.com/espana/catalunya/2020-08-24/cataluna-prohibe-las-reuniones-sociales-de-mas-de-10-personas.html (accessed Apr. 10, 2021).

[44] "BOE.es - BOE-A-2020-12898 Real Decreto 926/2020, de 25 de octubre, por el que se declara el estado de alarma para contener la propagación de infecciones causadas por el SARS-CoV-2." https://boe.es/diario_boe/txt.php?id=BOE-A-2020-12898 (accessed Apr. 10, 2021).

[45] "Los usuarios de las rondas de Barcelona son los que sufren más atascos." https://www.elperiodico.com/es/sociedad/20191120/rondas-barcelona-atascos-racc-7743562 (accessed Apr. 10, 2021).

[46] "Dades COVID Documentació." https://dadescovid.cat/documentacio (accessed Apr. 10, 2021).

[47] "Equipaments de Catalunya | Dades obertes de Catalunya." https://analisi.transparenciacatalunya.cat/Urbanisme-infraestructures/Equipaments-de-Catalunya/8gmd-gz7i (accessed Apr. 11, 2021).

[48] "Instal·lacions esportives de la ciutat de Barcelona - Conjunts de dades - Open Data Barcelona." https://opendata-ajuntament.barcelona.cat/data/ca/dataset/esports-instal-lacions-esportives (accessed Apr. 11, 2021).

[49] "Informació de les estacions del nou Bicing de la ciutat de Barcelona - Conjunts de dades - Open Data Barcelona." https://opendata-ajuntament.barcelona.cat/data/ca/dataset/informacio-estacions-bicing (accessed Apr. 11, 2021).

[50] "Residències per a la gent gran de la ciutat de Barcelona - Conjunts de dades - Open Data Barcelona." https://opendata-ajuntament.barcelona.cat/data/ca/dataset/serveissocials-residenciesgentgran (accessed Apr. 11, 2021).

[51] "Autoritzacions de terrasses a l'espai d'ús públic de la ciutat de Barcelona - Conjunts de dades - Open Data Barcelona." https://opendata-ajuntament.barcelona.cat/data/ca/dataset/terrasses-comercos-vigents (accessed Apr.

11, 2021).

[52] “Cercador d’equipaments. gencat.cat.” https://web.gencat.cat/ca/equipaments (accessed Apr. 11, 2021).

[53] “Població de Catalunya per municipi, rang d’edat i sexe | Dades obertes de Catalunya.” https://analisi.transparenciacatalunya.cat/en/Demografia/Poblaci-de-Catalunya-per-municipi-rang-d-edat-i-se/b4rr-d25b (accessed Apr. 11, 2021).

[54] “Observatori de districtes.” https://www.bcn.cat/estadistica/castella/documents/districtes/index.htm (accessed Apr. 11, 2021).

[55] “Observatori de districtes.” https://www.bcn.cat/estadistica/castella/documents/districtes/index.htm (accessed Mar. 13, 2021).

[56] “Correlation Coefficients.” https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm (accessed Apr. 11, 2021).

[57] H. Akoglu, “User’s guide to correlation coefficients,” doi: 10.1016/j.tjem.2018.08.001.

[58] D. General de Estadística, “Impacto socioeconómico de la COVID-19 sobre las mujeres en la Comunidad de Madrid.” Accessed: Apr. 11, 2021. [Online]. Available: www.madrid.org/iestadis.

[59] “More Correlation Coeficients.” https://www.andrews.edu/~calkins/math/edrm611/edrm13.htm#POINTB (accessed Apr. 11, 2021).

[60] S. Varma, “PRELIMINARY ITEM STATISTICS USING POINT-BISERIAL CORRELATION AND P-VALUES.”

[61] “Machine Learning - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/machine-learning (accessed Feb. 25, 2021).

[62] P. Lison, “An introduction to machine learning.”

[63] M. E. Celebi and K. Aydin, *Unsupervised learning algorithms*. Springer International Publishing, 2016.

[64] F. Murtagh and P. Contreras, “Methods of Hierarchical Clustering,” 2011.

[65] U. Von Luxburg, “A tutorial on spectral clustering,” *Stat Comput*, vol. 17, pp. 395–416, 2007, doi: 10.1007/s11222-007-9033-z.

[66] M. Steinbach, G. Karypis, and V. Kumar, “A Comparison of Document Clustering Techniques,” 2000. Accessed: Mar. 17, 2021. [Online]. Available: www.yahoo.com.

[67] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” 2003.

Accessed: Apr. 12, 2021. [Online]. Available: www.elsevier.com/locate/patcog.

[68] “sklearn.metrics.silhouette_score — scikit-learn 0.24.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (accessed Apr. 12, 2021).

[69] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?,” *Pattern Recognit.*, vol. 93, pp. 95–112, 2019, doi: 10.1016/j.patcog.2019.04.014.

[70] A. Setiawan *et al.*, “Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster Related content Inflation data clustering of some cities in Indonesia IMPLEMENTATION OF K-MEANS CLUSTERING METHOD FOR ELECTRONIC LEARNING MODEL Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster,” doi: 10.1088/1757-899X/336/1/012017.

[71] “Tabla de coeficientes de amortización lineal. - Agencia Tributaria.” https://www.agenciatributaria.es/AEAT.internet/Inicio/_Segmentos_/Empresas_y_profesionales/Empresas/Impuesto_sobre_Sociedades/Periodos_impositivos_a_partir_de_1_1_2015/Base_imponible/Amortizacion/Tabla_de_coeficientes_de_amortizacion_lineal_.shtml (accessed Apr. 16, 2021).

[72] “El área metropolitana cuadruplica en contaminación a Barcelona.” https://www.lavanguardia.com/local/barcelona/20210405/6627365/area-metropolitana-cuadruplica-contaminacion-barcelona.html (accessed Apr. 17, 2021).

[73] “Ordenació territorial per sectors sanitaris – Consorci Sanitari de Barcelona.” https://www.csb.cat/coneix-el-csb/ordenacio-per-districtes/ (accessed Apr. 14, 2021).

ANNEX

In this annex you can find the software developed in Python language for data processing, computation of correlation techniques and implementation of the machine learning algorithm for clustering.

Several libraries are imported and used throughout the processing:

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O
import matplotlib as mpl
import matplotlib.pyplot as plt
```

1. MOBILITY BIG DATA IMPORTING AND FILTERING

Several functions used for filtering the MITMA database are defined. The function import_from_MITMA_dataset() imports the database files one by one according to their name and local location. It is also given other parameters such as the code of the districts to be filtered from the rest so that within this function filter_by_districts() can be called to perform this filtering:

```
################################################################################################
###########                                USED FUNCTIONS                                ###########
################################################################################################

def filter_by_districts(global_dataset,districts,data_type):

    check_origin_district=global_dataset['valid'].copy()
    check_destination_district=global_dataset['valid'].copy()
    if data_type==1:
        for postal_code in districts['postal_code']:
            check_origin_district_aux = global_dataset['origen'] == postal_code
            check_origin_district=check_origin_district | check_origin_district_aux
        for postal_code in districts['postal_code']:
            check_destination_district_aux = global_dataset['destino'] == postal_code
            check_destination_district=check_destination_district                          |
check_destination_district_aux
        check_district = check_origin_district & check_destination_district
    elif data_type==2:
        for postal_code in districts['postal_code']:
            check_district = global_dataset['distrito'] == postal_code

    global_dataset['valid'] = global_dataset['valid'] | check_district

    global_dataset=global_dataset.drop(global_dataset[global_dataset.valid==False].index)

    return global_dataset



def    import_from_MITMA_dataset(global_dataset,    file_name,    file_directory,    districts,
```

```
data_type):

    print('Importing file ' + file_name + '.txt' + ' ...')
    file_ubi = file_directory+file_name+'.txt'

    if data_type==1:
        data2= pd.read_csv(file_ubi, sep='|',header=0, decimal=',', converters={"fecha": str,
"origen": str, "destino": str, "periodo": str, "distancia": str, "viajes_km": str, "viajes":
str})
        del data2['edad']

    elif data_type==2:
        data2= pd.read_csv(file_ubi, sep='|',header=0, converters={'distrito': str})

    data2['valid']=False
    print('Filtering file ' + file_name + '.txt' + ' ...')
    data2 = filter_by_districts(data2,districts,data_type)
    print('File ' + file_name + '.txt' + ' filtered successfully')

    if global_dataset.size == 0:
        rslt=data2
    else:
        rslt=global_dataset.append(data2)
    return rslt



def save_global_dataset(global_dataset,file):
    global_dataset.to_csv(file,sep='|',float_format='%.3f',index=False)
    print('>>>> global_dataset was saved <<<<')


def open_global_dataset(global_dataset,file):
    print(file)
    global_dataset=pd.read_csv(file,sep='|', header=0)
    print('>>>> global_dataset was imported <<<<')
    return global_dataset
```

Files with the names of the database files are imported:

```
imp_dates=pd.read_csv('../data_folder/import_dates_mitma_1_3rd_try.csv', header=None,
names=['dates'], converters={'dates': str})
bcn_districts=pd.read_csv('../data_folder/AMB_MitmaCOD.txt',header=0,
names=['OBJECTID','postal_code','DISTRITO','Shape_Length','Shape_Area'],
converters={'postal_code':str})
del bcn_districts['OBJECTID']
del bcn_districts['DISTRITO']
del bcn_districts['Shape_Length']
del bcn_districts['Shape_Area']
file_directory='../data_folder/Datos_mitma/maestra_1_mitma_distritos/'
data_type=1
MITMA_dataset_1=pd.DataFrame()

for j in range(len(imp_dates)):
    file_name=imp_dates.loc[j,'dates']

MITMA_dataset_1=import_from_MITMA_dataset(MITMA_dataset_1,file_name,file_directory,bcn_distr
icts, data_type)
    print('File ' + file_name + '.txt' + ' dataset imported and filtered successfully')
```

Created databases relating the zones with different codes used are imported:

```
########################################
### OPENING THE TIME PERIODS DATASET ###
########################################
file1='../data_folder/year_time_periods.csv'
year_time_periods=pd.read_csv(file1,sep=';', header=0)
del year_time_periods['covid_restrict_descrip']

############################
### OPENING THE DATASET ###
############################
file1='../data_folder/zonification_name_zone_code_comreg.csv'
zonification_name_zone_code_com=pd.read_csv(file1,sep=';', header=0,converters={"district":
str, "zone": str, "code_tfm": str, "com_reg": str, "fst_mob_clust":str})

############################
### OPENING THE DATASET ###
############################
file1='../data_folder/zonification_name_codi_postal.csv'
zonification_name_codi_postal=pd.read_csv(file1,sep=';', header=0,converters={"district":
str, "codi_postal": str, "zone": str, "code_tfm": str, "comarca": str})
```

2. MOBILITY DATA PROCESSING

The mobility database is created and processed by replacing the origin-destination pairs by the mobility indicator (internal, inwards or outwards trips):

```
############################################
### CALCULATION TO COMPUTE TRIPS PER HOUR ###
############################################
MITMA_dataset['tripshour']='Nan'
MITMA_dataset['tripshour'].loc[(MITMA_dataset['franja']=='Night')]=round(MITMA_dataset['viaj
es']/8)
MITMA_dataset['tripshour'].loc[(MITMA_dataset['franja']=='Morning_rush')]=round(MITMA_datase
t['viajes']/3)
MITMA_dataset['tripshour'].loc[(MITMA_dataset['franja']=='Afternoon_rush')]=round(MITMA_data
set['viajes']/4)
MITMA_dataset['tripshour'].loc[(MITMA_dataset['franja']=='Off-
peak')]=round(MITMA_dataset['viajes']/9)
MITMA_dataset

############################################
### CONVERSION FROM POSTAL CODE TO CODI_TFM ###
############################################
for i in range(len(zonification_name_zone_code_com['zone'])):
    zone=zonification_name_zone_code_com.loc[i,'zone']
    print(zone)
    #print(zonification_name_zone_code_com.loc[i,'comarca'])

MITMA_dataset['origen'].loc[(MITMA_dataset['origen']==zone)]=zonification_name_zone_code_com
.loc[i,'code_tfm']

MITMA_dataset['destino'].loc[(MITMA_dataset['destino']==zone)]=zonification_name_zone_code_c
om.loc[i,'code_tfm']
MITMA_dataset=MITMA_dataset.groupby(['fecha','origen','destino','distancia','franja'],
as_index=False)['tripshour'].sum()

#####################################################
### COMPUTING INTERNAL TRIPS OF MOBILITY INDICATOR ###
#####################################################
dataset_internal=MITMA_dataset.drop(MITMA_dataset[MITMA_dataset['origen']!=MITMA_dataset['de
stino']].index)
dataset_internal=dataset_internal.groupby(['fecha','origen','destino','distancia','franja'],
as_index=False)['tripshour'].sum()
dataset_internal['zona']=dataset_internal['destino']
del dataset_internal['destino']
del dataset_internal['origen']
dataset_internal['mobility_indicator']='internal'
```

```
#####################################################
### COMPUTING INWARDS TRIPS OF MOBILITY INDICATOR ###
#####################################################
dataset_inwards=MITMA_dataset.drop(MITMA_dataset[MITMA_dataset['origen']==MITMA_dataset['des
tino']].index)
dataset_inwards['zona']=dataset_inwards['origen']
del dataset_inwards['destino']
del dataset_inwards['origen']
dataset_inwards['mobility_indicator']='inwards'

#####################################################
### COMPUTING OUTWARDS TRIPS OF MOBILITY INDICATOR ###
#####################################################
dataset_outwards=MITMA_dataset.drop(MITMA_dataset[MITMA_dataset['origen']==MITMA_dataset['de
stino']].index)
dataset_outwards['zona']=dataset_outwards['destino']
del dataset_outwards['destino']
del dataset_outwards['origen']
dataset_outwards['mobility_indicator']='outwards'

###########################################
### CONCATENATING THE MOBILITY INDICATOR ###
###########################################
MITMA_dataset_mob_ind=pd.concat([dataset_inwards, dataset_outwards,dataset_internal],
ignore_index=True)
MITMA_dataset_mob_ind=MITMA_dataset_mob_ind.groupby(['fecha','zona','mobility_indicator','di
stancia','franja'], as_index=False)['tripshour'].sum()
MITMA_dataset_mob_ind.columns=['date','zone','mobility_indicator','distance','timezone','tri
pshour']


MITMA_dataset_mob_ind['tripshour']=pd.to_numeric(MITMA_dataset_mob_ind['tripshour'],
downcast='integer')
MITMA_dataset_mob_ind['covid_restrict']='Nan'
for i in range(len(year_time_periods['covid_restrict'])):

MITMA_dataset_mob_ind['covid_restrict'].loc[MITMA_dataset_mob_ind['date']==year_time_periods
.loc[i,'date']]=year_time_periods.loc[i,'covid_restrict']
MITMA_dataset_mob_ind=MITMA_dataset_mob_ind.groupby(['zone','covid_restrict','mobility_indic
ator','distance','timezone'], as_index=False)['tripshour'].mean()

MITMA_dataset_mob_ind
```

3. MOBILITY DATA FIRST ANALYSIS

The first processing of the data for the first mobility data analysis is carried out:

```
############################
### OPENING THE DATASET ###
############################
file1='../data_folder/zonification_name_zone_code_comreg.csv'
zonification_name_zone_code_com=pd.read_csv(file1,sep=';', header=0,converters={"district":
str, "codi_postal":str, "zone": str, "code_tfm": str, "com_reg": str, "fst_mob_clust":str})
zonification_name_zone_code_com

############################
### OPENING THE DATASET ###
############################
file1='../data_folder/MITMA_dataset_minimum_rows_15-12-20.csv'
MITMA_dataset_1=pd.read_csv(file1,sep=',', header=0,converters={'fecha': str,'origen':
str,'destino': str})
del MITMA_dataset_1['Unnamed: 0']
MITMA_dataset_1=MITMA_dataset_1.groupby(['fecha','origen','destino'],
as_index=False)['viajes'].sum()
MITMA_dataset_1

MITMA_dataset_1['origin_fst_mob_clust']='Nan'
MITMA_dataset_1['destino_fst_mob_clust']='Nan'
```

```
for i in range(len(zonification_name_zone_code_com['fst_mob_clust'])):
MITMA_dataset_1['origin_fst_mob_clust'].loc[(MITMA_dataset_1['origen']==zonification_name_zo
ne_code_com.iloc[i]['zone'])]=zonification_name_zone_code_com.iloc[i]['fst_mob_clust']

MITMA_dataset_1['destino_fst_mob_clust'].loc[(MITMA_dataset_1['destino']==zonification_name_
zone_code_com.iloc[i]['zone'])]=zonification_name_zone_code_com.iloc[i]['fst_mob_clust']
MITMA_dataset_1=MITMA_dataset_1.groupby(['fecha','origin_fst_mob_clust','destino_fst_mob_clu
st'], as_index=False)['viajes'].sum()
MITMA_dataset_1


dates_prepandemic=pd.read_csv('../data_folder/dates_prepandemic.csv', header=None,
names=['dates'], converters={'dates': str})
dates_lockdown=pd.read_csv('../data_folder/dates_lockdown.csv', header=None,
names=['dates'], converters={'dates': str})
dates_postlockdown=pd.read_csv('../data_folder/dates_postlockdown.csv', header=None,
names=['dates'], converters={'dates': str})
dates_prepandemic=np.array(dates_prepandemic['dates'])
dates_lockdown=np.array(dates_lockdown['dates'])
dates_postlockdown=np.array(dates_postlockdown['dates'])


OD_prepandemic=MITMA_dataset_1.drop(MITMA_dataset_1[~MITMA_dataset_1['fecha'].isin(dates_pre
pandemic)].index)
OD_lockdown=MITMA_dataset_1.drop(MITMA_dataset_1[~MITMA_dataset_1['fecha'].isin(dates_lockdo
wn)].index)
OD_postlockdown=MITMA_dataset_1.drop(MITMA_dataset_1[~MITMA_dataset_1['fecha'].isin(dates_po
stlockdown)].index)


MITMA_dataset_2=MITMA_dataset_1.groupby(['origin_fst_mob_clust','destino_fst_mob_clust'],
as_index=False)['viajes'].mean()
OD_prepandemic=OD_prepandemic.groupby(['origin_fst_mob_clust','destino_fst_mob_clust'],
as_index=False)['viajes'].mean()
OD_lockdown=OD_lockdown.groupby(['origin_fst_mob_clust','destino_fst_mob_clust'],
as_index=False)['viajes'].mean()
OD_postlockdown=OD_postlockdown.groupby(['origin_fst_mob_clust','destino_fst_mob_clust'],
as_index=False)['viajes'].mean()
```

4. COVID-19 DATA PROCESSING

The COVID-19 database is created and processed:

```
############################
### OPENING THE DATASET ###
############################
file1='../data_folder/ABS_casos_covid_per_dia_per_sexe2.csv'
ABS_covid_dataset=pd.read_csv(file1,sep=';', header=0)

############################
### OPENING THE DATASET ###
############################
file1='../data_folder/aga_arees_gestio_assistencial.csv'
ABS_codis=pd.read_csv(file1,sep=';', header=0,
converters={"codi_abs":int,"nom_abs":str,"codi_abs_tfm":str})
ABS_codis

ABS_covid_dataset=ABS_covid_dataset.drop(ABS_covid_dataset[~ABS_covid_dataset['ABSCodi'].isi
n(ABS_codis['codi_abs'])].index)

del ABS_covid_dataset['RegioSanitariaCodi']
del ABS_covid_dataset['RegioSanitariaDescripcio']
del ABS_covid_dataset['SectorSanitariCodi']
del ABS_covid_dataset['SectorSanitariDescripcio']
del ABS_covid_dataset['ABSDescripcio']
del ABS_covid_dataset['SexeCodi']
```

```
ABS_covid_dataset['zone']='Nan'
for i in range(len(ABS_codis['codi_abs'])):
    ABS_code=ABS_codis.loc[i,'codi_abs']
    #print(ABS_code)

ABS_covid_dataset['zone'].loc[(ABS_covid_dataset['ABSCodi']==ABS_code)]=ABS_codis.loc[i,'cod
i_abs_tfm']
    #print(ABS_codis.loc[i,'codi_abs_tfm'])

del ABS_covid_dataset['ABSCodi']
ABS_covid_dataset.columns=['date', 'sex','test_description','num_cases','zone']
ABS_covid_dataset=ABS_covid_dataset.groupby(['date','zone','test_description','sex'],
as_index=False)['num_cases'].sum()
#ABS_covid_dataset['date']=ABS_covid_dataset.date.str.slice(start=6,
stop=10)+'/'+ABS_covid_dataset.date.str.slice(start=3,
stop=5)+'/'+ABS_covid_dataset.date.str.slice(start=0, stop=2)
ABS_covid_dataset=ABS_covid_dataset.sort_values(by=['date'])

ABS_covid_dataset['covid_restrict']='Nan'
for i in range(len(year_time_periods['covid_restrict'])):

ABS_covid_dataset['covid_restrict'].loc[ABS_covid_dataset['date']==year_time_periods.loc[i,'
date']]=year_time_periods.loc[i,'covid_restrict']

ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='None')]=ABS_covid_
dataset['num_cases']/29
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='HC')]=ABS_covid_da
taset['num_cases']/51
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='OE')]=ABS_covid_da
taset['num_cases']/14
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='Fase1')]=ABS_covid
_dataset['num_cases']/21
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='Fase2')]=ABS_covid
_dataset['num_cases']/11
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='Fase3')]=ABS_covid
_dataset['num_cases']/1
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='NN')]=ABS_covid_da
taset['num_cases']/66
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='MR_10p')]=ABS_covi
d_dataset['num_cases']/62
ABS_covid_dataset['num_cases'].loc[(ABS_covid_dataset['covid_restrict']=='CFW')]=ABS_covid_d
ataset['num_cases']/37

#ABS_covid_dataset['com_reg']='Nan'
#for i in range(len(zonification_name_zone_code_com['code_tfm'])):
#
ABS_covid_dataset['com_reg'].loc[(ABS_covid_dataset['zone']==zonification_name_zone_code_com
.loc[i,'code_tfm'])]=zonification_name_zone_code_com.loc[i,'com_reg']


ABS_covid_dataset=ABS_covid_dataset.groupby(['zone','covid_restrict','test_description','sex
'], as_index=False)['num_cases'].sum()
```

5. COVID-19 DATA PROCESSING

The first processing of the data for the first analysis of COVID-19 data is performed:

```
############################
### OPENING THE DATASET ###
############################
file1='../data_folder /ABS_covid_dataset.csv'
ABS_covid_dataset=pd.read_csv(file1,sep=';',
header=0,converters={"date":str,"zone":str,"test_description":str,"sex":str,"num_cases":int}
)
ABS_covid_dataset

del ABS_covid_dataset['test_description']
ABS_covid_dataset=ABS_covid_dataset.groupby(['date','zone','sex'],
as_index=False)['num_cases'].sum()
```

```
ABS_covid_dataset_sex=ABS_covid_dataset.copy()
del ABS_covid_dataset_sex['date']
del ABS_covid_dataset_sex['zone']
ABS_covid_dataset_sex=ABS_covid_dataset_sex.groupby(['sex'],
as_index=False)['num_cases'].sum()

ABS_covid_dataset_date=ABS_covid_dataset.copy()
del ABS_covid_dataset_date['zone']
del ABS_covid_dataset_date['sex']
ABS_covid_dataset_date=ABS_covid_dataset_date.groupby(['date'],
as_index=False)['num_cases'].sum()

ABS_covid_dataset_zones=ABS_covid_dataset.copy()
del ABS_covid_dataset_zones['sex']
del ABS_covid_dataset_zones['date']
ABS_covid_dataset_zones=ABS_covid_dataset_zones.groupby(['zone'],
as_index=False)['num_cases'].sum()

ABS_covid_dataset_zones_sex=ABS_covid_dataset.copy()
del ABS_covid_dataset_zones_sex['date']
ABS_covid_dataset_zones_sex=ABS_covid_dataset_zones_sex.groupby(['zone','sex'],
as_index=False)['num_cases'].sum()

ABS_covid_dataset_clusters=ABS_covid_dataset.copy()
del ABS_covid_dataset_clusters['date']
ABS_covid_dataset_clusters['zone_fst_mob_clust']='Nan'
for i in range(len(zonification_name_zone_code_com['fst_mob_clust'])):

ABS_covid_dataset_clusters['zone_fst_mob_clust'].loc[(ABS_covid_dataset_clusters['zone']==zo
nification_name_zone_code_com.iloc[i]['code_tfm'])]=zonification_name_zone_code_com.iloc[i][
'fst_mob_clust']
ABS_covid_dataset_clusters=ABS_covid_dataset_clusters.groupby(['zone_fst_mob_clust'],
as_index=False)['num_cases'].sum()
```

The data of the facilities that were finally discarded is processed:

```
##########################################
### OPENING THE GARDENS AND PARKS DATASET ###
##########################################
file1='../data_folder/C006_Parcs_i_jardins.csv'
datasets_parcs_jardins=pd.read_csv(file1,sep=';', header=0,converters={"codi_tfm": str,
"NOM_DISTRICTE": str})
datasets_parcs_jardins['cont_parcs']=1
datasets_parcs_jardins=datasets_parcs_jardins.groupby(['codi_tfm','NOM_DISTRICTE'],
as_index=False)['cont_parcs'].sum()
del datasets_parcs_jardins['NOM_DISTRICTE']
datasets_parcs_jardins=datasets_parcs_jardins.set_index(['codi_tfm'])

#####################################
### OPENING THE INSTALATIONS DATASET ###
#####################################
file1='../data_folder/ES003_Installacions_esportives.csv'
datasets_instalacions_esportives=pd.read_csv(file1,sep=';',
header=0,converters={"CODI_EQUIPAMENT": str, "codi_tfm": str, "instalacions_esportives":
str})
datasets_instalacions_esportives=datasets_instalacions_esportives.reset_index()
##########################################
### FILTRATGE PER INSTALACIONS REPETIDES ###
##########################################
a=[0]
ers=[]
for i in range(len(datasets_instalacions_esportives['CODI_EQUIPAMENT'])):
    cond=(datasets_instalacions_esportives.iloc[i]['CODI_EQUIPAMENT']) in a
    #print(str(i)+","+str(cond))
    if cond:
        ers.append(i)
    else:
        a.append(datasets_instalacions_esportives.iloc[i]['CODI_EQUIPAMENT'])
datasets_instalacions_esportives=datasets_instalacions_esportives.drop(ers)
datasets_instalacions_esportives['cont_instal']=1
datasets_instalacions_esportives=datasets_instalacions_esportives.groupby(['codi_tfm','insta
```

```
lacions_esportives'], as_index=False)['cont_instal'].sum()
datasets_instalacions_esportives

####################################
### OPENING THE GERIATRIC DATASET ###
####################################
file1='../data_folder/J004_ResidenciesGentGran.csv'
datasets_residencies_gent_gran=pd.read_csv(file1,sep=';',
header=0,converters={"CODI_EQUIPAMENT": str, "codi_tfm": str, "residencies_gent_gran": str})
##########################################
### FILTRATGE PER GERIATRIC REPETITS ####
##########################################
a=[0]
ers=[]
for i in range(len(datasets_residencies_gent_gran['CODI_EQUIPAMENT'])):
    cond=(datasets_residencies_gent_gran.iloc[i]['CODI_EQUIPAMENT']) in a
    #print(str(i)+","+str(cond))
    if cond:
        ers.append(i)
    else:
        a.append(datasets_residencies_gent_gran.iloc[i]['CODI_EQUIPAMENT'])
datasets_residencies_gent_gran=datasets_residencies_gent_gran.drop(ers)
datasets_residencies_gent_gran['cont_resi']=1
datasets_residencies_gent_gran=datasets_residencies_gent_gran.groupby(['codi_tfm','residenci
es_gent_gran'], as_index=False)['cont_resi'].sum()


################################################
### MERGING TERRITORY AND POPULATION DATASETS ###
################################################
datasets_parcs_jardins_piv=pd.pivot_table(datasets_parcs_jardins, values=['cont_parcs'],
index=['codi_tfm'])
datasets_instalacions_esportives_piv=pd.pivot_table(datasets_instalacions_esportives,
values=['cont_instal'], index=['codi_tfm'],columns=['instalacions_esportives'])
datasets_residencies_gent_gran_piv=pd.pivot_table(datasets_residencies_gent_gran,
values=['cont_resi'], index=['codi_tfm'],columns=['residencies_gent_gran'])
datasets_terrasses_autoritzacions_piv=pd.pivot_table(datasets_terrasses_autoritzacions,value
s=['seats'], index=['codi_tfm'],columns=['tipus_terrassa'])

#dataset_territori_piv2 = pd.merge(datasets_parcs_jardins,
datasets_instalacions_esportives_piv, on=['codi_tfm'])
#dataset_territori_piv2 = pd.merge(dataset_territori_piv2,
datasets_residencies_gent_gran_piv, on=['codi_tfm'])
#dataset_territori_piv2 = pd.merge(dataset_territori_piv2,
datasets_terrasses_autoritzacions_piv, on=['codi_tfm'])

#dataset_territori_piv2=dataset_territori_piv2.fillna(0)
```

6. DEMOGRAFIC AND FACILITIES DATA PROCESSING

The demographic and facilities dataset are created and processed:

```
####################################
### OPENING THE EQUIPMENT DATASET ###
####################################
file1='../data_folder/Equipaments_de_Catalunya_4.csv'
datasets_equipaments_catalunya=pd.read_csv(file1,sep=';', header=0,converters={"CPOSTAL":
str, "equipaments": str})
datasets_equipaments_catalunya['cont_equips']=1
datasets_equipaments_catalunya
################################################
### DATASET EQUIPAMENTS - FILTRATGE PER ZONES ###
################################################
datasets_equipaments_catalunya['bool_val']=False
datasets_equipaments_catalunya['codi_tfm']='Nan'
for i in range(len(zonification_name_codi_postal['codi_postal'])):
    #print(zonification_name_codi_postal.iloc[i]['code_tfm'])
    #print(zonification_name_codi_postal.loc[i,'codi_postal'])
    #print('--------------')
```

```
datasets_equipaments_catalunya['bool_val'].loc[datasets_equipaments_catalunya['CPOSTAL']==zo
nification_name_codi_postal.loc[i,'codi_postal']]=True

datasets_equipaments_catalunya['codi_tfm'].loc[datasets_equipaments_catalunya['CPOSTAL']==zo
nification_name_codi_postal.loc[i,'codi_postal']]=zonification_name_codi_postal.iloc[i]['cod
e_tfm']

datasets_equipaments_catalunya=datasets_equipaments_catalunya.drop(datasets_equipaments_cata
lunya[datasets_equipaments_catalunya['bool_val']==False].index)
#datasets_equipaments_catalunya['bool_val']
datasets_equipaments_catalunya=datasets_equipaments_catalunya.reset_index()
datasets_equipaments_catal=datasets_equipaments_catalunya
##########################################
### FILTRATGE PER EQUIPAMENTS REPETITS ####
##########################################



a=[0]
for i in range(len(datasets_equipaments_catalunya['IDEQUIPAMENT'])-1):
    #print(i)
    cond=(datasets_equipaments_catalunya.iloc[i]['IDEQUIPAMENT']) in a
    #print(cond)
    if cond:
        datasets_equipaments_catalunya=datasets_equipaments_catalunya.drop(i)
    else:
        a.append(datasets_equipaments_catalunya.iloc[i]['IDEQUIPAMENT'])
datasets_equipaments_catalunya=datasets_equipaments_catalunya.groupby(['codi_tfm','equipamen
ts'], as_index=False)['cont_equips'].sum()


#####################################
### OPENING THE POPULATION DATASET ###
#####################################
file1='../data_folder/dataset_districtes_bcn2.csv'
dataset_poblacio_bcn=pd.read_csv(file1,sep=';', header=0,converters={"codi_tfm": str})
del dataset_poblacio_bcn['name']
dataset_poblacio_bcn=dataset_poblacio_bcn.set_index(['codi_tfm'])


################################################
### MERGING TERRITORY AND POPULATION DATASETS ###
################################################
datasets_equipaments_catal_piv=pd.pivot_table(datasets_equipaments_catalunya,values=['cont_e
quips'], index=['codi_tfm'],columns=['equipaments'])
dataset_territori_piv1 = pd.merge(dataset_poblacio_bcn, datasets_equipaments_catal_piv,
on=['codi_tfm'])
dataset_territori_piv1=dataset_territori_piv1.fillna(0)

dataset_territori_piv1
```

7. CORRELATION ANALYSIS

Mobility, COVID-19 and the territory datasets are processed and merged for quantitative correlation analysis:

```
dataset_territori_piv1=dataset_territori_piv1.reset_index()
dataset_territori_piv1['zone']=dataset_territori_piv1['codi_tfm']
del dataset_territori_piv1['codi_tfm']

dataset_anlys=pd.merge(MITMA_dataset_mob_ind, ABS_covid_dataset, on=["zone",
"covid_restrict"])
dataset_anlys=dataset_anlys.groupby(['zone','covid_restrict','mobility_indicator','distance'
,'timezone','test_description','sex'], as_index=False)['tripshour','num_cases'].sum()
dataset_anlys=pd.merge(dataset_anlys, dataset_territori_piv1, on=["zone"])
```

```
del dataset_anlys['covid_restrict']
del dataset_anlys['mobility_indicator']
del dataset_anlys['distance']
del dataset_anlys['timezone']
del dataset_anlys['test_description']
del dataset_anlys['sex']
del dataset_anlys['zone']

dataset_anlys=dataset_anlys.fillna(0)

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler() #Standardize features by removing the mean and scaling to unit
variance#
datadropnorm = scaler.fit_transform(dataset_anlys.astype(float))

datadropnorm=pd.DataFrame(datadropnorm)
datadropnorm.columns=dataset_anlys.columns
datadropnorm.index=dataset_anlys.index

dataset_anlys_correlation=datadropnorm.corr(method ='pearson')
dataset_anlys_listed_correlations=pd.DataFrame()

fig, ax = plt.subplots(figsize=(25,25))
sns.heatmap(dataset_anlys_correlation, vmax=1, square=True, annot=True)


tripshour_pearson=pd.DataFrame()
tripshour_pearson['feature']=dataset_anlys_correlation.columns
tripshour_pearson['correlation']=np.array(dataset_anlys_correlation.loc['tripshour'])
tripshour_pearson=tripshour_pearson.sort_values(by=['correlation'],ascending=False)

covid_pearson=pd.DataFrame()
covid_pearson['feature']=dataset_anlys_correlation.columns
covid_pearson['correlation']=np.array(dataset_anlys_correlation.loc['num_cases'])
covid_pearson=covid_pearson.sort_values(by=['correlation'],ascending=False)
```

Mobility and COVID-19 datasets are processed and merged for qualitative correlation analysis:

```
from scipy.stats import pointbiserialr

dataset_anlys=pd.merge(MITMA_dataset_mob_ind, ABS_covid_dataset, on=["zone",
"covid_restrict"])
for num in ['15','18','21','34','40','42','43']:
    dataset_anlys=dataset_anlys.drop(dataset_anlys[dataset_anlys['zone']==num].index)

del dataset_anlys['num_cases']

#dataset_anlys['zone']=pd.to_numeric(dataset_anlys['zone'],downcast='integer')

dataset_anlys = pd.get_dummies(data=dataset_anlys, drop_first=False)

i=0
tripshour_pointbiserial=pd.DataFrame()
tripshour_pointbiserial['feature']='Nan'
tripshour_pointbiserial['correlation']='Nan'
tripshour_pointbiserial['p_value']='Nan'
for columna in dataset_anlys.columns:
    i=i+1
    tripshour_pointbiserial.loc[i,'feature']=columna
    aux=pointbiserialr(dataset_anlys[columna], dataset_anlys['tripshour'])
    tripshour_pointbiserial.loc[i,'correlation']=aux[0]
    tripshour_pointbiserial.loc[i,'p_value']=aux[1]
tripshour_pointbiserial=tripshour_pointbiserial.sort_values(by=['correlation'],ascending=Fal
se)


from scipy.stats import pointbiserialr

dataset_anlys=pd.merge(MITMA_dataset_mob_ind, ABS_covid_dataset, on=["zone",
"covid_restrict"])

for num in ['15','18','21','34','40','42','43']:
```

```
    dataset_anlys=dataset_anlys.drop(dataset_anlys[dataset_anlys['zone']==num].index)

del dataset_anlys['tripshour']

dataset_anlys = pd.get_dummies(data=dataset_anlys, drop_first=False)

i=0
covid_pointbiserial=pd.DataFrame()
covid_pointbiserial['feature']='Nan'
covid_pointbiserial['correlation']='Nan'
covid_pointbiserial['p_value']='Nan'
for columna in dataset_anlys.columns:
    i=i+1
    covid_pointbiserial.loc[i,'feature']=columna
    aux=pointbiserialr(dataset_anlys[columna], dataset_anlys['num_cases'])
    covid_pointbiserial.loc[i,'correlation']=aux[0]
    covid_pointbiserial.loc[i,'p_value']=aux[1]
covid_pointbiserial=covid_pointbiserial.sort_values(by=['correlation'],ascending=False)
```

8. PATTERN RECOGNITION THROUGH MACHINE LEARNING TECHNIQUES

The 3 previous datasets are pivoted and merged to form a single dataset:

```
#########################################################
### PIVOTING MOBILITY, COVID AND CERTIFICATES DATASET ###
#########################################################
MITMA_dataset_mob_ind_piv=pd.pivot_table(MITMA_dataset_mob_ind, values=['tripshour'],
index=['zone'],columns=['covid_restrict','mobility_indicator','distance','timezone'])
ABS_covid_dataset_piv=pd.pivot_table(ABS_covid_dataset, values=['num_cases'],
index=['zone'],columns=['covid_restrict','test_description','sex'])

#########################################################
### MERGING MOBILITY, COVID AND CERTIFICATES DATASET ###
#########################################################
dataset_mob_covid_cert_population_pivoted = pd.concat([MITMA_dataset_mob_ind_piv,
ABS_covid_dataset_piv],axis=1)
#dataset_mob_covid_cert_population_pivoted =
pd.concat([dataset_mob_covid_cert_population_pivoted,
certificats_autoresponsabilitat_piv],axis=1)
dataset_mob_covid_cert_population_pivoted =
pd.concat([dataset_mob_covid_cert_population_pivoted, dataset_territori_piv1],axis=1)
dataset_mob_covid_cert_population_pivoted
```

Data are processed, records are merged to compensate for missing COVID-19 data in 7 zones and subsequently removed. Finally, null values are replaced by zero values:

```
dataset_mob_covid_cert_population_pivoted['zone']=dataset_mob_covid_cert_population_pivoted[
'Unnamed: 0']
del dataset_mob_covid_cert_population_pivoted['Unnamed: 0']

dataset_mob_covid_cert_population_pivoted.iloc[32-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[15-1, 32-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[29-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[18-1, 29-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[26-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[21-1, 26-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[44-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[34-1, 44-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[39-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[40-1, 39-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[28-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[42-1, 28-1], :].sum()
dataset_mob_covid_cert_population_pivoted.iloc[24-
1,:]=dataset_mob_covid_cert_population_pivoted.iloc[[43-1, 24-1], :].sum()
```

```
dataset_mob_covid_cert_population_pivoted=dataset_mob_covid_cert_population_pivoted.drop([15
,18,21,34,40,42,43])
dataset_mob_covid_cert_population_pivoted=dataset_mob_covid_cert_population_pivoted.fillna(0
)

dataset_mob_covid_cert_population_pivoted
```

The values are transformed by scaling with the mean and standard deviation:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler() #Standardize features by removing the mean and scaling to unit
variance
datadropnorm = scaler.fit_transform(dataset_mob_covid_cert_population_pivoted.astype(float))
```

The elbow technique is implemented with the original features scaled to have a tool to support the decision of how many clusters to take:

```
from sklearn import metrics
from sklearn import cluster

N=7
inrt=np.zeros(N)
sil_kmeans=np.zeros(N)

for i in range(N):
        print('k:',i+2)
        clf = cluster.KMeans(init='k-means++', n_clusters=i+2, random_state=0)
        clf.fit(datadropnorm)
        inrt[i]=clf.inertia_
        sil_kmeans[i]=metrics.silhouette_score(datadropnorm, clf.labels_,
metric='euclidean')
        print(inrt[i])
        print(sil_kmeans[i])
        print('-------')

plt.plot(np.arange(N),scaler.fit_transform(sil_kmeans.reshape(-1, 1)),'r',
label='silhouette')
plt.plot(np.arange(N),scaler.fit_transform(inrt.reshape(-1, 1)),'b', label='inertia')
plt.legend()
fig=plt.gcf()
fig.set_size_inches((16,5))
```

The k-means algorithm is implemented for 4 clusters with the scaled original features:

```
k=4
clf = cluster.KMeans(init='k-means++', n_clusters=k, random_state=0)
clf.fit(datadropnorm)
inrt_opt=clf.inertia_
sil_opt=metrics.silhouette_score(datadropnorm, clf.labels_,  metric='euclidean')
print('Optimal inertia:', inrt_opt, 'Optimal silhouette:', sil_opt)


y_pred2 = clf.predict(datadropnorm)
num_clusters2=pd.DataFrame()
num_clusters2['name_zone']='Nan'
num_clusters2['zone']=dataset_mob_covid_cert_population_pivoted.index
num_clusters2['cluster']=y_pred2


for i in range(len(zonification_name_zone_code_com['code_tfm'])):
    codi_tfm=zonification_name_zone_code_com.loc[i,'code_tfm']

num_clusters2['name_zone'].loc[(num_clusters2['zone']==int(codi_tfm))]=zonification_name_zon
e_code_com.loc[i,'district']

num_clusters2=num_clusters2.sort_values(by=['cluster', 'zone'])
num_clusters2
```

Data is transformed into its principal components:

```
from sklearn.decomposition import PCA
pca = PCA(.95)
pca.fit(datadropnorm)
datadropnorm = pca.transform(datadropnorm)
attributes=dataset_mob_covid_cert_population_pivoted.columns.tolist()
```

The elbow technique is implemented with the principal components of the original features to have a tool to support the decision of how many clusters to take:

```
from sklearn import metrics
from sklearn import cluster

N=7
inrt=np.zeros(N)
sil_kmeans=np.zeros(N)

for i in range(N):
        print('k:',i+2)
        clf = cluster.KMeans(init='k-means++', n_clusters=i+2, random_state=0)
        clf.fit(datadropnorm)
        inrt[i]=clf.inertia_
        sil_kmeans[i]=metrics.silhouette_score(datadropnorm, clf.labels_,
metric='euclidean')
        print(inrt[i])
        print(sil_kmeans[i])
        print('-------')


plt.plot(np.arange(N),scaler.fit_transform(sil_kmeans.reshape(-1, 1)),'r',
label='silhouette')
plt.plot(np.arange(N),scaler.fit_transform(inrt.reshape(-1, 1)),'b', label='inertia')
plt.legend()
fig=plt.gcf()
fig.set_size_inches((16,5))
```

The k-means algorithm is implemented for 4 clusters with the principal components of the original features:

```
k=4
clf = cluster.KMeans(init='k-means++', n_clusters=k, random_state=0)
clf.fit(datadropnorm)
inrt_opt=clf.inertia_
sil_opt=metrics.silhouette_score(datadropnorm, clf.labels_,  metric='euclidean')
print('Optimal inertia:', inrt_opt, 'Optimal silhouette:', sil_opt)


y_pred2 = clf.predict(datadropnorm)
num_clusters2=pd.DataFrame()
num_clusters2['name_zone']='Nan'
num_clusters2['zone']='Nan'
num_clusters2['cluster']='Nan'
num_clusters2['zone']=dataset_mob_covid_cert_population_pivoted.index
num_clusters2['cluster']=y_pred2


for i in range(len(zonification_name_zone_code_com['code_tfm'])):
    codi_tfm=zonification_name_zone_code_com.loc[i,'code_tfm']
    #print(codi_tfm)

num_clusters2['name_zone'].loc[(num_clusters2['zone']==int(codi_tfm))]=zonification_name_zon
e_code_com.loc[i,'district']

num_clusters4=num_clusters2.sort_values(by=['cluster', 'zone'])
num_clusters4
```