EUREKA CLUSTER ITEA 3 Software intensive systems and services

# Descriptive and Predictive Models for City Decision Support Tools

Deliverable 5.2

POLicy Data Exploitation & Re-use

| | |
|---|---|
| Project Identifier | POLDER |
| Project Title | POLicy Data Exploitation & Re-use |
| | |
| Document Version | 1.6.1 |
| Planned Delivery Date | 30/07/2021 (M32) |
| Actual Delivery Date | M36 |
| Document Title | Descriptive and Predictive Models for City Decision Support Tools |
| Work Package | WP5 |
| Document Type | Word |
| | |
| Abstract | Report describing the results of the different types of the data analysis and modelling work: exploratory data analysis results, exploration of different modelling approaches and preliminary results of the models selected for the case studies. |
| Keywords | Data inventory, algorithms, objectives, outputs, ontology |

| Function | Name | Entity |
|---|---|---|
| Author | Carlos Blasco | NOMMON |
| Editors | Irene Torrego | ACCURO |
| | - | - |
| Contributors | Ali KAFALI | ACD |
| | Irene Torrego | ACCURO |
| | Iván Romero | Starflow |
| | Carlos Blasco | NOMMON |
| | Ülkü Esra OKUYAN | ARD Group |
| | Bilge BAŞDEMİR | ARD Group |
| | --- | Mantis |
| | | |
| | | |
| | | |
| | | |

# EXECUTIVE SUMMARY

Report describing the results of the different types of the data analysis and modelling work: exploratory data analysis results, exploration of different modelling approaches and preliminary results of the models selected for the case studies.

## PARTNERS CONTRIBUTIONS RECORD

| # | Partners | Contributor on Phase 1 | Date of Contribution | Contributor on Phase 2 | Date of Contribution2 |
|---|---|---|---|---|---|
| 1 | Fortearge | | | | |
| 2 | Accuro | X | 29/06/2021 | | |
| 3 | ACD | X | 10/06/2021 | | |
| 4 | Mantis | | 22/09/2021 | | |
| 5 | Nommon | X | 23/07/2021 | | |
| 6 | Ard group | X | 30/07/2021 | | |
| 7 | Starflow | X | 12/07/2021 | | |
| 8 | | | | | |
| 9 | | | | | |
| 10 | | | | | |
| 11 | | | | | |
| 12 | | | | | |
| 13 | | | | | |
| 14 | | | | | |
| 15 | | | | | |

## CHANGES RECORD

| Version | Date | Entity | Description of Changes |
|---|---|---|---|
| 1.0 | 19/05/2021 | NOMMON | Initial Version |
| 1.1 | 10/06/2021 | ACD | Added ACD's models and algorithms in Section 2.2.1 |
| 1.2 | 29/06/2021 | ACCURO | Added ACCURO's models and algorithms in Section 2.1.1 |
| 1.3 | 12/07/2021 | STARFLOW | Added STARFLOW'S models and algorithms in Section 2.1.3 |
| 1.4 | 23/07/2021 | NOMMON | Added NOMMON'S models and algorithms in Section 2.1.2 |
| 1.5 | 30/07/2021 | ARD | Added ARD's contributions in Section 2.3.2 |
| 1.5.1 | 13/09/2021 | ACCURO | Edited format and updated Changes record table |
| 1.6 | 22/09/2021 | MANTIS | Added Mantis's models and algorithms in Section 2.3.1 |
| 1.6.1 | 08/11/2021 | ACCURO | Edited format and updated Changes record table |

# CONTENT

# List of figures

## List of tables

# 1 INTRODUCTION

## 1.1 Document Objectives and Scope

## 1.2 Document Structure

This document contains one section for each case study, divided in the following subsections:

- Case Study #1
    - Introduction
    - Partner #1
        - Model #1
            - Objective
            - General Approach
                - Input Data
                - Expected Output
                - Proposed Methodology
                - Tests
                - Conclusion
        - Model #2
            - Objective
            - General Approach
                - Input Data
                - Expected Output
                - Proposed Methodology
                - Tests
                - Conclusion

# 2 MODELLING

## 2.1 Smart Tourism (Tourism)

Smart Tourism use case aims to provide the tourism sector with smart tools to improve its quality and its capacity, leading to a better customer experience.

For this purpose, the POLDER project partners have developed different AI models based on neural networks that are able to process data from a wide variety of sources, providing the POLDER platform with a very broad view of different aspects of the tourism sector. This will allow to forecast the tourism demand in a given area, propose new innovative services for citizens and tourists, improve the resource management for touristic destinations, raise tourist context awareness, and detect behavioral patterns.

### 2.1.1 Accuro Technology

Accuro has developed four algorithms for object detection in video images based on the use of convolutional neural networks. These algorithms use the same data model to identify the objects contained in the images, and thus only one image recognition model will be explained in the following subsections.

#### 2.1.1.1 Model 1: Image recognition algorithm

The image recognition model created by Accuro processes and analyses images and videos to gather information about the environment that can be useful in the tourism sector. More specifically, it focuses on detecting the number of people (with and/or without facemask, depending on the context in which is applied), different types of vehicles (cars, bicycles, buses, etc.) and objects that tourists usually carry, such as suitcases and backpacks.

These objects will be used to estimate the number of tourists visiting the monitored areas and subsequently analyse their behaviour and preferences, so that the services offered can be improved and a more pleasant experience can be provided.

##### 2.1.1.1.1 Objective

The objective of Accuro is to monitor and understand the behaviour of complex and dynamic population groups in urban environments, relating the identified and labelled components, entities or subjects. By understanding the behaviour and relationships between people and the actions they take in a tourist area, it will be possible to enhance the quality of the services provided not only for tourists, but also for residents.

In this context, for sub-use cases have been identified within the use case of population dynamics monitoring addressed by Accuro, attending to different problems or needs, being these:

- Sub use-case 1: Tourists who want to visit a place open to the public (e.g., a museum).
- Sub use-case 2: Control of the number of people in a pedestrian area with restricted capacity.
- Sub use-case 3: Monitoring access to sites, premises and buildings according to COVID regulations.

- Sub use-case 4: Detection and classification of people and vehicles in tourist areas: beaches, restaurants, pedestrian areas, hotels, etc.

Accuro proposes the use of video image analysis and recognition algorithms for obtaining pedestrian and vehicle traffic flows. These algorithms can be easily adapted to other applications in different Smart Cities domains.

#### 2.1.1.1.2 General Approach

The image recognition algorithms, which are based on the YOLO technique and the Darknet convolutional neural network, follow a two-stage activity detection system. The first stage detects events and labels them; the second stage classifies them.

Following this procedure, they are able to detect, label and classify a wide range of objects of different classes in live or recorded videos or in static images.

##### 2.1.1.1.2.1 Input Data

The developed algorithm takes recorded or real-time video images as input. Although it is also able to use static images, it will preferably use live videos as the aim is to make real-time analysis.

##### 2.1.1.1.2.2 Expected Output

The images used by the algorithm will not be recorded, but will be used to generate labels of the detected objects. It is possible to detect several objects of the same type in the same frame, so the algorithm will provide the total number of objects of the same type detected within the same frame, and this number will be updated each time a new element is detected in the image.

Thus, the algorithm shall provide as output the count of the total number of detected objects of the same type at each time a new detection of that object occurs. The expected outputs of each sub use-case are show on *Table 1*.

Table 1. Accuro's image recognition algorithms expected output

| **Sub use-case 1** | **Sub use-case 2** | | **Sub use-case 3** | **Sub use-case 4** | |
|---|---|---|---|---|---|
| Total number of… | Total number of… | | Total number of… | Total number of… | |
| People | People | Dogs | People with facemask | Single persons | Cars |
| Handbags | Bicycles | Horses | People without facemask | 2-people groups | Trucks |
| Suitcases | Cars | Handbags | | 3-people groups | Motorcycles |
| Backpacks | Motorcycles | Backpacks | | 4-people groups | Bicycles |

| Sub use-case 1 | Sub use-case 2 | | Sub use-case 3 | Sub use-case 4 | |
|---|---|---|---|---|---|
| Total number of… | Total number of… | | Total number of… | Total number of… | |
| | Buses | Handbags | | 5-people groups | Buses |
| | Trucks | Scooters | | 6-people groups | Bathers / swimmers |
| | | | | Groups with more than 6 people | |

#### 2.1.1.1.2.3 Proposed Methodology

The developed algorithms follow a two-stage activity detection system: in the first stage, the videos are processed to generate event proposals to spatially and temporally locate the candidates and the activity they are performing; in the second stage, features are extracted, and spatio-temporal classification and post-processing is performed in order to generate the activity detection results.

- Stage 1: The labels of the detected events (detection of a car, person, bicycle, etc.) are generated.
- Stage 2: The labels are classified in the database.

Figure 1 represents the model architecture, where the two-stage detection system is illustrated.

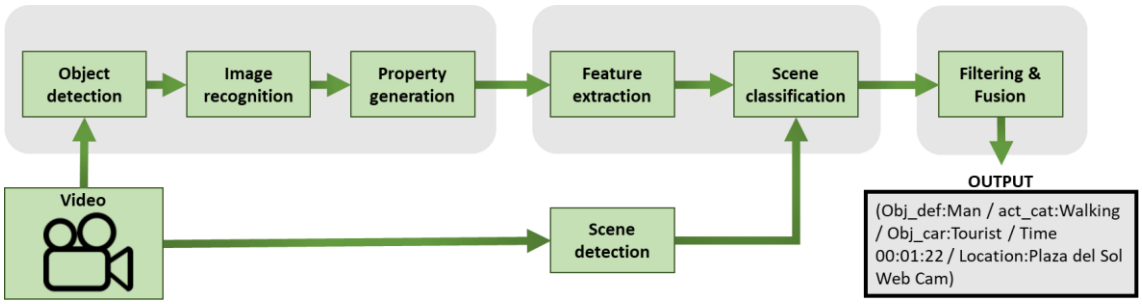



*Figure 1. Process of subjects/objects detection and labelling in images*

To perform the object detection, a POLDER dataset has been created with at least 10 000 images of each object class to be detected. Images taken from different angles and distances, and with different colours and shapes for the same object have been used in order to increase the probability of detecting the objects in case any of these variations occur. In addition, different ages, races, genders and clothing have been taken into account to detect people.

For each image an XML file containing the objects that appear in the image and their position within the image have been created. The dimensions of the object (height and width) in pixels are also indicated in this XML file.

The object detection model has been developed using the YOLO (*You Only Look Once*) technique, the fastest real-time video processing technique available up until now. YOLO

makes use of the Darknet convolutional neural network for image classification. The model has been trained with the images and XML files previously mentioned.

Once the training is complete, the algorithms use this model to identify the corresponding object classes corresponding to each sub use-case and count the total number of objects of each type found in the image.

#### 2.1.1.1.2.4 Tests

The image recognition algorithms have been tested with several videos obtained from public IP cameras, videos found on Youtube and videos recorded by Accuro staff.

Figure 2 shows an example of how object detection works. This image shows a street with people walking and cars driving on the road. When the algorithm detects one of the objects defined for the corresponding use case, it marks it with a rectangle and indicates the type of object at the top of the rectangle. The probability that the identified object is actually the one of the indicated classes is displayed next to the object name. The closer this number is to 1, the higher the probability of having correctly identified the object.


*Figure 2. Object detection algorithm performance*

#### 2.1.1.1.2.5 Conclusion

The performance of the image recognition model has been successful, as after training it has been able to achieve a very low error during the validation and testing processes. This indicates that the algorithms will correctly detect and classify objects present in an image with a very high probability.

A phenomenon that has been observed is the capability to detect objects reflected in glass, which leads to erroneous detections as objects are detected in duplicate. Therefore, when installing the cameras for the pilot, it will be necessary to avoid mirrors, windows and glass in general in their field of view.

## 2.1.2 Nommon

Nommon has developed a model based on unsupervised machine learning for the classification of the zones of the city, for a given zoning configuration, in terms of their tourism attractiveness. Further to the classification of the zones, the attributes (number of points of interest, hotels, etc.) of each class are provided.

### 2.1.2.1 Cities' zones classification and characterization

#### 2.1.2.1.1 Objective

The objective of this work is to provide private and public tourism as well as other city administrations with relevant tools for the planning and management of tourism and tourism related measures.

In particular, the objective of the model developed is the classification of different zones of the city according to the visitors they attract (e.g., primary residents, international visitors, etc.) and the activities performed there as well as the identification of those attributes that characterise them and single them out from zones belonging to other categories, e.g., identify which are the attributes of those zones that result more attractive to international visitors vs national visitors or to visitors vs residents. This classification will allow public administrations, for instance, to identify those resident areas with a higher tourism pressure and hence are more likely to present conflicts between residents and tourists. Or for tourism services providers to identify the type of services that attract different types of tourists or to foresee which zones will receive higher demand of services for a given date or occasion to be (i.e., national holiday, international holiday, football match, etc.). It will also allow services providers to identify which are those services which attract visitors of a given profile.

The final outcome of the model will be a set of classes with a list of the zones belonging to each class and the characteristics of each class.

#### 2.1.2.1.2 General Approach

The development of the model is based on the hypothesis that the offer of different services in the city attracts a distinct combination of visitors and residents. Hence each zone of the city may be classified according to the mix of visitors it receives. Since these classes are not known a priori, unsupervised machine learning algorithms will be used to discover the patterns of visitors. More concretely clustering techniques will be used to identify different classes of zones (clusters) which share similar characteristics in terms of visitors mixtures.

The first challenge to build the model consists in identifying the relevant attributes that can be calculated for each zone that is relevant for its tourism characterization. This information will later on be analysed iteratively to obtain the indicators that better describe the touristic attributes of the study zones.

The proposed clusters will be crossed with attributes relative to the zones, like number of museums per zone or airbnbs per zone, to find the features that better define the clusters behaviour and explain the presence of the different visitors.

##### 2.1.2.1.2.1 Input Data

The data used to develop the study is contained in the following files:

- Presence indicators: Information related to the presence of persons in each study zone in the study area. This data set measures the number of unique persons in every hourly interval for the given study dates. This information has been segmented with fields related to the agents and to the performed activities, as it is described below. These indicators have been obtained applying the algorithms for activity and mobility extraction from mobile phone data developed by Nommon and are available in the following format:

  "zone_id","date","interval","target_population","non_residents_residence","residents_residence","activity_type","gender","age","income","stay_length","indicators","value"

  This way, the instances of the file are of the form:

  "0","2019-04-01","00:00-00:59","residents","Spain","0704801002","frequent","female","25-44","income_4","4-inf","persons","4"

- Overnight indicators: Information related to the daily volume of persons having an overnight in each study zone. Just like the presence indicators, the information has been segmented as follows:

  “study_date”,”zone_id”,”nationality”,”home_province”,”home_zone”,”age”,’gender’,”overnights”

  This way, the instances of the file are of the form:

  “20190428","0","Spain","28","13","A1","Male","2.69813377084”

  “20190421","0","Spain","31","0","A2","Male","3.05134919731”

The time period covered by both files comprises April 2019, in order to capture different touristic behaviours in the study area, like tourism during ordinary weekends or tourism during holidays. In Spain, easter holidays is a national holiday that, in 2019, took place during april, so regular weekdays can be compared to holiday weekdays to obtain comparative indicators.

The fields contained in the files are described as follows:

- zone_id: zone identifier
- date/study_date: study date in format YYYY-MM-DD and YYYYMMDD, respectively
- Interval:hourly intervals in format HH:MM
- target_population: population will be characterised as resident or non-resident, depending on if they live in Spain or not
- non_residents_residence/nationality: residence country
- residents_residence: census tract of residence for national agents according to INE codes (www.ine.es)
- activity_type: activities will be classified based on a longitudinal analysis of the agents as:
  - home - if the activity is taking place in the users identified home
  - work - if the activity is taking place in the users identified workplace
  - frequent - if the activity is taking place in a users recurrent area
  - non_frequent - if the activity is taking place in a non recurrent area
- gender: national agents will be segmented into males and females

- age: national agents will be grouped into the following age groups
    - 0-24
    - 25-44
    - 45-64
    - 65-inf
- income: average income of the census tract of residence
    - income_1: 0-7000 €
    - income_2: 7000-10000 €
    - income_3: 10000-12000 €
    - income_4: 12000-15000 €
    - income_5: >15000 €
- stay_length: duration of the activity in the study zone
    - 0-1: less than an hour
    - 1-2: between one and two hours
    - 2-4: between two and four hours
    - 4-inf: more than four hours
- Indicators: “persons”
- value:  number of people
- home_province: province of residence for national agents, according to INE codes
- home_zone: identifier of the zone where the agents home has been detected
- overnights: number of overnights

Other three files were considered for the study:

- InsideAirbnb: Data set with information of the exact location for every airbnb in Madrid is extracted.
- Overpass turbo: Osrm API, where study zones can be extracted in a shape file. From this data set relevant information like the location of museums, hotels or restaurants is extracted.
- POI location: Dataset with the most important tourist attractions in Madrid.

#### 2.1.2.1.2.2 Expected Output

When completed, this model should offer two outputs.

The first output will contain information about what zones have formed each cluster. The information would be grouped as follows:

Cluster_1: zone_1, zone_4, zone_5…

Cluster_2: zone_2, zone_3…

…

The second output would contain the information of the analysis about what characteristics define each cluster:

Cluster_1: residential cluster with mainly national visitors and residents.

Cluster_2: …

…

#### 2.1.2.1.2.3 Methodology

The general methodology for the study consists of the following steps.

#### 2.1.2.1.2.4 Definition of the zones

The case study is applied to the study city of Madrid. Madrid is one of the most important cities in Spain, where more than 3 million people live. The supply of possible activities and attractions in Madrid attracted more than 10 million visitors during 2019 and that is why Madrid is an interesting study case, as it combines a great amount of possible activities for residents as well as for visitors.

Tourist and residential areas can easily be told apart for the people living in Madrid. Nevertheless, the main challenge will be identifying if these touristic zones can be differentiated from one another as well as understanding what features have a bigger impact on a zone's tourism characteristics.

For this characterization, the municipality of Madrid was divided into a grid of 1000*1000m, resulting in 688 zones (see Figure 1). This grid size had the better tradeoff in not grouping zones with different tourism characteristics into one and not compromising the quality of the tourism indicators calculated with mobile phone data.

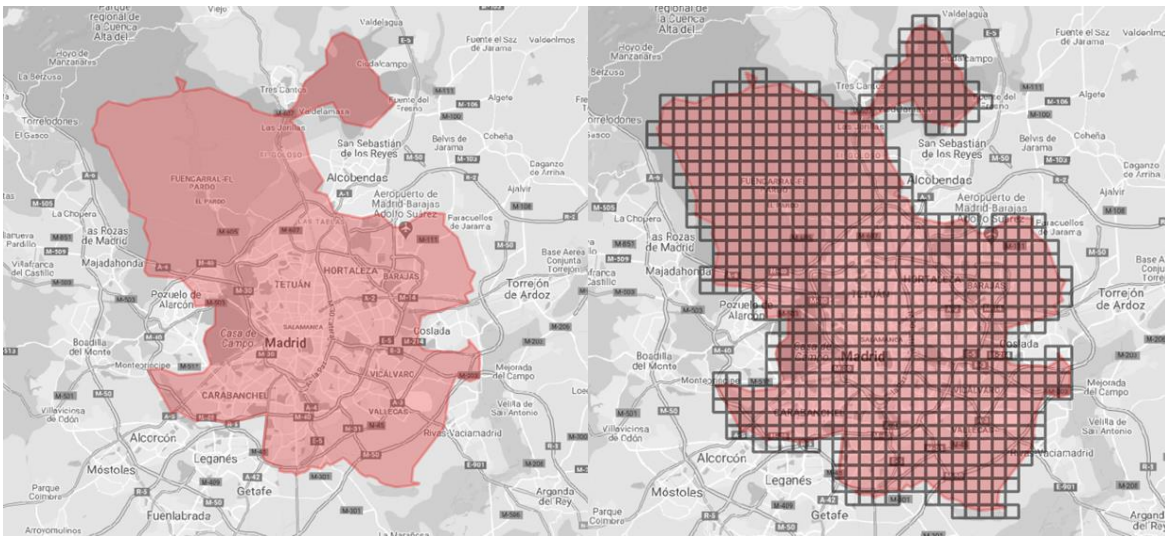



*Figure 3. Division of the municipality of Madrid into a grid of 688 zones of 1000*1000m.*

For each one of these zones, presence and overnight indicators have been extracted using the solution developed by Nommon. These indicators, described in section **1.1.1.1.2.1,** offer

information about the spatial and temporal distribution of the agents, as well as the zones where these agents are spending the night.

#### 2.1.2.1.2.5 Selection of variables for clustering

Next, in order to extract each zone's tourist characteristics, a series of variables have been defined and computed using the information available in the presence and overnight indicator files. With this information we can identify the kind of people that visit or spend the night in each zone, namely, residents (residents in C. Madrid), and visitors (non-resident people in C. Madrid), distinguishing for this last case between national visitors (residents in Spain who do not reside in C. Madrid) and foreign visitors (non-resident people in Spain). As we can see, the information of the fields *value* and *overnights,* of the presence and overnights indicators, respectively*,* already provides the total number of each kind of person per zone. However, we do not directly use these total numbers as we want to identify which group of persons dominate each zone and the difference in the distribution of each group of persons among the zones, and we consider that these data do not provide accurate enough information for that and that better variables can be defined to address those questions.

For that, a series of variables were defined and computed, reflecting the proportion in which each group of persons is present or spend the night in each zone and the ratios between these proportions. These values will allow us to better capture and identify the behaviour of each zone by means of the presence and overnight stays of each group of people.

The information used to compute the variables is the one appearing in the fields *value* and *overnights*, while the information of the fields *zone_id, date, interval, target_population, non_residents_residence, study_date,* and *nationality* is used to separate or filter among the kind of person and the time period per zone.

With this idea, three kinds of variables per zone have been defined: presence variables, overnight stay variables and hourly presence variables. Next, we describe each of them in detail.

**Presence variables**

Using the presence information available in the presence indicators file, we defined three variables to represent the kind of people that visit a zone, and the proportion they represent compared to the total. With them, we aim at identifying and characterizing the tourist presence and distribution along the zones of Madrid. All of them are normalised such that, when defined, their values lie in the interval [0,1], in order to have the information scaled and bounded:

- **Visitors/residents presence ratio per zone**. This ratio measures the proportion of visitors with respect to the total number of persons in a zone. It is defined as the quotient between the number of visitors and the total number of persons in each zone:

$$\text{ratio visitors/residents} = \frac{\#\text{visitors}_i}{\#\text{visitors}_i + \#\text{residents}_i},$$

  where $\#\text{visitors}_i$ and $\#\text{residents}_i$ stand for the total number of visitors and residents in zone i for the entire month, respectively.

- **Foreign visitors/national visitors presence ratio per zone**: This ratio measures the proportion of foreign visitors with respect to the total number of visitors in a zone. It is defined as

the quotient between the number of foreign visitors and the total number of visitors in each zone:

$$\text{ratio foreign visitors/national visitors}_i = \frac{\#\text{foreign visitors}_i}{\#\text{foreign visitors}_i + \#\text{national visitors}_i},$$

where #foreign visitors$_i$ and #national visitors$_i$ stand for the total number of foreign and national visitors in zone i for the entire month, respectively.

- Seasonality presence ratio per zone: This ratio measures the visits seasonality component, i.e., the increase or decrease in the number of visitors between holidays and normal days in a zone. It is defined as the quotient between the number of visitors on April 20, 2019 (the Saturday of Spanish Holy Week) and the total number of visitors on April 6 (a standard Saturday of April, i.e., non-holiday or close-to-holiday day) and April 20, 2019 in each zone:

$$\text{ratio seasonality} = \frac{\#\text{visitors(20 April)}_i}{\#\text{visitors(20 April)}_i + \#\text{visitors(6 April)}_i},$$

where #visitors(20 April)$_i$ and #visitors(6 April)$_i$ stand for the number of visitors on April 6 and April 20 in zone i, respectively. With this ratio, we measure the proportion of visitors on April 20 with respect to the total number of visitors on April 6 and 20.

Since the number of residents per zone is typically much higher than the number of visitors, and the same may happen with the number of national visitors over foreign visitors, we decided to compute these ratios using relative values, rather than absolute ones. For that, we normalised them by dividing the total number of each kind of person in that zone by the total number of persons of that kind in all the zones. This way, what we are actually comparing is the percentage of each kind of person present in that zone. With this approach, the definition of the variables becomes:

- **Visitors/residents presence ratio per zone**. This ratio is defined as the quotient between the percentage of visitors and the percentage of total persons in each zone:

$$\text{ratio visitors/residents normalised}_i = \frac{\frac{\#\text{visitors}_i}{\sum_i \#\text{visitors}_i}}{\frac{\#\text{visitors}_i}{\sum_i \#\text{visitors}_i} + \frac{\#\text{residents}_i}{\sum_i \#\text{residents}_i}},$$

where, as before, #visitors$_i$ and #residents$_i$ stand for the total number of visitors and residents in zone i for the entire month, respectively.

- **Foreign visitors/national visitors presence ratio per zone**: This ratio is defined as the quotient between the percentage of foreign visitors and the percentage of total visitors in each zone:

$$\text{ratio foreign visitors/national visitors normalised}_i = \frac{\frac{\#\text{foreign visitors}_i}{\sum_i \#\text{foreign visitors}_i}}{\frac{\#\text{foreign visitors}_i}{\sum_i \#\text{foreign visitors}_i} + \frac{\#\text{national visitors}_i}{\sum_i \#\text{national visitors}_i}},$$

where, as before, #foreign visitors$_i$ and #national visitors$_i$ stand for the total number of foreign and national visitors in zone i for the entire month, respectively.

- **Seasonality presence ratio per zone**: This ratio is defined as the quotient between the percentage of visitors on April 20, 2019 (the Saturday of Spanish Holy Week) and the percentage of total visitors on April 6 (a standard Saturday) and April 20, 2019 in each zone:

$$\text{ratio seasonality normalised}_i = \frac{\frac{\#\text{visitors(20 April)}_i}{\sum_i \#\text{visitors(20 April)}_i}}{\frac{\#\text{visitors(20 April)}_i}{\sum_i \#\text{visitors(20 April)}_i} + \frac{\#\text{visitors(6 April)}_i}{\sum_i \#\text{visitors(6 April)}_i}},$$

  where, as before, $\#\text{visitors(20 April)}_i$ and $\#\text{visitors(6 April)}_i$ stand for the number of visitors on April 6 and April 20 in zone i, respectively.

Additionally to those ratios and for analysis purposes, we also computed the total number of residents, visitors, national visitors and foreign visitors present in each zone along the month.

### Overnight stay variables

Using the overnight information available in the overnight indicators file, we defined three variables to represent the kind of people who spend the night in a zone, and the proportion they represent with respect to the total. With them, we aim at identifying and characterizing the tourists' overnight stays and their distribution along the zones of Madrid. As with the presence ratios, all of them are defined in such a way that, when defined, their values lie in the interval [0,1], in order to have the information scaled and bounded:

- **Visitor/resident overnight stays ratio per zone**. This ratio measures the proportion of visitor overnight stays with respect to the total number of overnight stays in a zone. It is defined as the quotient between the number of visitor overnight stays and the total number of overnight stays in each zone:

$$\text{ratio visitors/residents overnights}_i = \frac{\#\text{visitor overnights}_i}{\#\text{visitor overnights}_i + \#\text{resident overnights}_i},$$

  where $\#\text{visitor overnights}_i$ and $\#\text{resident overnights}_i$ stand for the total number of visitors and residents overnight stays in zone i along the entire month, respectively.

- **Foreign visitor/national visitor overnight stays ratio per zone**: This ratio measures the proportion of foreign visitor overnight stays with respect to the total number of visitor overnight stays in a zone. It is defined as the quotient between the number of foreign visitor overnight stays and the total number of visitor overnight stays in each zone:

$$\text{ratio foreign visitors/national visitors overnights}_i = \frac{\#\text{foreign visitor overnights}_i}{\#\text{foreign visitor overnights}_i + \#\text{national visitor overnights}_i},$$

  where $\#\text{foreign visitor overnights}_i$ and $\#\text{national visitor overnights}_i$ stand for the total number of foreign and national visitor overnight stays in zone i along the entire month, respectively.

- **Seasonality overnight stays ratio per zone**: This ratio measures the overnight stays seasonality component, i.e., the increase or decrease in the number of overnight stays between holidays and normal days in a zone. It is defined as the quotient between the number of visitor overnight stays on April 20, 2019 (the Saturday of Spanish Holy Week) and the total number of visitor overnight stays on April 6 (a standard Saturday) and April 20, 2019 in each zone:

$$\text{ratio seasonality overnights}_i = \frac{\#\text{visitor overnights(20 April)}_i}{\#\text{visitor overnights(20 April)}_i + \#\text{visitor overnights(6 April)}_i},$$

where #visitor overnights(20 April)$_i$ and #visitor overnight(6 April)$_i$ stand for the number of visitor overnight stays on April 6 and April 20 in zone i, respectively.

Since the number of resident overnight stays per zone is typically much higher than the number of visitor overnight stays, and the same may happen with the number of national visitor overnight stays over foreign visitor overnight stays, we decided to compute these ratios using relative values, rather than absolute ones. For that, we normalised them by dividing the total number of each kind of person spending the night in that zone by the total number of persons of that kind spending the night in all the zones. This way, what we are actually comparing is the percentage of each kind of person spending the night in that zone. With this approach, the definition of the variables becomes:

- **Visitor/resident overnight stays ratio per zone**. This ratio measures the proportion of visitor overnight stays with respect to the total proportion of overnight stays of each kind in a zone. It is defined as the quotient between the percentage of visitors overnight stays and the percentage of total overnight stays in each zone:

$$\text{ratio visitors/residents overnights normalised}_i = \frac{\frac{\#\text{visitor overnights}_i}{\sum_i \#\text{visitor overnights}_i}}{\frac{\#\text{visitor overnights}_i}{\sum_i \#\text{visitor overnights}_i} + \frac{\#\text{resident overnights}_i}{\sum_i \#\text{resident overnights}_i}},$$

where, as before, #visitor overnights$_i$ and #resident overnights$_i$ stand for the total number of visitors and residents overnight stays in zone i along the entire month, respectively.

- **Foreign visitor/national visitor overnight stays ratio per zone**: This ratio measures the proportion of foreign visitor overnight stays with respect to the total proportion of visitor overnight stays of each kind in a zone:

$$\text{ratio foreign visitors/national visitors overnights normalised}_i = \frac{\frac{\#\text{foreign visitor overnights}_i}{\sum_i \#\text{foreign visitor overnights}_i}}{\frac{\#\text{foreign visitor overnights}_i}{\sum_i \#\text{foreign visitor overnights}_i} + \frac{\#\text{national visitor overnights}_i}{\sum_i \#\text{national visitor overnights}_i}},$$

where, as before, #foreign visitor overnights$_i$ and #national visitor overnights$_i$ stand for the total number of foreign and national visitor overnight stays in zone i along the entire month, respectively.

- **Seasonality overnight stays ratio per zone**: This ratio measures the overnight stays seasonality component, i.e., the increase or decrease in the number of overnight stays between holidays and normal days in a zone. It is defined as the quotient between the percentage of visitors overnight stays on April 20, 2019 (the Saturday of Spanish Holy Week) and the percentage of the total number of visitors overnight stays on April 6 (a standard Saturday) and April 20, 2019 in each zone:

$$\text{ratio seasonality overnights normalised}_i = \frac{\frac{\#\text{visitor overnights(20 April)}_i}{\sum_i \#\text{visitor overnights(20 April)}_i}}{\frac{\#\text{visitor overnights(20 April)}_i}{\sum_i \#\text{visitor overnights(20 April)}_i} + \frac{\#\text{visitor overnights(6 April)}_i}{\sum_i \#\text{visitor overnights(6 April)}_i}},$$

where #visitor overnights (20 April)i and #visitor overnight(6 April)i stand for the number of visitor overnight stays on April 6 and April 20 in zone i, respectively.

Note that we defined the ratios in an analogue way to the presence ratios in order to make the clustering results comparable.

Additionally, to those ratios and for analysis purposes, we also computed the total number of residents, visitors, national visitors and foreign visitors overnight stays in each zone along the month.

**Hourly presence variables**

Finally, the last kind of variables computed correspond to the amount of persons present in a zone each hourly interval. With this information, we aim at characterising the zones according to the time of the day of higher activity (i.e night time activity zones, working hours activities, etc.). This will allow us to better characterise or identify the behaviour in that zone and the kind of activities carried out.

Using the presence information available in the presence indicators file, the time series, representing the amount (percentage) of persons present in a zone each hourly interval along the day, were computed according to the following procedure:

i. select the length of the time interval to compute the time series, in our case, hourly presence along a day;
ii. in order to get more representative and robust information about the hourly presence in each zone, we take a range of days similar to the kind of day selected to compute the time series. For instance, to compute the hourly presence of a standard working day of April, we can take the information of 3 standard working days of that month, meaning days not holidays or close to holidays;
iii. compute the sum over the selected range of days of the number of persons present in the zone for each hour of the time period selected;
iv. compute the average of those hourly values for scaling purposes. For instance, if we take all the Mondays and Saturdays of April to compare their hourly presence, it is necessary to compute the mean of those values per kind of day to scale them, since April 2019 has more Mondays than Saturdays;
v. compute the percentage of presence in each hour of the time interval with respect to the total presence of the whole day.

Following this procedure, three time series were computed:

1. The time series of a standard working day of April, taking three standard working days (Tuesday 2nd, Wednesday 3rd and Thursday 4th).
2. The time series of a standard weekend day of April, taking a standard weekend (Saturday 6th and Sunday 7th).
3. The junction of the two previous time series, i.e., the time series of a standard working day and a standard weekend day of April.

With these three kinds of time series, we capture the hourly behaviour of each zone for the two types of week days, working days and weekend days, both independently and jointly. This will allow us for a more complete analysis of the zones.

#### 2.1.2.1.2.6 Analysis and cleansing

Once all the ratios defined are computed, we analysed the complete dataset. First of all, we performed a spatial analysis of the variables computed.

Next, by the definition used to compute the presence and overnight stay variables, some of them are not defined for all the zones. This is the case, for instance, of the foreign visitor/national visitor overnight stays ratio, which is not defined for those zones with no visitor overnight stays. We defined the ratio value as -1 for these zones, to represent this behaviour.

On the other hand, some of the hourly presence time series considered are not defined either for some zones, for instance, for those zones with no visitors on the range of days selected. In this case, we defined the time series as identically zero, i.e., a vector of zeros, as the presence in those zones is zero.

Then, we performed a statistical analysis of the less representative variables according to the spatial analysis, to see how scattered the values are. We computed the number of zones for which the ratio is defined, its mean, standard deviation, first, second and third quartile and the minimum and maximum values of the variable.

Finally, we analysed the total number of residents and visitors per zone. We used that information to set thresholds for the number of residents and visitors per zone, in order to remove noisy zones, meaning zones which are barely visited, from the study. To fix the values of the thresholds, we analysed the distribution of zones remaining for different values of the threshold and applied the elbow method. This method consists in finding the point where the decrease in the variable of interest, in this case, the number of zones, begins to slow. Note that, as the threshold values increase, the number of zones decreases, thus, this method consists in finding a trade-off between the variables.

We only considered the number of visitors and residents to filter the zones, instead of considering the number of overnight stays as well, because the overnight stays are restricted to the presence of accommodations of some kind, hence, some relevant zones for tourism without (enough) accommodations may disappear from the study, such as park and commercial zones.

#### 2.1.2.1.2.7 Clustering analysis

Next, we performed clustering analyses for each of the different sets of variables computed. Different clustering techniques and combinations of them are considered. We followed two different approaches.

First, we applied k-means clustering. This algorithm requires the selection of the number of clusters, $k$, in which the data will be grouped. It starts by randomly choosing the centroids of each cluster. Then, each point is assigned to the closest centroid. Once the clusters are formed, the centroids for each cluster are computed again. The algorithm iteratively repeats these two steps until the centroids do not change or any other alternative relaxed convergence criterion is met.

Taking in mind the shape of the data (zones of Madrid) and the kind of shape expected for the clusters, this algorithm seems a good option, given its ability to find convex-like clusters.

To obtain the optimal number of clusters, the algorithm is run for different values of $k$, and for each one of the results a score is obtained to measure the quality of the clusters. The score selected was the inertia, which measures the distance between each data point and its centroid, computes the square of this distance and sums these squares over a cluster. To find the optimal value of $k$ the elbow method was used. Note that, as the number of clusters increases, the inertia decreases, thus, it is essential to find a good trade-off between the number of clusters and their quality.

For the implementation of the k-means algorithm, the *scikit-learn* Python library was used, specifically its sub-module *cluster*, which includes the function named *KMeans*. This function receives as parameters the number of clusters *k*, the maximum number of iterations to run (set by default to 300) and the random seed to generate the initial random centroids (set to 0 by default). The rest of the parameters of the algorithm can be found in the DOCUMENTATION. The distance function used by this algorithm is the Euclidean distance.

This first approach was applied to the three sets of variables computed independently.

The second approach considered consists of a two-level clustering analysis. With this approach, k-means was first applied, and then, hierarchical clustering was applied to the centroids of the clusters obtained with k-means. The first clustering layer (k-means) groups the zones into groups of similar zones, to perform a preliminary clustering. This preliminary clustering is refined through a second clustering layer. For that, the centroids of the clusters are computed, and each zone is identified with the centroid of the cluster it belongs to. These new attributes for each zone are then used to apply a hierarchical clustering with which the zones of Madrid are finally classified.

Hierarchical clustering does not require to specify in advance the number of clusters. Moreover, the algorithm allows the plotting of dendrograms to visualize the clusters' distribution, enabling the choice of the most appropriate number of clusters. However, it is necessary to define the maximum distance for a point to be considered part of a cluster, called distance threshold. These distances can be computed using different methods (single link, complete link, ...). We applied agglomerative clustering, a category of hierarchical clustering in which each point is initially considered a single cluster and the algorithm iteratively merges the ones that are closer to each other until there is only one cluster.

To select the number of clusters, the dendrogram is computed and based on the clusters' distribution, the threshold value is fixed. We selected different threshold values and kept the one that gives the best results.

For the implementation of the agglomerative clustering algorithm, the Python library used was *scikit-learn*, and its sub-module *cluster*, which includes the function named *AgglomerativeClustering*. This function receives as parameters the distance threshold to stop aggregating points to clusters and the linkage method, i.e., the way to compute the distances between points, among others. The rest of the parameters of the algorithm can be found in the DOCUMENTATION. For this experiment, we take linkage = 'ward' (the default one). When this is the linkage criterion, only Euclidean distance can be chosen for this algorithm.

This second approach was also applied to the three sets of variables computed independently.

We notice that when applying both clustering approaches to the time series, we are not comparing the similarity of the time series per se, i.e., the similarity of their temporal profiles (as shown in Figures 5, 6 and 7), but the similarity of the vectors defining them, i.e., the 24-variables vector or 48-variables vector containing the percentage of people present in a zone per hour, by means of the Euclidean distance.

Once the clusters are computed, a spatial analysis of the clustering results is performed in order to visualize the distribution in clusters of the zones and analyse the quality and consistency of the classification. For the presence and overnight stay variables we also performed a statistical analysis of their values within each cluster, and for the hourly presence

variables, we plotted the mean, filled with the standard deviation, of the time series of the zones belonging to the same cluster.

#### 2.1.2.1.2.8 Interpretation of the results

Lastly, we interpret the final clustering results using the following information:

1. the location of a set of points of interests of Madrid, and
2. the number of hotels, pubs, guest houses and airbnbs located in each zone.

The set of points of interest were selected in collaboration with UNWTO (United Nations World Tourism Organisation), to analyse their cluster location for each kind of clustering classification. The distribution of hotels, pubs, guest houses and airbnbs is used to analyse in more detail the clustering results obtained with the overnight stays variables.

#### 2.1.2.1.2.9 Tests

In this section we show the results obtained.

##### 2.1.2.1.2.9.1 Data analysis and cleansing

First, the variables defined in Section 2.1.2.1.2.5 were computed and analysed through a spatial analysis of their values.

**Presence variables**

Figure 4.shows the comparison of the values of the visitors/residents ratio and the visitors/residents ratio normalised for each zone of Madrid.

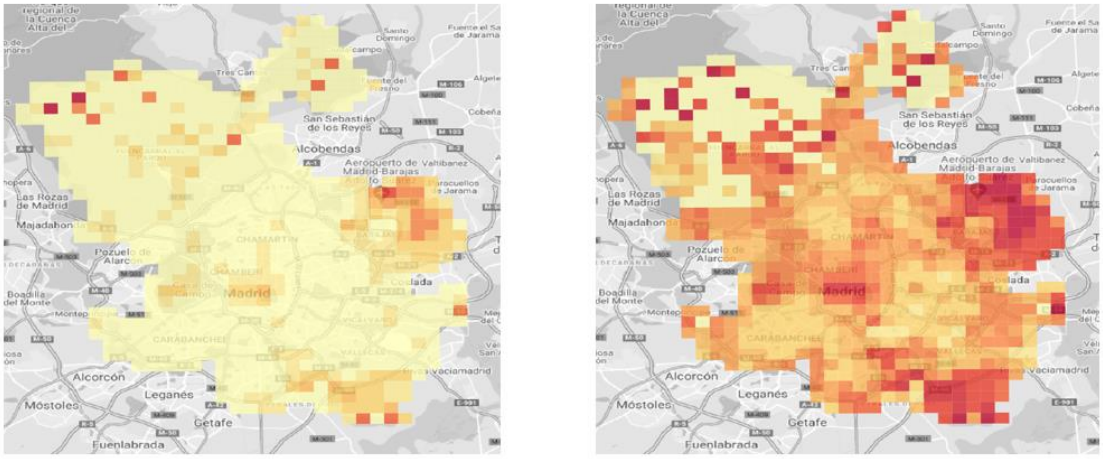



*Figure 4. Comparison of the values of the visitors/residents ratio for each zone of Madrid considering non-normalised and normalised ratios, respectively, where yellow color stands for values close to zero and red color, for values close to one. On the left we have the representation of the ratio using the non-normalised values; and on the right, the normalised values.*

As we can see, the ratio values computed using the non-normalised values seem more uniform along the zones, and much lower. This is because the number of residents in a zone is much higher than the number of visitors, as we mentioned before. Hence, the normalised values seem more informative and they are the ones we have kept for the clustering analysis.

Figure 5 shows the values distribution of the three ratios for the zones of Madrid.

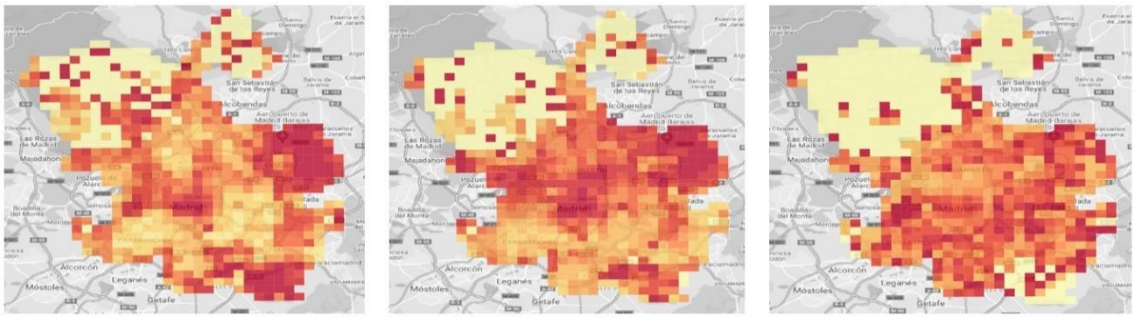


*Figure 5. Values distribution of the three ratios for the zones of Madrid, where yellow color stands for values close to zero and red color, for values close to one. On the left, the visitors/residents ratio, in the middle, the foreign/national visitors ratio, and on the right, the seasonality ratio.*

As we can see in the first two images of

, the visitors/residents and the foreign/national visitors ratios seem to follow a similar pattern among the zones, taking higher values in the city centre and the airport zones, and distinguishing groups of zones. However, the seasonality ratio values do not seem as informative as the other two ratios, as no clear structure is visible. We will further analyse that in this section.

### Overnight stay variables

Having in mind the results obtained with the presence ratios using non-normalised and normalised values (see Figure 4), we decided to compute the overnight variables using normalised values directly.

Figure 6 shows the values distribution of the three overnight stay ratios for the zones of Madrid. We observe that, in the foreign/national visitor overnight stays ratio and the seasonality overnight stays ratio representation, there are zones without value (in grey). These are zones where the ratio is not defined, i.e., zones where there are no visitor overnight stays.

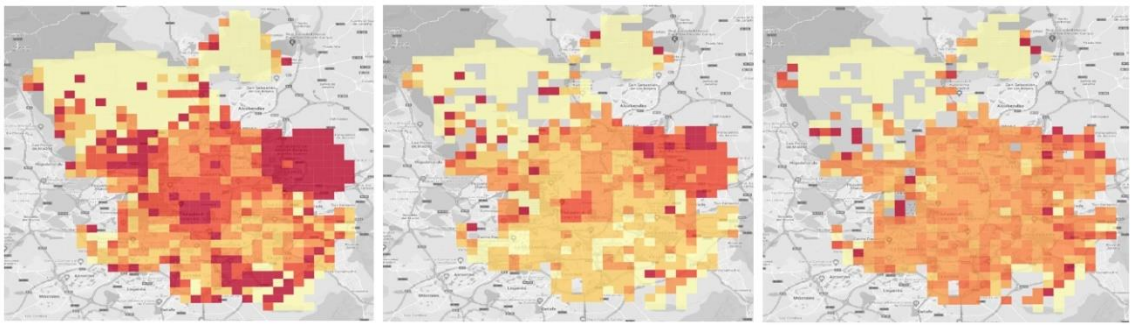


*Figure 6. Values distribution of the three ratios for the zones of Madrid, where yellow color stands for values close to zero and red color, for values close to 1. On the left, the visitor/resident overnight stays ratio, in the middle, the foreign/national visitor overnight stays ratio, and on the right, the seasonality overnight stays ratio.*

As we can see in the first two images of Figure 6, the seasonality ratio values seem rather uniform among the zones, which does not seem quite informative (as happened with the presence seasonality ratio). We will further analyse that in this section.

### Hourly presence variables

Three time series were computed, as described in Section 2.1.2.1.2.5. Next, a sample of the time series obtained for four zones of Madrid (the same four zones in all cases) is shown.

a. The time series of a standard working day of April, taking three standard working days (Tuesday 2nd, Wednesday 3rd and Thursday 4th). Figure 7shows a sample of the time series obtained for four zones of Madrid. As we can see, the distribution of people throughout the day is pretty different for each zone. The first one (blue one) presents higher activity during the evening, while the second one (orange one) presents higher activity during the early morning, the third one (green one), during the early morning, evening and night times, and finally, the fourth one (red one) presents higher activity during the day hours.

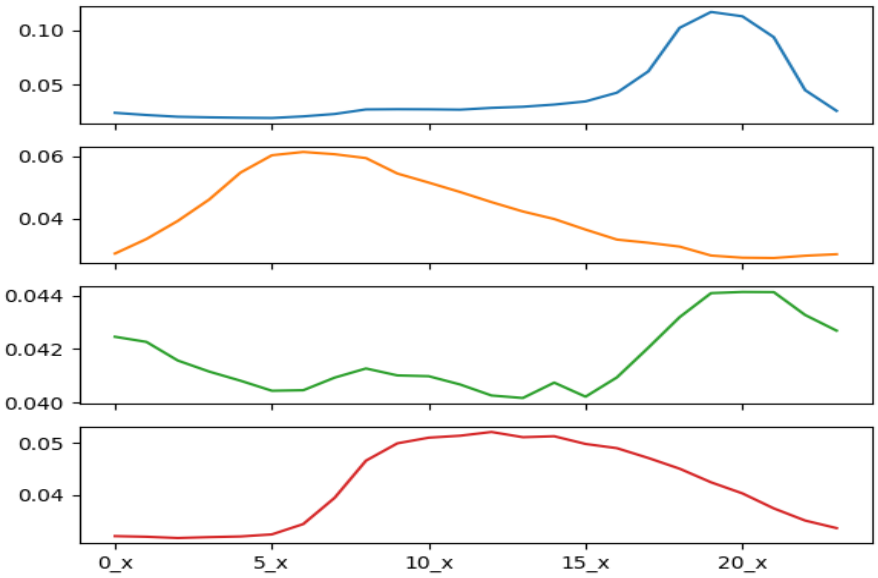



*Figure 7. Time series of a standard working day of April for four zones of Madrid.*

b. The time series of a standard weekend day of April, taking a standard weekend (Saturday 6th and Sunday 7th). Figure 6 shows a sample of the times series obtained for four zones of Madrid. As we can see, the distribution of people throughout the day in the blue, green and red zones is more similar than it was for the previous case (a standard working day). The first one (blue one) presents higher activity during the day times and the first hours of the night, while the second one (orange one) presents higher activity during the early morning, the third one (green one), during the day times, and finally, the fourth one (red one) presents higher activity during the day times and the first hours of the night.

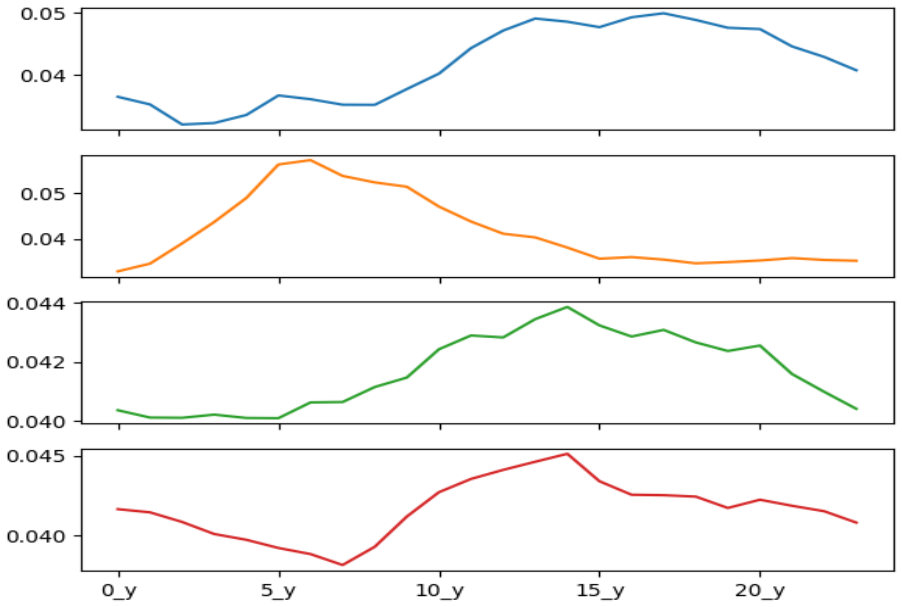



*Figure 8. Time series of a standard weekend day of April for four zones of Madrid.*

c. The junction of the two previous time series, i.e., the time series of a standard working day and a standard weekend day of April. Figure 9 shows a sample of the times series obtained for four zones of Madrid. These time series correspond to the concatenation of the time series of the same colour of Figure 7 and Figure 8.

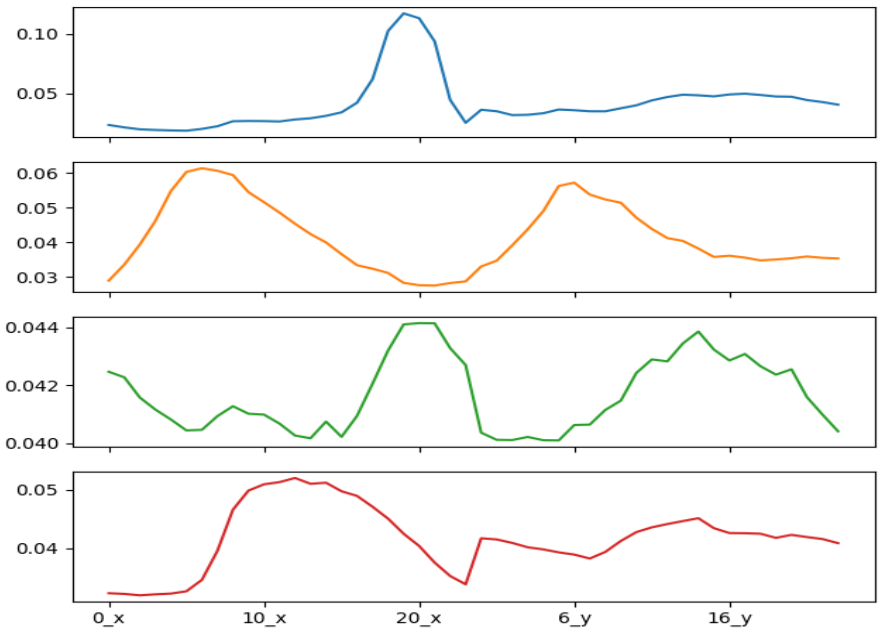


*Figure 9. Time series of a standard working day and a standard weekend day of April for four zones of Madrid.*

Then, we set the presence and overnight stay ratios to -1 and the time series to a vector of zeros for those zones in which they are not defined.

As the visual spatial analysis of the seasonality ratios for both presence and overnight stays cases showed a quite uninformative distribution (see Figure 4 and Figure 5), we performed a statistical analysis of both ratios to see how scattered the values are. As mentioned in Section 2.1.2.1.2.6, we computed the number of zones for which the ratio is defined (count in Figures 8 and 9), its mean (mean in Figures 8 and 9), standard deviation (std in Figure 10 and Figure 11), first, second and third quartile (25%, 50% and 75%, respectively, in Figure 10 and Figure 11) and the minimum and maximum values of the variable (min and max, respectively, in Figure 10 and Figure 11).

The statistical analysis of the seasonality presence ratio, shown in Figure 10, showed that this feature is not very informative, since more than half of its values lie in the interval [0.41, 0.55] (first and third quartile). This implies that for more than half of the zones there is almost no variation between the proportion of visitors present on April 6 and April 20. Moreover, this ratio is only defined for 498 out of the 688 zones.

```
Statistical description of seasonality presence ratio
count    498.000000
mean       0.481244
std        0.175432
min        0.000000
25%        0.413180
50%        0.486309
75%        0.553020
max        1.000000
Name: seasonality_ratio_normalized, dtype: float64
```

*Figure 10. Statistical analysis of the seasonality presence ratio.*

The statistical analysis of the seasonality overnight stays ratio, shown in Figure 11, showed that this feature is not very informative either, since more than half of its values lie in the interval [0.45, 0.55]. This implies that for more than half of the zones there is almost no variation between the proportion of visitors that spent the night on April 6 and April 20. Moreover, this ratio is only defined for 466 out of the 688 zones. Given this information, we decided not to consider this ratio for the analysis.

```
Statistical description of seasonality_overnight_norm ratio
count    466.000000
mean       0.500322
std        0.164376
min        0.000000
25%        0.453074
50%        0.503018
75%        0.549244
max        1.000000
Name: seasonality_overnight_norm, dtype: float64
```

*Figure 11. Statistical analysis of the seasonality overnight stays ratio.*

Finally, analysing the total number of residents and visitors per zone information, we found that out of the 688 zones in which the City of Madrid is divided, 93 have zero visitors during the whole period of the study. As described in Section 2.1.2.1.2.6**,** we set thresholds to remove noisy zones like those by plotting the distribution of zones remaining for different values of the threshold (see Figure 12) and applying the elbow method.

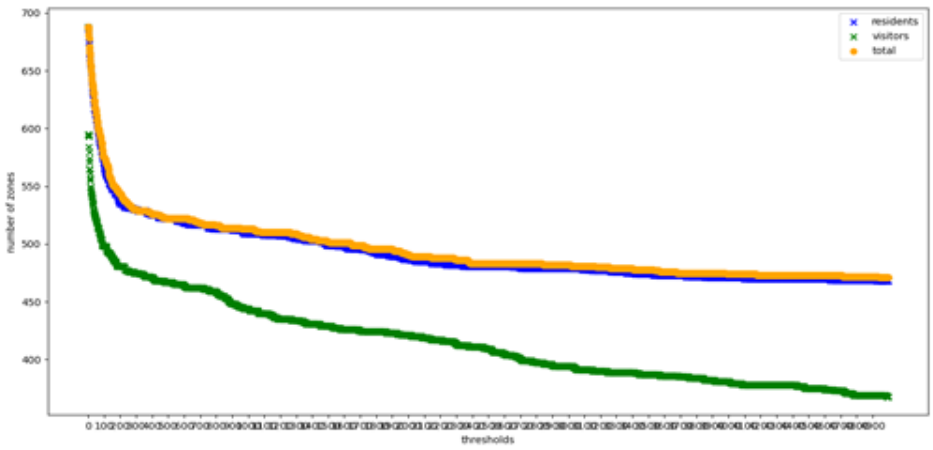



*Figure 12. Relation between the number of zones and the threshold values.*

According to Figure 12, there are elbows for both threshold values at approximately 300, hence, we decided to fix both at 300, which seems to give a good trade-off between the zones and the number of residents and visitors per zone. This way, we kept the zones which have been visited by at least 300 residents and 300 visitors (this translates into 10 residents and 10 visitors per day on average, which we think is a very appropriate and conservative number).

Figure 13 shows the zones remaining before and after applying the residents and visitors threshold. As we can see, the removed zones mostly correspond to forest zones of El Pardo, the Tajo river, and the north of Madrid. As this study aims to analyse the tourism characteristics of each zone, these zones do not seem quite relevant for the present study, as they are barely visited. After this process, 475 zones remained (out of 688).

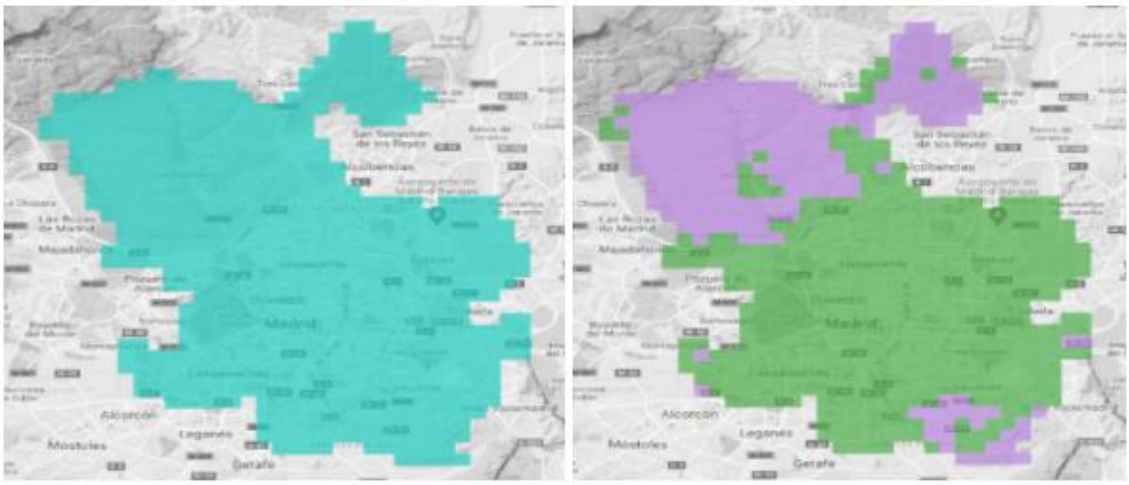


*Figure 13. Zones remaining before and after applying the residents and visitors threshold. On the left, all the initial zones, on the right, in purple the zones removed and in green, the zones kept.*

#### 2.1.2.1.2.9.2 Clustering analysis

Next, we describe in detail the results obtained for each clustering analysis approach.

### K-means clustering

We applied the first clustering approach described in Section 2.1.2.1.2.7, i.e., the application of k-means clustering, to the three sets of variables computed.

### Presence variables

In the first place, we analysed the zones by means of the presence variables. We run the algorithm with different values of *k* and plot in a graphic the inertia score for each value of k in order to apply the elbow method to select the optimal number of clusters. The results obtained are depicted in Figure 14.

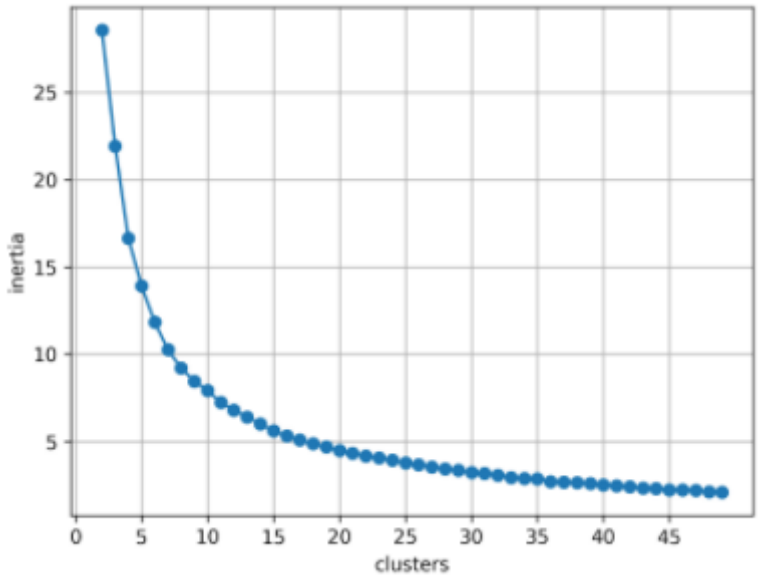



*Figure 14. Inertia values for different numbers of clusters for the presence ratios.*

As we can see, there is no clear “elbow” in the graphic, i.e., there is no clear value for *k*, which means that this criterion is not a good method to find the optimal number of clusters for these data.

According to the results obtained for the statistical analysis of the seasonality ratio (Figure 10), we also considered the set of ratios without the seasonality one, i.e., taking just the visitors/residents presence ratio per zone and the foreign visitors/national visitors presence ratio per zone. The results obtained with this combination are depicted in Figure 15.

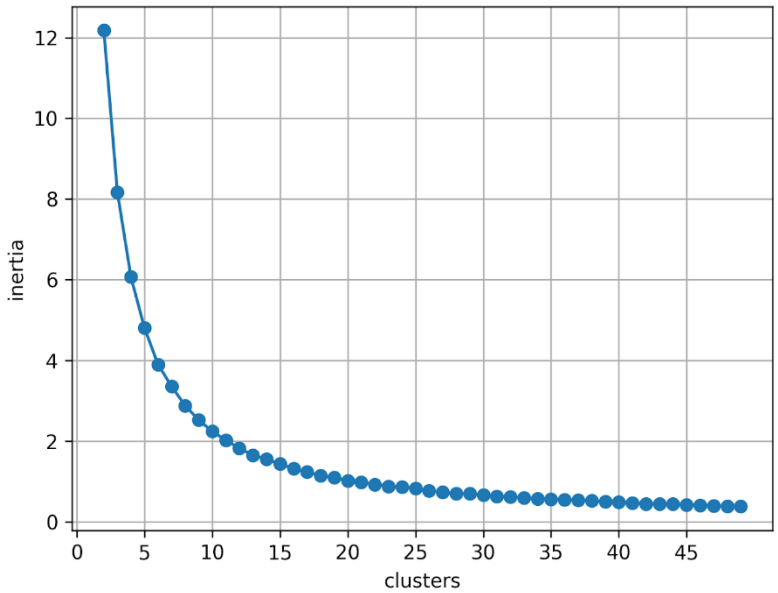



*Figure 15. Inertia values for different numbers of clusters for the presence ratios without the seasonality one.*

Even though with this combination of ratios there is also no clear value for *k*, we see that the values of the inertia for these two ratios are much lower for each value of *k*. This means that these two indicators alone give more structure to the data than the three of them. Thus, we did not consider the seasonality presence ratio per zone for the next clustering analysis.

The results obtained for the overnight stay ratios and the time series are pretty similar: there is neither a clear value for *k* in these cases.

According to these results and in the absence of a clear value for *k* in any of the cases, we considered another approach and applied a two level clustering, in order to capture the structure of the data more appropriately.

### Two levels clustering: K-means + hierarchical clustering

In light of the clustering results obtained following the previous approach, we decided to follow a two-level clustering approach as described in Section 2.1.2.1.2.7. We applied this approach to the three sets of variables:

- Presence variables: visitors/residents presence ratio per zone and foreign visitors/national visitors presence ratio per zone.
- Overnight stay variables: visitor/resident overnight stays ratio per zone and foreign visitor/national visitor overnight stays ratio per zone.
- hourly presence variables:
    - the time series of a standard working day of April, taking three standard working days (Tuesday 2nd, Wednesday 3rd and Thursday 4th),
    - the time series of a standard weekend day of April, taking a standard weekend (Saturday 6th and Sunday 7th),
    - the junction of the two previous time series, i.e., the time series of a standard working day and a standard weekend day of April.
    - According to the graphics obtained for the inertia values with the previous clustering approach, we took 25 clusters for the k-means algorithm in all the cases as it seems a good trade-off for all of them. Next, we describe the results obtained for each set.

### 1. Presence variables

In the first place, we analysed the zones by means of the presence ratios. After applying the k-means algorithm with 25 clusters and identifying each zone with the centroid of the ratios it belongs to, we applied agglomerative clustering to the 25 centroid points.

In order to select the final number of clusters for the agglomerative clustering algorithm, we computed the dendrogram, shown in Figure 14. According to this figure, we took as distance threshold to define the clusters 0.35. With this value, six clusters were obtained. We also considered other threshold values: 0.5, yielding 4 clusters, and 0.22, yielding 9 clusters; but these classifications led to worse results. To analyse and interpret these clusters, we performed a spatial analysis of the clustering classification and we performed a statistical analysis of the values of the ratios used to compute them and the total number of residents, national visitors and foreign visitors per cluster. The spatial analysis and the results obtained using that information are shown in Figure 17 and Table 2.

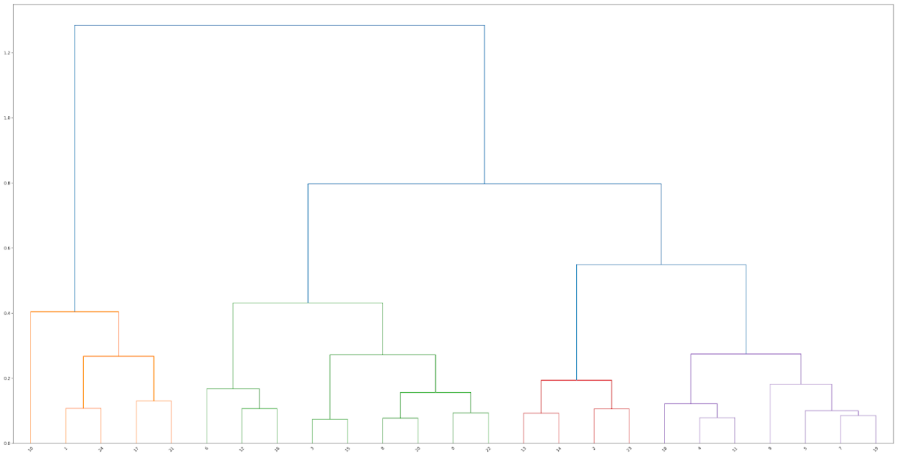


*Figure 16. Dendrogram for the presence ratios.*

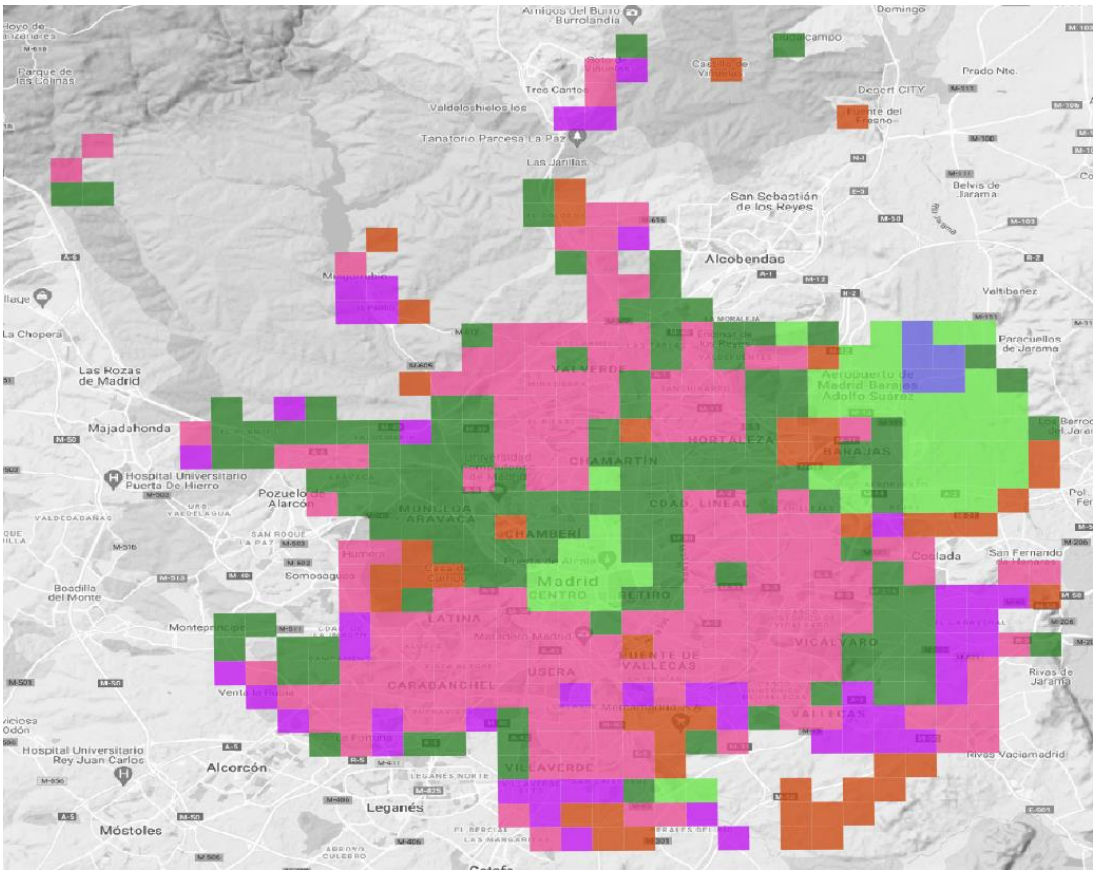



*Figure 17. Clustering classification using the presence ratios, where dark green represents cluster 0, pink represents cluster 1, light green represents cluster 2, orange-brown represents cluster 3, purple represents cluster 4, and blue represents cluster 5.*

*Table 2. Statistical analysis of the presence ratios values within each cluster.*

| Cluster | Characteristics | Visitors vs residents | Foreign vs national visitors | Nº of residents | Nº of national visitors | Nº of foreign visitors |
|---|---|---|---|---|---|---|
| 0 | nº zones | 139 | 139 | 139 | 139 | 139 |
| | mean | 0,49 | 0,47 | 492021,10 | 20735,84 | 23208,18 |
| | std | 0,07 | 0,08 | 736642,75 | 33760,82 | 40551,01 |
| | min | 0,32 | 0,30 | 5877 | 299 | 229 |
| | max | 0,63 | 0,65 | 3225420 | 169875 | 247375 |
| 1 | nº zones | 187 | 187 | 187 | 187 | 187 |
| | mean | 0,31 | 0,34 | 656718,27 | 15616,60 | 9586,52 |
| | std | 0,075 | 0,08 | 525855,37 | 13456,88 | 10361,00 |
| | min | 0,11 | 0,10 | 8948 | 224 | 118 |
| | max | 0,45 | 0,49 | 2695722 | 80952 | 65184 |
| 2 | nº zones | 48 | 48 | 48 | 48 | 48 |
| | mean | 0,78 | 0,63 | 678821,92 | 61105,71 | 114194,19 |
| | std | 0,10 | 0,07 | 1146534,13 | 114446,35 | 213001,38 |
| | min | 0,61 | 0,51 | 963 | 133 | 287 |
| | max | 0,97 | 0,74 | 5043235 | 621134 | 1163051 |
| 3 | nº zones | 47 | 47 | 47 | 47 | 47 |
| | mean | 0,70 | 0,36 | 114517,83 | 13414,00 | 7971,94 |
| | std | 0,07 | 0,06 | 156134,63 | 16862,18 | 9414,69 |
| | min | 0,58 | 0,25 | 1966 | 199 | 120 |
| | max | 0,87 | 0,50 | 656359 | 77574 | 35962 |
| 4 | nº zones | 49 | 49 | 49 | 49 | 49 |

| Cluster | Characteristics | Visitors vs residents | Foreign vs national visitors | Nº of residents | Nº of national visitors | Nº of foreign visitors |
|---|---|---|---|---|---|---|
| | mean | 0,45 | 0,22 | 162038,18 | 7751,04 | 2452,16 |
| | std | 0,07 | 0,044 | 203358,04 | 8898,88 | 2542,71 |
| | min | 0,34 | 0,11 | 4632 | 281 | 68 |
| | max | 0,60 | 0,29 | 925398 | 42950 | 10449 |
| 5 | nº zones | 5 | 5 | 5 | 5 | 5 |
| | mean | 0,91 | 0,92 | 20583,40 | 1354,80 | 15464,20 |
| | std | 0,03 | 0,05 | 25008,36 | 1902,45 | 11332,74 |
| | min | 0,86 | 0,85 | 2647 | 143 | 2146 |
| | max | 0,94 | 0,96 | 64498 | 4728 | 30158 |

From the results obtained, we can establish a preliminary characterization or description of each cluster:

- Cluster 0. Balanced mix of visitors: This cluster presents a balanced distribution of residents, national visitors and international visitors. As can be seen in Table 2, both ratios defined, residents vs visitors and national visitors vs foreign visitors are near 0.5. The zones belonging to this cluster are widely distributed in the city and include areas such as Chamartín and a variety of municipalities of Madrid, such as Vicálvaro, Pozuelo, Aravaca or Barajas, as well as part of Universidad Complutense de Madrid (see Figure 17). These areas are characterised by holding large residential areas but also university and office areas, as well as some tourist attractions. This will be discussed and analysed in the following section.
- Cluster 1. Residents mostly: This cluster presents a predominant presence of resident people. As can be seen in Table 2, the first ratio defined, residents vs visitors, is 0.31. The zones belonging to this cluster are distributed in the north and south of the city and include areas such as Chamartín, Carabanchel, Usera or Vallecas (see Figure 17). These zones are characterised by holding large residential areas and neighborhood shops. This will be discussed and analysed in the following section.
- Cluster 2. National and mostly foreign visitors: This cluster presents a predominant presence of visitors, mostly foreign ones. As can be seen in Table 1, both ratios defined are above 0.6. This cluster mostly comprises the city center and Barajas Airport, as well as El Capricho park (see Figure 17). These areas are characterised by concentrating the most relevant tourist attractions of Madrid and the airport. This will be discussed and analysed in the following section.

- Cluster 3. National visitors: This cluster presents a predominant presence of national visitors. As can be seen in Table 2, the first ratio defined, residents vs visitors, is 0.70, and the second one, national visitors vs foreign visitors, is 0.36. The zones belonging to this cluster are widely distributed in the periphery of the city and include areas such as Barajas, Casa de Campo, Mercamadrid, Valdemingómez recycling plant or the Castillo de Viñuelas (see Figure 17). These zones are characterised by holding large logistic and industrial areas, but also little relevant monuments and points of interest.
- Cluster 4. Balanced mix of residents and national visitors: This cluster presents a balanced distribution of residents and national visitors, with few foreign visitors. As can be seen in Table 2, the first ratio defined, residents vs visitors, is 0.45, and the second one, national visitors vs foreign visitors, is 0.22. The zones belonging to this cluster are widely distributed in the periphery of the city and include areas such as Vallecas, Villaverde or El Pardo (see Figure 17). These areas are characterised by holding residential areas. This will be discussed and analysed in the following section.
- Cluster 5. Foreign visitors: This cluster presents a predominant presence of foreign visitors. As can be seen in Table 2, both ratios defined are above 0.9. The zones belonging to this cluster correspond to the satellite building of Terminal 4 of Barajas airport (T4S) zones (see Figure 17). According to the BARAJAS AIRPOT WEB PAGE, this building “is reserved for all international flights outside the Schengen Area”. These areas are characterised by their large flow of foreign visitors. This will be discussed and analysed in the following section.

### 2. Overnight stay variables

Next, we analysed the zones by means of the overnight stay ratios. After applying the k-means algorithm with 25 clusters and identifying each zone with the centroid of the ratios it belongs to, we applied agglomerative clustering to the 25 centroid points.

*In order to select the final number of clusters for the agglomerative clustering algorithm, we computed the dendrogram, shown in Figure 18. According to this figure, we took as distance threshold to define the clusters 0.65. With this value, seven clusters were obtained. We also considered other threshold values: 0.8, yielding 6 clusters, and 0.49, yielding 8 clusters; but these classifications led to worse results. To analyse and interpret these clusters, we performed a spatial analysis of the clustering classification and a statistical analysis of the values of the same ratios that we used to compute them and the total number of residents, national visitor and foreign visitor overnight stays per cluster. The visual geographical representation and the results obtained using that information are shown in Figure 19 and*

Table 3.

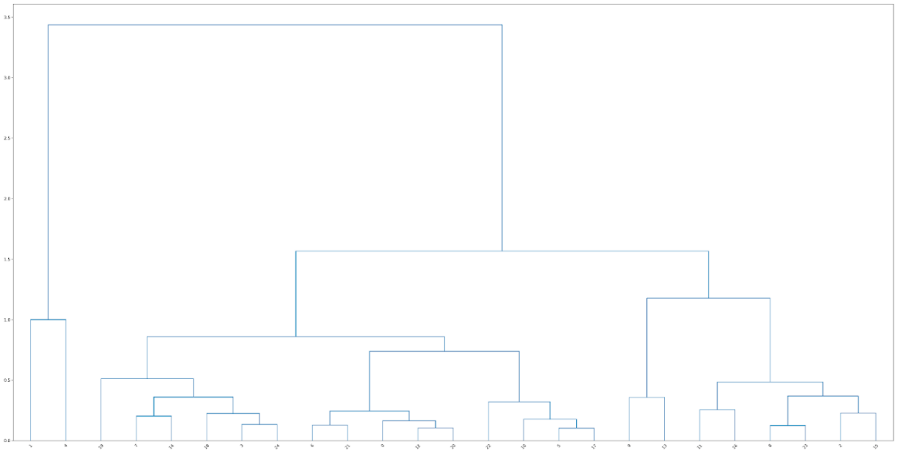


*Figure 18. Dendrogram for the overnight stays ratios.*

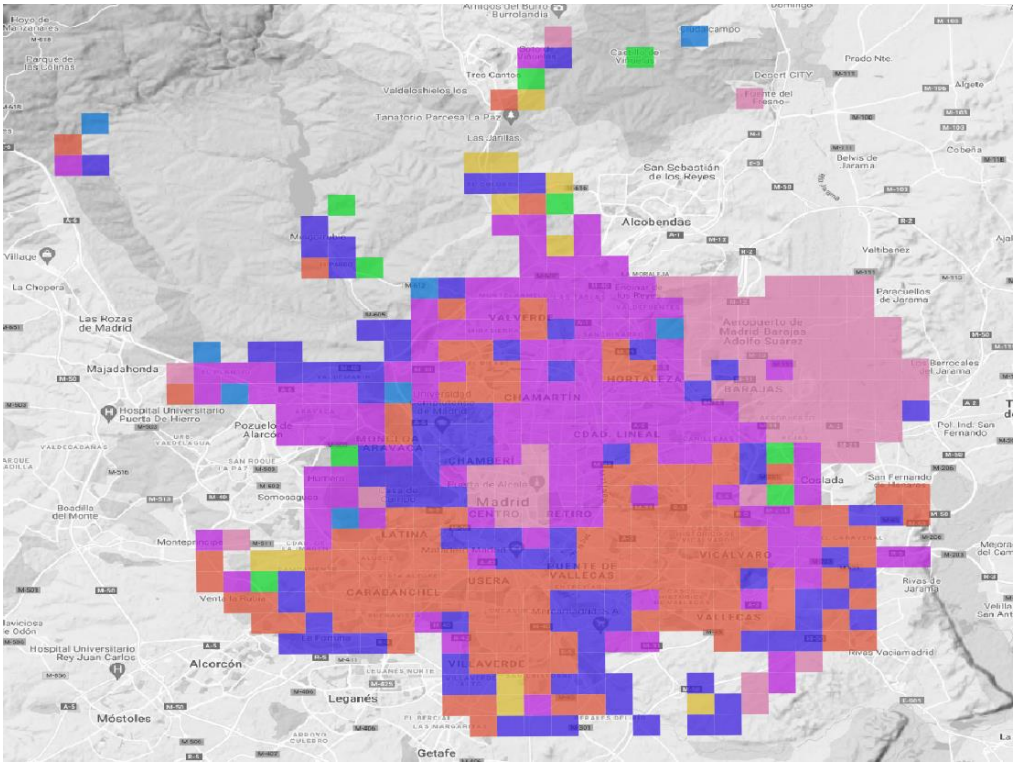



*Figure 19. Clustering classification using the overnight stays ratios, where dark blue represents cluster 0, pink represents cluster 1, orange represents cluster 2, light blue represents cluster 3, yellow represents cluster 4, green represents cluster 5, and purple represents cluster 6.*

*Table 3. Statistical analysis of the overnight stays ratios values within each cluster.*

| Cluster | Characteristics | Visitors vs residents ratio | Foreign vs national visitors ratio | Number of resident overnights | Number of national visitors overnights | Number of foreign visitors overnights |
|---|---|---|---|---|---|---|
| 0 | nº zones | 96 | 96 | 96 | 96 | 96 |
| | mean | 0,57 | 0,26 | 106327,18 | 5123,58 | 1620,27 |
| | std | 0,14 | 0,11 | 213304,74 | 9267,55 | 3140,95 |
| | min | 0,39 | 0 | 0 | 3,72 | 0 |
| | max | 1 | 0,49 | 1174702,23 | 55099,34 | 19221,04 |
| 1 | nº zones | 76 | 76 | 76 | 76 | 76 |
| | mean | 0,79 | 0,72 | 118279,21 | 11381,33 | 17510,50 |
| | std | 0,14 | 0,11 | 277531,85 | 31822,05 | 51060,11 |
| | min | 0,40 | 0,48 | 0 | 0 | 6,78 |
| | max | 1 | 1 | 1415221,64 | 221650,33 | 390258,58 |
| 2 | nº zones | 140 | 140 | 140 | 140 | 140 |
| | mean | 0,30 | 0,26 | 311211,02 | 7600,85 | 1892,67 |
| | std | 0,07 | 0,08 | 244968,98 | 6437,11 | 1801,16 |
| | min | 0,04 | 0 | 148,46 | 2,05 | 0 |
| | max | 0,40 | 0,48 | 972018,19 | 27843,82 | 8087,31 |
| 3 | nº zones | 8 | 8 | 8 | 8 | 8 |

| Cluster | Characteristics | Visitors vs residents ratio | Foreign vs national visitors ratio | Number of resident overnights | Number of national visitors overnights | Number of foreign visitors overnights |
|---|---|---|---|---|---|---|
| | mean | 0,09 | 0,74 | 29524,58 | 124,46 | 164,77 |
| | std | 0,06 | 0,17 | 35938,44 | 196,63 | 264,95 |
| | min | 0,02 | 0,58 | 33,74 | 0 | 0,139994316 |
| | max | 0,16 | 1 | 84964,13 | 478,215136 | 685,666382 |
| 4 | nº zones | 11 | 11 | 11 | 11 | 11 |
| | mean | 0 | -1 | 111,93 | 0 | 0 |
| | std | 0 | 0 | 81,37 | 0 | 0 |
| | min | 0 | -1 | 9,94 | 0 | 0 |
| | max | 0 | -1 | 262,64 | 0 | 0 |
| 5 | nº zones | 9 | 9 | 9 | 9 | 9 |
| | mean | -1 | -1 | 0 | 0 | 0 |
| | std | 0 | 0 | 0 | 0 | 0 |
| | min | -1 | -1 | 0 | 0 | 0 |
| | max | -1 | -1 | 0 | 0 | 0 |
| 6 | nº zones | 135 | 135 | 135 | 135 | 135 |
| | mean | 0,46 | 0,48 | 205699,84 | 8143,86 | 5220,63 |
| | std | 0,08 | 0,08 | 242531,82 | 11084,05 | 7744,10 |

| Cluster | Characteristics | Visitors vs residents ratio | Foreign vs national visitors ratio | Number of resident overnights | Number of national visitors overnights | Number of foreign visitors overnights |
|---|---|---|---|---|---|---|
| | min | 0,23 | 0,35 | 74,13 | 2,63 | 2,06 |
| | max | 0,62 | 0,73 | 1073389,29 | 49211,62 | 45771,08 |

First of all, looking at Table 3 we notice that there are some clusters with mean equals to 0 or -1. As we explained in Section 2.1.2.1.2.6, we selected the zones for the study by means of the number of residents and visitors per zone using the presence indicators, so, in this case there are zones where the number of overnight stays is zero for some or all the total values (as columns “Number of resident overnights”, “Number of national visitors overnights” and “Number of foreign visitors overnights” show). In these cases, the ratios are not defined, hence, we defined them to be -1. The algorithm perfectly identifies those zones with no visitor overnight stays and with no overnight stays and groups them together (clusters 4 and 5, respectively). Looking at the geographical representation of the clusters (Figure 19), we notice that these zones correspond to park or forest zones or zones with little relevant monuments (such as the Castillo de Viñuelas) and forest, which is perfectly consistent with the ratios values.

From the results obtained, we can establish a preliminary characterization of each cluster:

- Cluster 0. Balanced mix of resident and national visitor overnight stays: This cluster presents a balanced distribution of resident and national visitor overnight stays, with few foreign visitors. As can be seen in Table 3, the first ratio defined, residents vs visitors, is 0.57, and the second one, national visitors vs foreign visitors, is 0.26. The zones belonging to this cluster are widely distributed in the city and include areas such as Mercamadrid, El Pardo and Mingorrubio, Universidad Complutense de Madrid and its residences, and the residences of Universidad Autónoma de Madrid (see Figure 19). These areas are characterised by holding logistic, industrial and educational areas, but also residential zones. This will be discussed and analysed in the following section.
- Cluster 1. National and mostly foreign visitor overnight stays: This cluster presents predominant visitors, mostly foreign ones, overnight stays. As can be seen in Table 3, both ratios defined, residents vs visitors and national visitors vs foreign visitors, are above 0.7. This cluster comprises the city center and Barajas Airport (see Figure 19). These areas are characterised by concentrating the most relevant tourist attractions of Madrid and the airport. This will be discussed and analysed in the following section.
- Cluster 2. Resident overnight stays mostly: This cluster presents predominant overnight stays of resident people. As can be seen in Table 3, the first ratio defined, residents vs visitors, is 0.30. The zones belonging to this cluster are widely distributed in the south of the city, and to a lesser extent in the north, and include areas such as Carabanchel, Usera, Vicálvaro or Hortaleza (see Figure 19). These zones are characterised by holding large residential areas and neighborhood shops. This will be discussed and analysed in the following section.
- Cluster 3. Resident overnight stays mostly: This cluster presents predominant overnight stays of residents. As can be seen in Table 3the first ratio defined, residents vs visitors, is 0.09. The zones belonging to this cluster are widely distributed in the periphery of the

city and in other municipalities such as Torrelodones, Ciudalcampo or Majadahonda (see Figure 19). These zones are characterised by holding residential and forest/park areas. This will be discussed and analysed in the following section.

- Cluster 4. No visitor overnight stays: This cluster presents no visitor overnight stays and very few resident overnight stays. As can be seen in Table 3, the first ratio defined is 0 and the second one, -1. The zones belonging to this cluster are distributed in the border of the city and include areas such as the surroundings of Universidad Autónoma de Madrid or San Cristobal industrial area (see Figure 19). These zones are characterised by holding industrial areas, but also road and rail, and forest areas. This will be discussed and analysed in the following section.
- Cluster 5. No overnight stays: This cluster presents no overnight stays. As can be seen in Table 3, both ratios computed are -1. The zones belonging to this cluster are distributed in the periphery of Madrid and other municipalities such as Tres Cantos and El Pardo (see Figure 19). These zones are characterised by holding large forest and park areas, but also little relevant monuments (such as the Castillo de Viñuelas). This will be discussed and analysed in the following section.
- Cluster 6. Balanced mix of overnight stays: This cluster presents a balanced distribution of overnight stays of residents, national visitors and international visitors. As can be seen in Table 3, both ratios defined, residents vs visitors and national visitors vs foreign visitors are near 0.5. The zones belonging to this cluster are widely distributed in the north of the city, and to a lesser extent in the south, and include areas such as Chamartín, El Pilar or Mirasierra (see Figure 19). These areas are characterised by holding large residential areas but also office areas. This will be discussed and analysed in the following section.

## 3. Hourly presence variables

Finally, we analysed the zones by means of their hourly presence behaviour. We applied the clustering process to each of the three time series. For all cases, the process followed is the same as before, i.e., the application of the k-means algorithm with 25 clusters and the identification of each zone with the centroid of the ratios it belongs to, and finally the application of agglomerative clustering to the 25 centroid points. The selection of the final number of clusters is done by means of the dendrogram obtained for each case. Once the number of clusters is chosen and the clusters are computed, a spatial analysis of the clustering results is performed to visualize the distribution in clusters of the zones and analyse the quality and consistency of the classification.

Next, we describe the results obtained for each one of the time series.

### *3.1. The time series of a standard working day of April, taking three standard working days (Tuesday 2nd, Wednesday 3rd and Thursday 4th).*

The dendrogram obtained appears in Figure 20. According to this figure, we took as distance threshold to define the clusters 0.1. With this value, nine clusters were obtained. We also considered other threshold values: 0.08, yielding 10 clusters, and 0.12, yielding 7 clusters; but these classifications led to worse results. To analyse and interpret these clusters, we performed a spatial analysis of the clustering classification, shown in Figure 21.

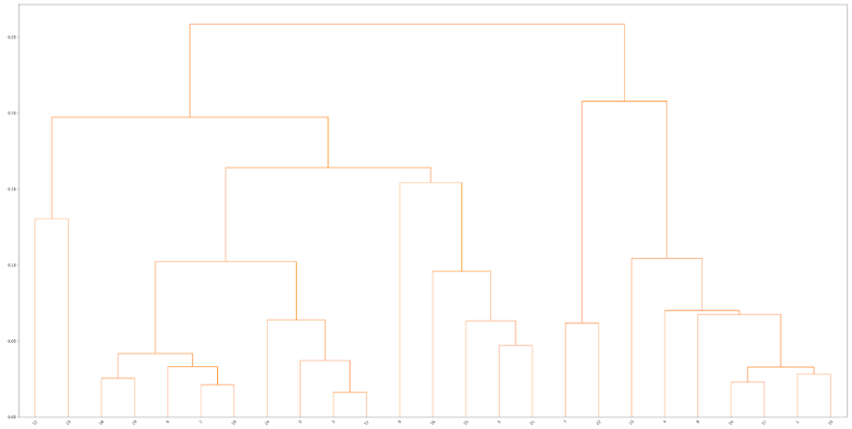


*Figure 20. Dendrogram for the time series of a standard working day of April.*

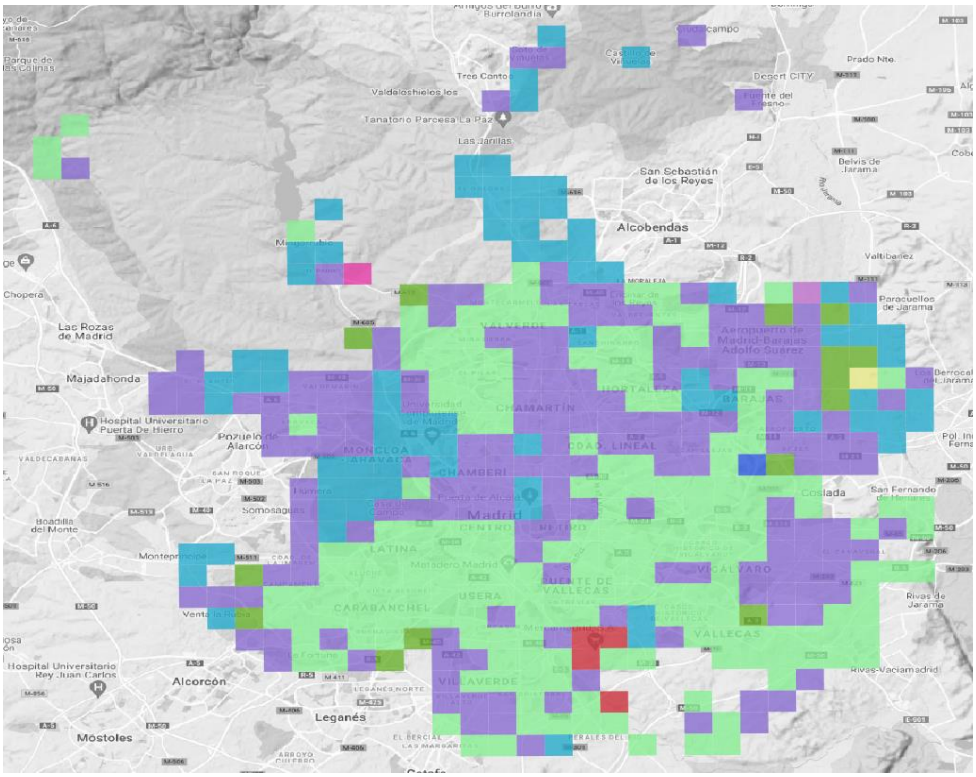



*Figure 21. Clustering classification using the time series of a standard working day of April, where dark green represents cluster 0, light blue represents cluster 1, red represents cluster 2, light green represents cluster 3, dark blue represents cluster 4, light purple represents cluster 5, yellow represents cluster 6, pink represents cluster 7, and purple represents cluster 8.*

To further analyse and interpret each cluster, we plotted the mean (filled with the standard deviation) of the time series of the zones belonging to the same cluster. These time series are shown in Figure 22.

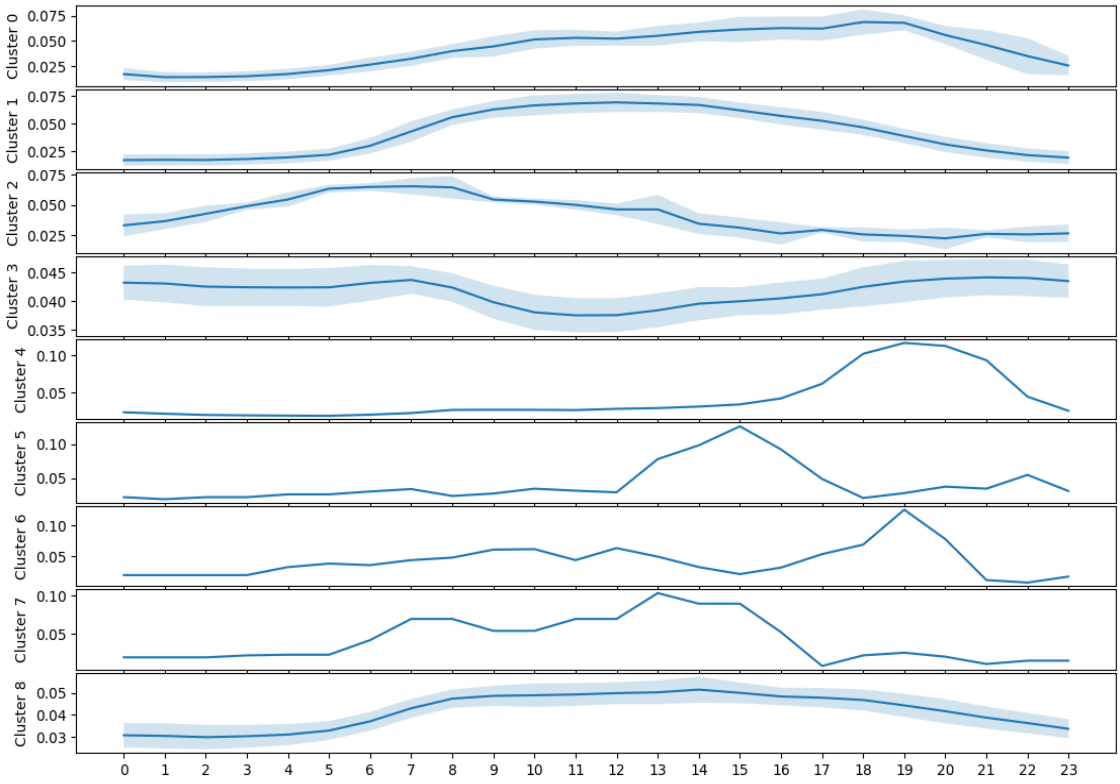



*Figure 22. Mean and standard deviation of the time series of a standard working day of April be-longing to each cluster.*

As we can see in Figure 21, there are 4 clusters that only contain one zone (clusters 4, 5, 6, 7). These zones have a very particular and distinct behaviour from the other zones.

- Cluster 4 corresponds to the Wanda Metropolitano Stadium zone, the stadium of the Atlético de Madrid football team. This zone has almost no presence during the day, except on the days of football matches, when the presence increases at the time of the match. This is the case of 2nd April 2019 (one of the days taken to compute the time series), where there was a Champions football match. As the time series (Figure 22) of this cluster shows, there are almost no people along the day, but between 6pm and 9pm this zone presents higher activity, corresponding to the hours of the match. We see that this zone is the only one with this behaviour.
- Cluster 5 corresponds to an airport logistic zone where there is a peak of presence between 1pm and 4pm, as shown in Figure 22.
- Cluster 6 corresponds to an airport logistic zone near Terminal 4 where there is a peak of presence between 6pm and 8pm, as shown in Figure 22.
- Cluster 7 corresponds to a restaurant and a recreational area at the entrance of El Pardo where people can go hiking and spend the day. The peak of presence, shown in Figure 20, may correspond to the Sun hours and with the restaurant opening hours.

Regarding the other clusters, we observe the following:

- Cluster 0. Logistic, industrial and commercial areas: This cluster presents higher activity between 10am and 9pm (see Figure 22). The temporal profile is consistent with what is shown by the geographical representation. According to Figure 21, this cluster consists of zones of the periphery of Madrid with isolated polygons and commercial areas (malls),

such as Islazul and La Gavia, sport zones, such as a riding club, and logistic zones of Vallecas. It also contains a few zones near the airport. These zones are filled during the working hours of the polygon areas and the opening hours of the malls or sport clubs, as Figure 22shows.

- Cluster 1. University: This cluster presents a higher activity between 8am and 7pm (see Figure 22). The temporal profile is consistent with what is shown by the geographical representation. According to Figure 19, this cluster contains Universidad Autónoma de Madrid, Universidad Complutense de Madrid and Universidad Politécnica de Madrid, as well as city center zones and sport clubs (of different kinds of the ones belonging to cluster 0), such as golf clubs. In these places the presence throughout the day is pretty uniform, for instance, from the beginning of the morning lessons until the end of the afternoon lessons in universities. While the presence during the night drops sharply, as Figure 22shows.
- Cluster 2. Mercamadrid area: This cluster presents a higher activity during the early morning, from 4am to 10am (see Figure 22). The temporal profile is consistent with what is shown by the geographical representation. According to Figure 21, this cluster contains Mercamadrid, where the peak of activity is early in the morning, as Figure 22 shows.
- Cluster 3. Residential: This cluster presents higher activity during the early morning and night times (see Figure 22). The temporal profile is consistent with what is shown by the geographical representation. According to Figure 21, this cluster contains neighborhoods such as Carabanchel, Usera and Vallecas. These are residential neighbourhoods, and many of their inhabitants move to other zones of Madrid to work (mostly zones belonging to cluster 8). These zones are emptied during the working hours, and filled later when people return home, as Figure 22shows.
- Cluster 8. Residential and office: This cluster presents higher activity during the day (from 7am to 8pm) (see Figure 22). The temporal profile is consistent with what is shown by the geographical representation. According to Figure 21, this cluster contains neighborhoods such as Chamartín and Chamberí, as well as the city centre, with many tourist attractions of Madrid. These are mainly residential and working neighbourhoods, with many offices and shops, and tourist zones. These zones receive many workers from other zones and cities of Madrid, for instance, from people living in cluster 3 and cities such as Móstoles, Alcorcón, Tres Cantos, etc. These zones are filled during the working hours, as Figure 22 shows.

We will deepen this analysis in the next section.

### *3.2. The time series of a standard weekend day of April, taking a standard weekend (Saturday 6th and Sunday 7th).*

The dendrogram obtained appears in Figure 23. According to this figure, we took as distance threshold to define the clusters 0.09. With this value, 11 clusters were obtained. We also considered other threshold values: 0.07, yielding 13 clusters, and 0.12, yielding 7 clusters; but these choices led to worse classification results. To analyse and interpret these clusters, we performed a spatial analysis of the clustering classification, shown in Figure 24.

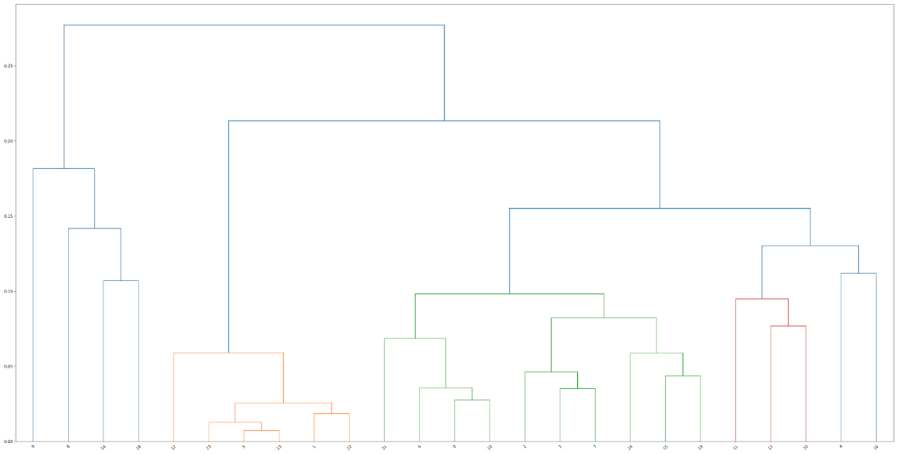


*Figure 23. Dendrogram for the time series of a standard weekend day of April.*

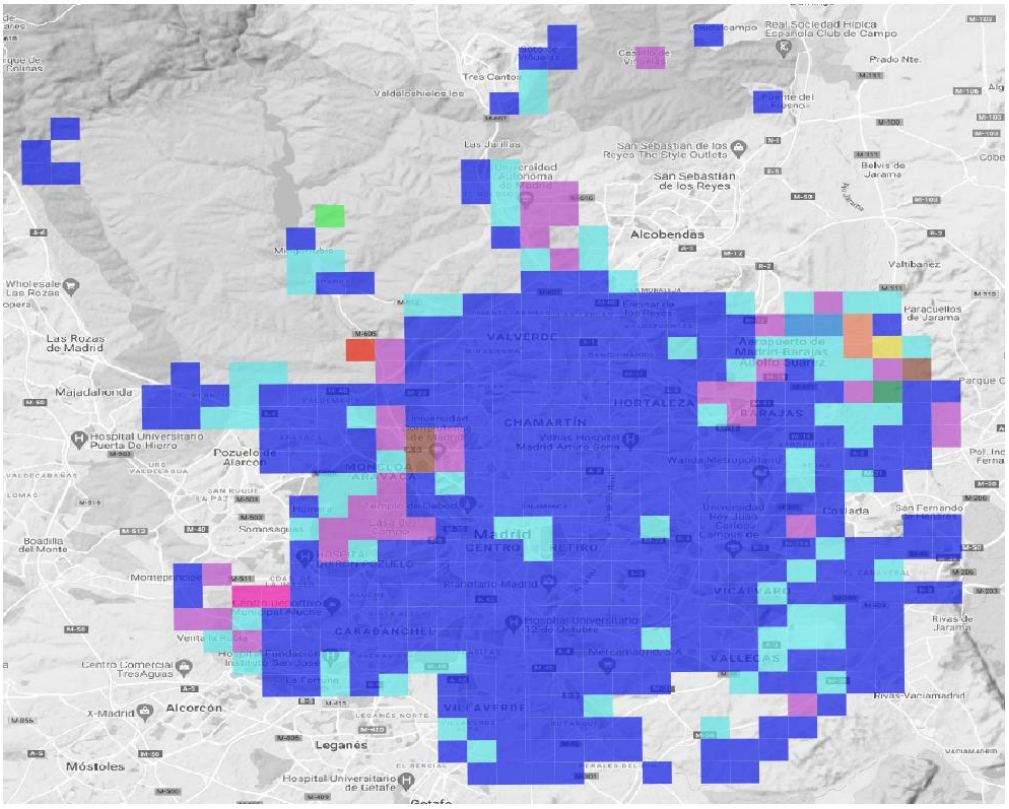



*Figure 24. Clustering classification using the time series of a standard weekend day of April, where purple represents cluster 0, light blue represents cluster 1, orange represents cluster 2, yellow represents cluster 3, dark green represents cluster 4, pink represents cluster 5, dark blue represents cluster 6, brown represents cluster 7, light green rep-resents cluster 8, red represents cluster 9, and blue represents cluster 10.*

As we can see, despite obtaining 11 clusters, the clustering algorithm is unable to separate among zones with these data. We also plotted the mean (filled with the standard deviation)

of the time series of the zones belonging to the same cluster. These time series are shown in Figure 25.

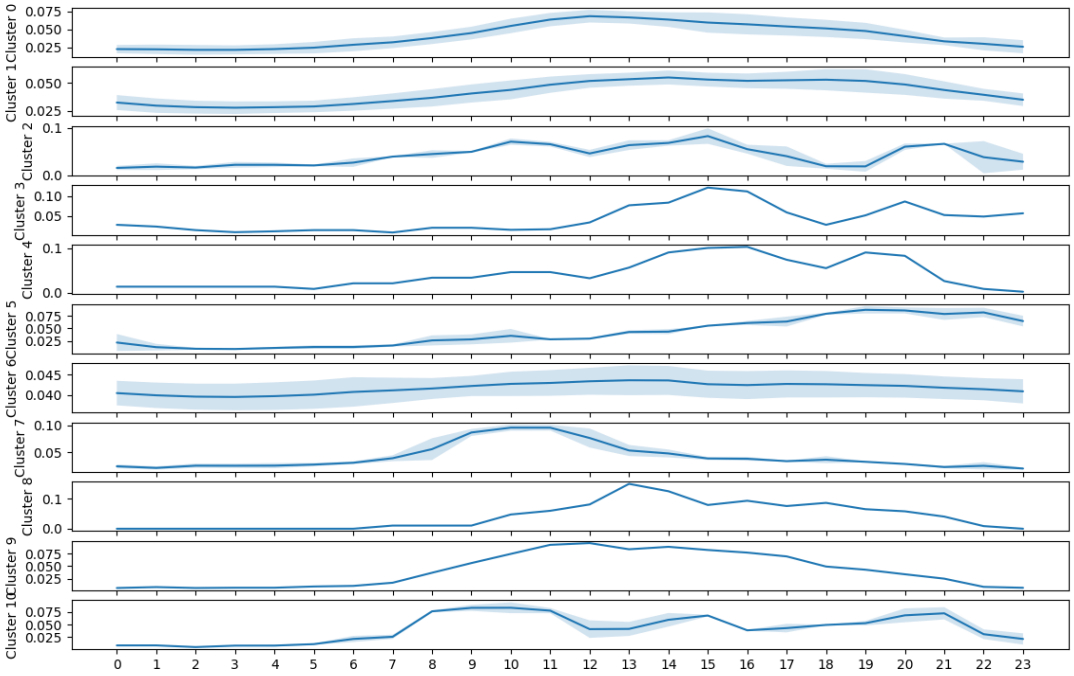



*Figure 25. Mean and standard deviation of the time series of a standard weekend day of April be-longing to each cluster.*

As we can see in Figure 25, the time series profile of clusters 0 and 1 are similar, while the standard deviation of the time series in cluster 6 is very large. Clusters 2, 3, 4, 5, 7, 8, 9, 10 contain few zones.

Reducing the threshold value would lead to a larger number of clusters, complicating the interpretation of the results, and yielding an unrepresentative and uninformative number of clusters. Hence no further analysis for this case is carried out.

### *3.3. The junction of the two previous time series, i.e., a time series of a standard working day and a standard weekend day of April.*

The dendrogram obtained appears in Figure 26. According to this figure, we took as distance threshold to define the clusters 0.15. With this value, nine clusters were obtained. We also considered other threshold values: 0.1, yielding 14 clusters, and 0.2, yielding five clusters; but these classifications led to worse results. To analyse and interpret these clusters, we performed a spatial analysis of the clustering classification, shown in Figure 27.

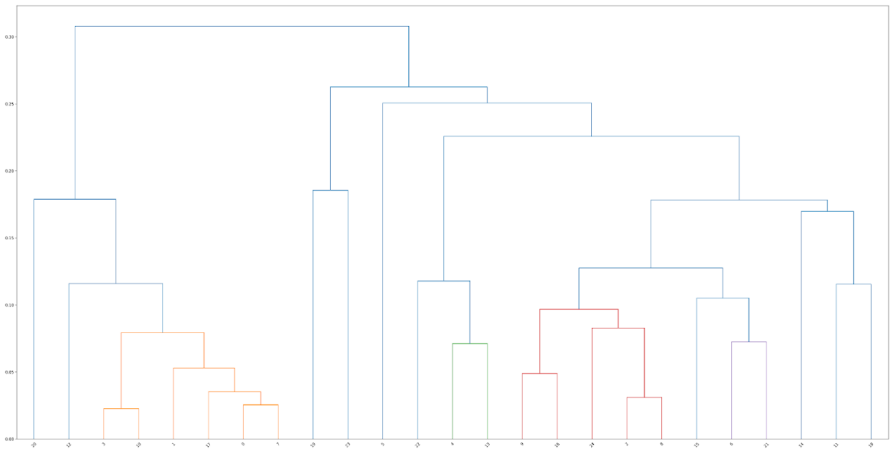


*Figure 26. Dendrogram for the time series of a standard working day and a standard weekend day of April.*

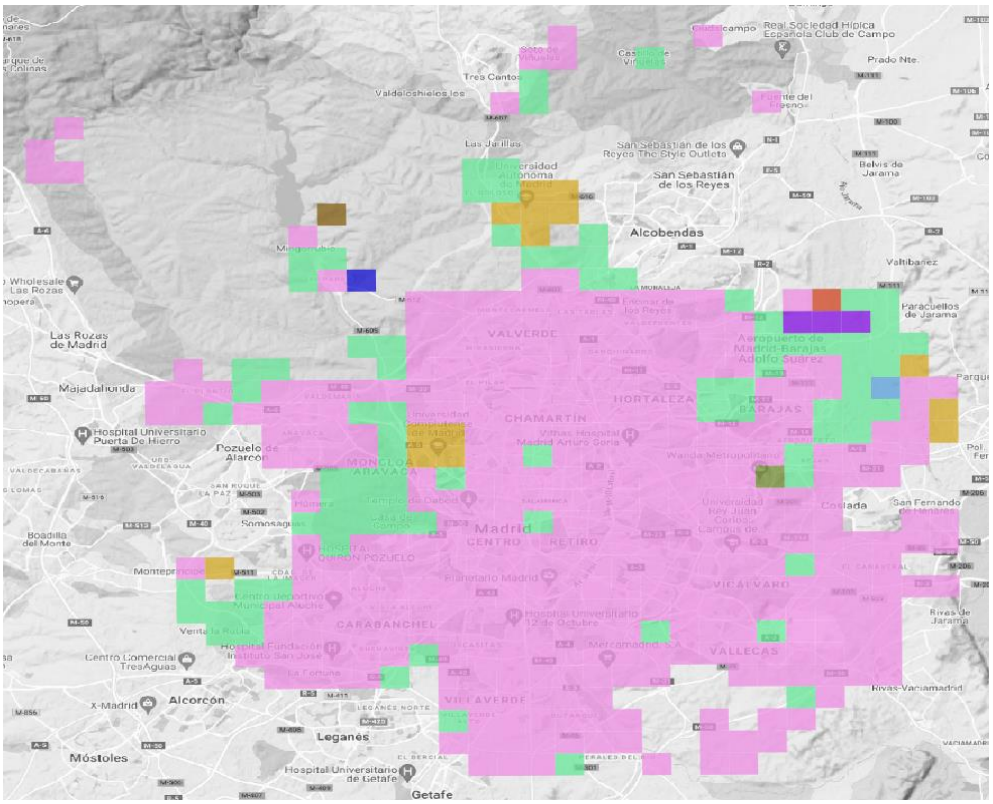



*Figure 27. Clustering classification using the time series of a standard working day and a stand-ard weekend day of April, where light green represents cluster 0, orange represents cluster 1, pink represents cluster 2, purple represents cluster 3, light blue represents cluster 4, dark green represents cluster 5, dark blue represents cluster 6, brown rep-resents cluster 7, and red represents cluster 8*

As we can see, the clustering algorithm is unable to separate among zones with these data as well. It barely separates university zones and Mercamadrid, for instance. The results are

similar to those of case 3.2, which make sense since this case combines the data of cases 3.1 and 3.2. We also plotted the mean (filled with the standard deviation) of the time series of the zones belonging to the same cluster. These time series are shown in Figure 28.

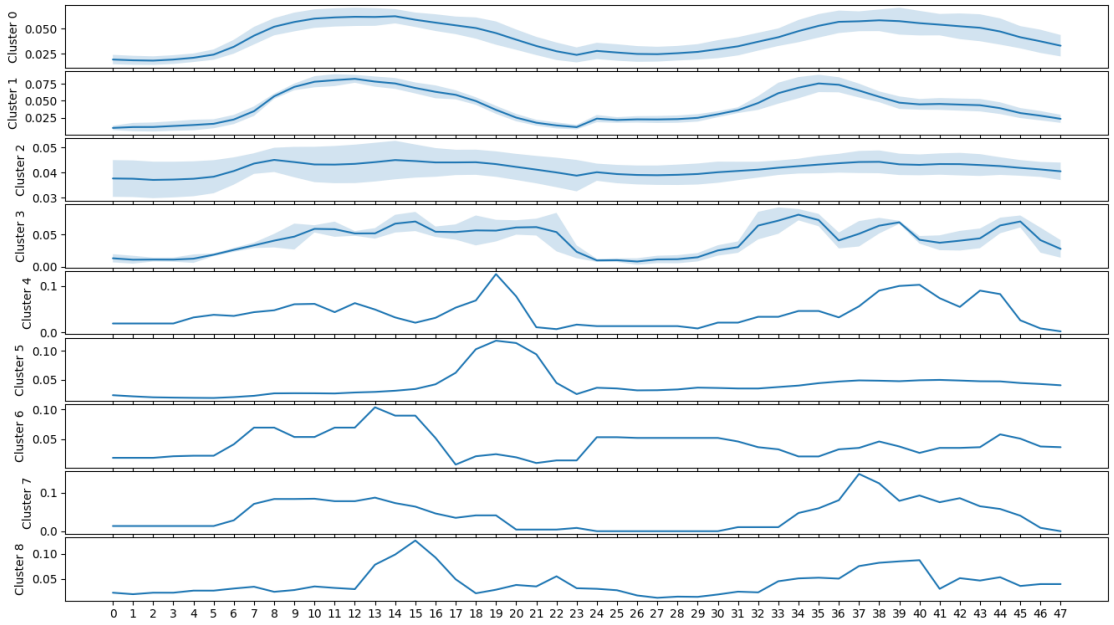



*Figure 28. Mean and standard deviation of the time series of a standard working day and a standard weekend day of April.*

No further analysis for this case is carried out.

As we can see, the only time series that gave good results are the ones corresponding to a normal working day, which really led to meaningful results. We were not able to obtain informative results for a weekend day, either independently (case 3.2) or together with a working day (case 3.3). This kind of day may be more disruptive and this approach does not enable to characterize the zones. Hence, this study is restricted to analysing the hourly presence on working days.

### 2.1.2.1.2.9.3 Interpretation of the results

The final results kept are the ones obtained with the presence and overnight stay ratios and with the time series of a standard working day of April, taking three standard working days (case 3.1). This way, we have a triple characterization of the zones of Madrid based on the kind of persons that are present or spend the night in the zones and the hourly presence.

As mentioned in Section 2.1.2.1.2.7 in order to interpret the clustering results obtained, we used two kinds of information:

a. the location of a set of points of interests (POIs) of Madrid, and
b. the number of hotels, pubs, guest houses and airbnbs located in each zone.

The selected POIs, together with the cluster they belong to are shown in Table 4.

*Table 4. Cluster location of the selected POIs for the three clustering classifications.*

| Type attribute | Name | Presence cluster | Overnight stays cluster | Hourly presence cluster |
|---|---|---|---|---|
| Museums | Prado National Museum | 2 | 1 | 8 |
| | Thyssen-Bornemisza Museum of Art | 2 | 1 | 1 |
| | Reina Sofia Museum | 2 | 1 | 8 |
| | Museum Sorolla | 2 | 1 | 8 |
| | National Archaeological Museum | 2 | 1 | 1 |
| Tourist attractions | Retiro Park | 0,2 | 1,6 | 1,8 |
| | Royal Palace of Madrid | 2 | 1 | 8 |
| | Santiago Bernabéu Stadium | 2 | 6 | 8 |
| | Gran Via | 2 | 1 | 8 |
| | Plaza Mayor | 2 | 1 | 8 |
| | Plaza de Cibeles | 2 | 1 | 1 |
| | Neighbourhood Salamanca | 0,2 | 1 | 1 |
| | Debod Temple | 0 | 1 | 8 |
| | El Capricho Park | 2 | 1 | 3 |
| | San Miguel Market | 2 | 1 | 8 |
| | Cibeles Palace | 2 | 1 | 1 |

| | | | | |
|---|---|---|---|---|
| | Plaza Santa Ana | 2 | 1 | 8 |
| | Puerta del Sol | 2 | 1 | 8 |
| | Crystal Palace | 0 | 6 | 8 |
| | Roof of Círculo de Bellas Artes | 2 | 1 | 8 |
| | Puerta Alcalá | 2 | 1 | 1 |
| | Almudena Cathedral | 2 | 1 | 8 |
| | Atocha Station | 0 | 1 | 8 |
| | San Antón Market | 2 | 1 | 8 |
| | Platform of Chamberí Station | 2 | 1 | 8 |
| | Plaza de España | 0,2 | 1 | 8 |
| | Neighbourhood La Latina | 2 | 6 | 3 |
| | Botanic Garden | 2 | 1 | 8 |
| | Plaza Callao | 2 | 1 | 8 |
| | Plaza de Oriente | 2 | 1 | 8 |
| | Fuencarral Street | 2 | 1 | 8 |
| | Congress Madrid | 2 | 1 | 8 |
| | Plaza de Colón | 2 | 1 | 1 |
| | Calle de Alcalá/Goya | 0,2 | 1,6 | 1,8 |

| | | | | |
|---|---|---|---|---|
| | Real Theatre/Opera | 2 | 1 | 8 |
| Universities | Universidad Autónoma de Madrid (UAM) | 1 | 0,2 | 1 |
| | Universidad Complutense de Madrid (UCM) | 0 | 0 | 1 |
| | Universidad Politécnica de Madrid (UPM) | 0 | 0 | 1 |
| | Universidad Rey Juan Carlos - Vicálvaro Campus (URJC) | 1 | 2 | 3 |
| Logistic areas | Mercamadrid | 3 | 0 | 2 |
| | Valdemingómez recycling plant | 3 | 0,1 | 3,8 |
| Airport | Airport | 2 | 1 | 0,1,5,6,8 |
| | Airport T4S (satellite building of Terminal 4) | 5 | 1 | 8 |
| Other points of interest | Madrid Río Park | 1 | 0 | 3 |
| | Plaza de Toros de las Ventas | 0 | 6 | 3 |
| | El Rastro | 2 | 1,6 | 3 |
| | Cable Car/Teleférico Madrid | 0 | 6 | 8 |
| | Trade Fair Madrid IFEMA | 3 | 0 | 1 |
| | Attraction Park Madrid | 1 | 2 | 3 |
| | Parque del Oeste/ East Park | 0,3 | 6 | 8 |
| | Planetarium Madrid | 1 | 0 | 3 |
| | Caixa Forum | 2 | 1 | 8 |

| | | | | |
|---|---|---|---|---|
| | Zoo Aquarium of Madrid | 0 | 6 | 8 |
| | Matadero Madrid | 1 | 0 | 3 |
| | Casa de Campo Lake | 0 | 0 | 8 |

Next, we analyse the information of Table 4 within each clustering classification.

### Presence clusters interpretation

According to Table 4, all the museums and the 90% of the tourist attractions, as well as the airport, are located in cluster 2, the tourist one. The only three tourist attractions not located in cluster 2 are the Debod Temple, the Crystal Palace and Atocha station, which are mostly visited by residents and national visitors. These three places are located in cluster 0, together with UCM, UPM and other points of interest such as Plaza de Toros de las Ventas, or the Zoo. These places are also mostly visited by residents and national visitors, in accordance with the initial cluster description.

On the other hand, UAM and URJC are located in cluster 1, as well as other points of interest, such as Madrid Rio Park and El Matadero, which are places typically visited by resident people, in accordance with the residential behaviour of these zones initially identified.

Places like Trade Fair Madrid IFEMA, Mercamadrid and Valdemingómez recycling plant are located in cluster 3, mainly visited by residents and national visitors for business or working purposes.

Finally, airport T4S belongs to cluster 5, as already noticed in the initial cluster description.

We notice that none of these places is located in cluster 4, which makes perfect sense as this cluster is composed of periphery zones of Madrid.

According to this information and to the initial description of the clusters, we can conclude with a final characterization or identification of each cluster:

- Cluster 0: heterogeneous cluster in which the presence of the three kinds of persons is balanced. Working, residential and a bit tourist cluster.
- Cluster 1: residential cluster with mainly national visitors and residents.
- Cluster 2: tourist cluster, with higher visitors presence, mainly foreign ones.
- Cluster 3: logistic cluster, with Mercamadrid and other relevant logistic areas, with mainly residents and national visitors. Cluster with no tourist interest.
- Cluster 4: residential cluster, with mainly resident people.
- Cluster 5: Satellite building of Terminal 4 of Barajas airport, foreign visitors cluster.

### Overnight stays clusters interpretation

According to Table 4, all the museums and 90% of the tourist attractions, as well as the airport and T4S, are located in cluster 1, the tourist one. The only three tourist attractions not located in cluster 1 are the Santiago Bernabéu Stadium, the Crystal Palace and La Latina. These

three places are located in cluster 6, together with the 41.7% of the other points of interest, with places such as Plaza de Toros de las Ventas, East park and the Zoo.

UAM, UPM and UCM are located in cluster 0, as well as the logistic places and the 41.7% of the other points of interest, with places such as IFEMA and Madrid Río. While cluster 2 also contains UAM, together with URJC and the Attraction Park.

Finally, we notice that none of these places is located in clusters 3, 4 or 5, which makes perfect sense as these clusters correspond to zones of Madrid with very little tourist interest and with very few (if any) overnight stays.

In addition to this information, we also considered the distribution of the hotels, guest houses, airbnb accommodations and pubs in each cluster. Table 4 shows the proportion of each of them in each cluster. The last row ("Total number") shows the total number of each of these attributes in Madrid. As we can see, cluster 1 contains the greatest proportion of hotels, guest houses and airbnb accommodations, in perfect accordance with its initial identification as the tourist cluster in overnight stays terms. On the other hand, clusters 2 and 3 contain the greatest proportion of pubs, which makes sense given that these zones are very residential and cover the most inhabited areas in the north and south of Madrid. Clusters 3, 4 and 5 have no accommodations or pubs. This suggests that very few visitors (if any) spend the night in the zones belonging to these clusters, which is consistent with the clusters interpretation made so far, as the zones belonging to these clusters correspond to forest zones mostly with no visitor overnight stays. Finally, in cluster 0 the proportion of pubs stands out, this makes sense given the location of the zones belonging to this cluster, near universities and tourist areas.

*Table 5. Distribution of hotels, guest houses, pubs and airbnb accommodations in each cluster using overnight stays ratios.*

| cluster | hotel | guest house | pub | airbnb |
|---|---|---|---|---|
| 0 | 0,06 | 0,02 | 0,13 | 0,08 |
| 1 | 0,71 | 0,95 | 0,21 | 0,45 |
| 2 | 0,02 | 0,01 | 0,36 | 0,19 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0,21 | 0,02 | 0,30 | 0,28 |

| | | | | |
|---|---|---|---|---|
| Total number | 224 | 254 | 657 | 17236 |

We also notice the huge difference in the number of airbnb accommodations and the other types of accommodations. This difference and the location of the airbnb accommodations affect where visitors spend the night. To deepen this idea, we analysed the distribution of the three kinds of accommodations (hotels, guest houses and airbnbs) within the zones of the same cluster. For that, we computed the number of hotels, guest houses and airbnbs per zone and divided them by the sum of the three values to see the percentage of accommodation per zone that corresponds to each type. Figure 29 shows the boxplots with the results obtained for each cluster.

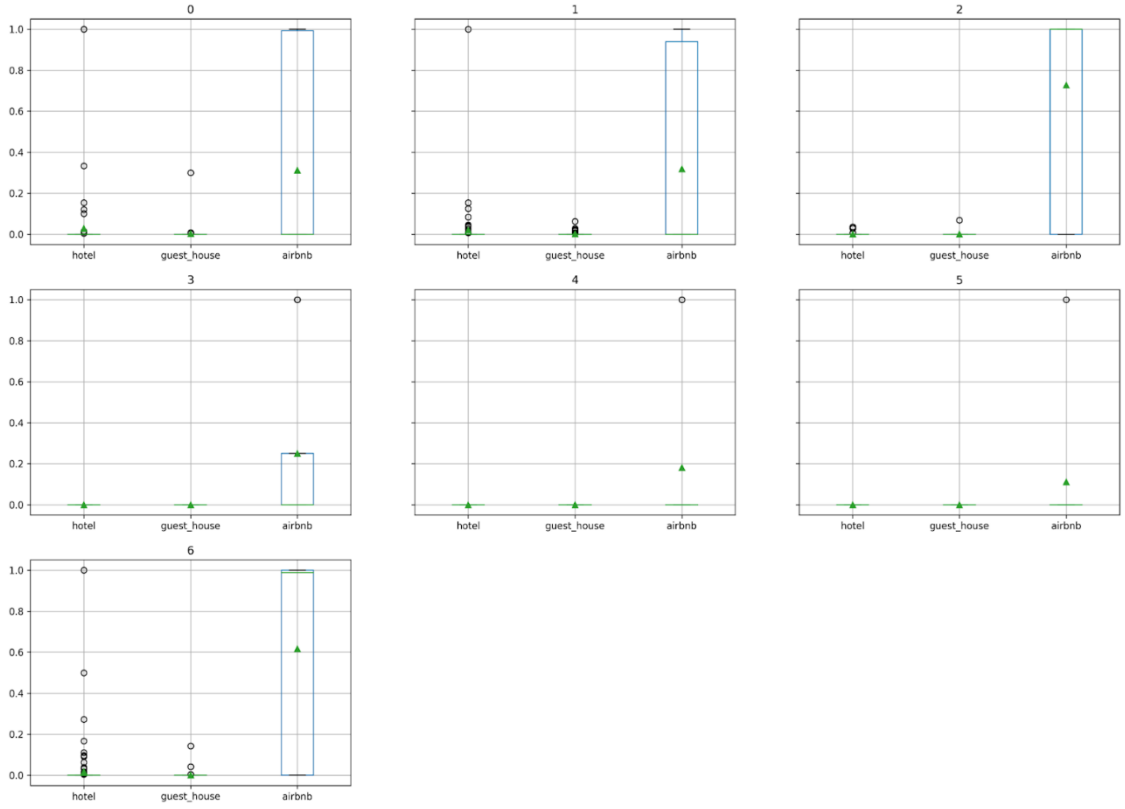



*Figure 29. Boxplots of the distribution of hotels, guest houses and airbnb accommodations in the zones of each cluster using overnight stays ratios. Green triangle shows the mean for each attribute.*

As we can see, the presence of the airbnb accommodations is predominant in each cluster, due to its high number. This may imply that most of the visitors who come to Madrid spend the night in this type of accommodation.

According to this information and to the initial description of the clusters, we can conclude with a final characterization or identification of each cluster:

- Cluster 0: university and logistic cluster, with residents and national visitors overnight stays mainly. Cluster with little tourist interest.
- Cluster 1: tourist cluster, with foreign visitors overnight stays mainly.

- Cluster 2: residential cluster.
- Cluster 3: periphery zones with mostly residents overnight stays. Cluster with no tourist interest.
- Cluster 4: cluster with no visitors overnight stays.
- Cluster 5: cluster with no overnight stays.
- Cluster 6: heterogeneous cluster, with balanced residents, national and foreign visitors overnight stays.

### Hourly presence clusters interpretation

According to Table 4, all the universities of Madrid are located in cluster 1, but URJC. We have to keep in mind that the Vicálvaro Campus of the URJC is in the middle of the Vicálvaro neighborhood, so, the university hourly behavior may be eclipsed by the hourly behaviour of the neighborhood itself. Moreover, IFEMA is also located in cluster 1. Some tourist attractions (Retiro, Cibeles and Salamanca, for instance) and museums also belong to this cluster. In these places, the hourly presence is also quite uniform throughout the day.

Mercamadrid belongs to cluster 2, as mentioned in the initial description of this cluster. This is the only place of Table 4in this cluster.

URJC and 50% of the other points of interest are located in cluster 3. Many of these places are located in residential areas (URJC, the Planetarium, Madrid Río or the Attraction Park, for instance).

Finally, 60% of the museums, 76.6% of the tourist attractions and 41.7% of the other points of interest belong to cluster 8, as well as airport T4S. This makes perfect sense given that this cluster is the most tourist one.

According to this information and to the initial description of the clusters, we can conclude with a final characterization or identification of each cluster:

- Cluster 0: isolated polygon and logistic, commercial (malls) and sports areas.
- Cluster 1: university cluster.
- Cluster 2: Mercamadrid cluster.
- Cluster 3: residential cluster.
- Cluster 4: Wanda Metropolitano Stadium cluster.
- Cluster 5: airport logistic cluster.
- Cluster 6: airport logistic cluster.
- Cluster 7: isolated hotel and recreation area.
- Cluster 8: working and tourist cluster, with offices and commercial zones.

## 2.1.2.1.2.10 Conclusion

Three kinds of variables were computed, presence, overnight stays and hourly presence, in order to characterize the zones of Madrid in terms of the type of person that visits them, namely, residents, national visitors and foreign visitors. As a result, a triple characterization of the zones of Madrid is achieved.

The final variables used to characterize the zones of Madrid are:

- presence variables: visitors/residents presence ratio per zone and foreign visitors/national visitors presence ratio per zone,

- overnight stay variables: visitor/resident overnight stays ratio per zone and foreign visitor/national visitor overnight stays ratio per zone,
- hourly presence variables: the time series of a standard working day of April, taking three standard working days (Tuesday 2nd, Wednesday 3rd and Thursday 4th).

The results obtained allow to group and characterize the zones of Madrid by means of the two kinds of indicators computed. Moreover, the distribution of some relevant zone attributes (hotels, pubs, airbnb accommodation and guest houses) and the location of the selected POIs provide a good and coherent interpretation of the results.

Finally, as we were not able to characterize the zones by means of their time series for weekend days, as future work it could be interesting to try other approaches, such as taking more days to compute the time series.

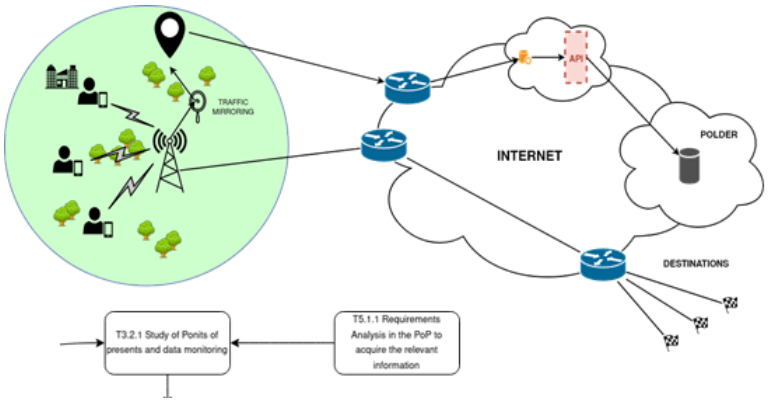
### 2.1.3 Starflow

This deliverable includes the following Starflow activities and tasks in POLDER project:

T5.2.- Data Model where the descriptive and predictive network data models have been defined, analyzed, generated, tested and validated including T5.2.1 Requirement Analysis and T5.2.2 API design

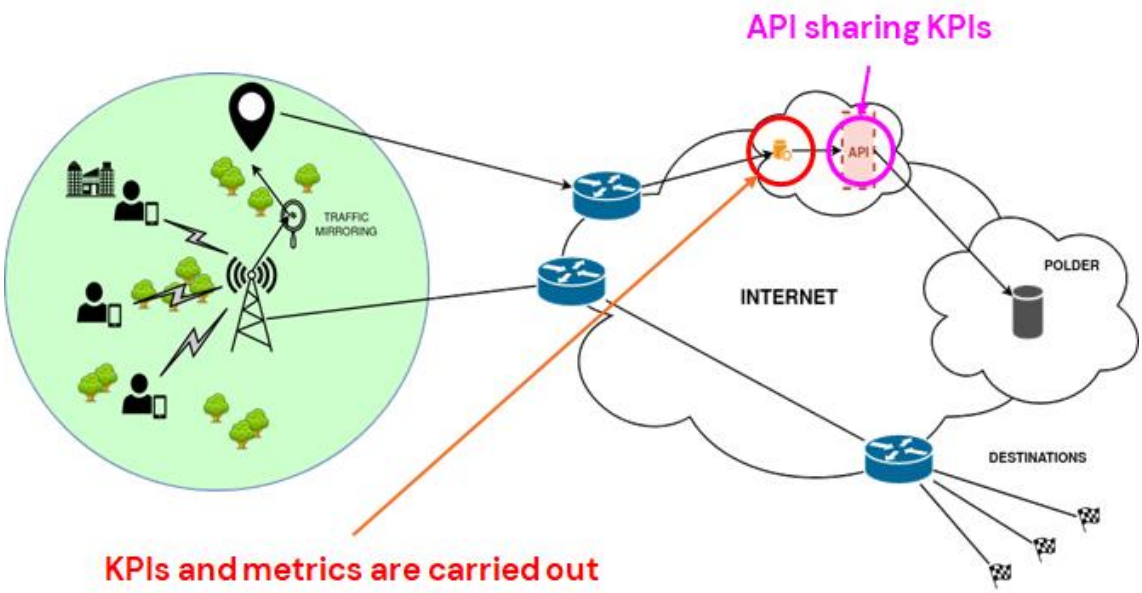
This task provides the data model for the T5.5 Descriptive and Predictive models where the network traffic patterns have been developed

Starflow collects network information from different devices that are connected to a point of presence.




| Recursos | Cantidad |
|---|---|
| Núcleos de CPU | 2 |
| RAM | 2GB |
| Disco | 10GB |

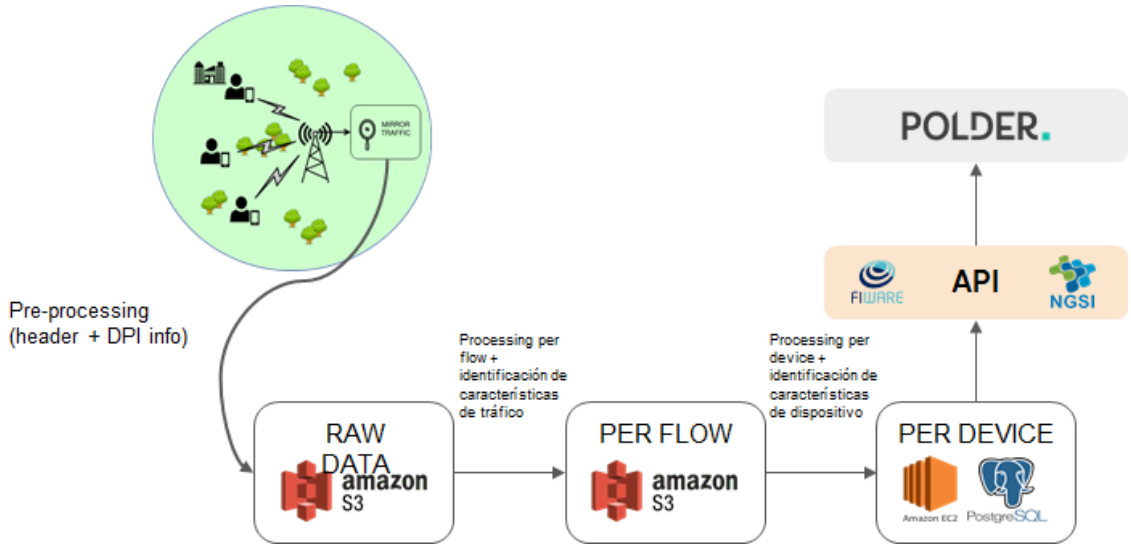
this information is stored in one of the tested cloud services in the project (AWS; GCP, Azure, IBM cloud) guaranteeing future massive deployment and the KPIS's and the metrics are processed in Cloud by Starflow developed AI algorithms



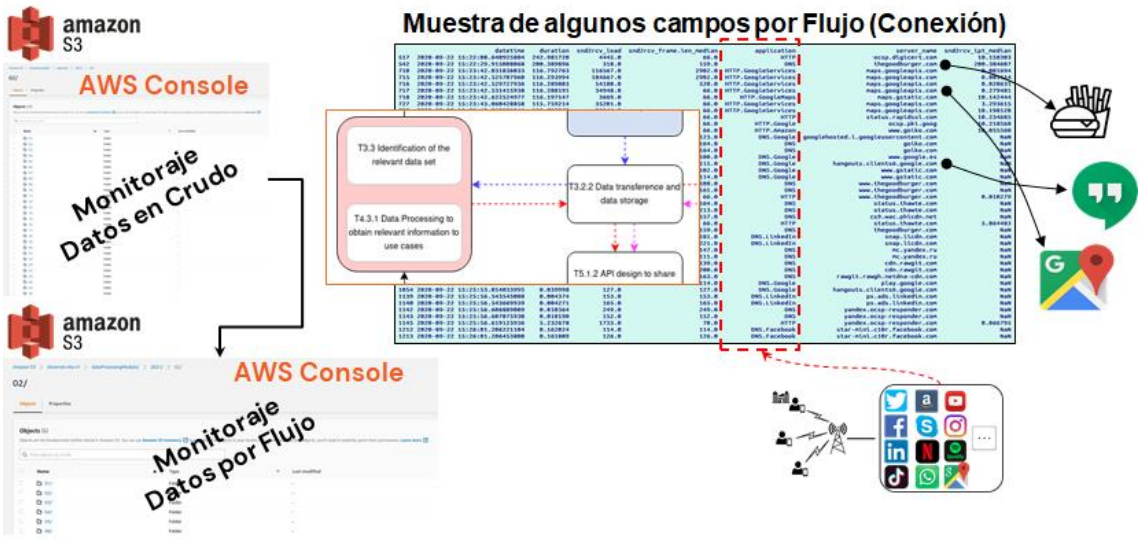

The network raw data obtained from the adapters is initially processed per flow allowing Starflow to identify network traffic characteristics.

Once this initial per flow process is completed a second processing per device allows also to identify the device characteristics.

The processed data is then sent to POLDER API




Among all the information that can be obtained from network traffic, we find the number of devices connected to a point of presence as a good indicator that can be used to help match supply and demand in sectors such as tourism, hotels, restoration and more.




Today, most people have at least one device connected to the Internet. For example, mobiles have applications that are continuously connected to the cloud to be evaluated, supported or updated. Additionally, there are multiple applications that keep people in contact with others through applications such as WhatsApp, Twitter, Instagram, etc. Therefore we can conclude that monitoring and predicting the number of devices connected to a point of presence is a reliable and effective indicator.

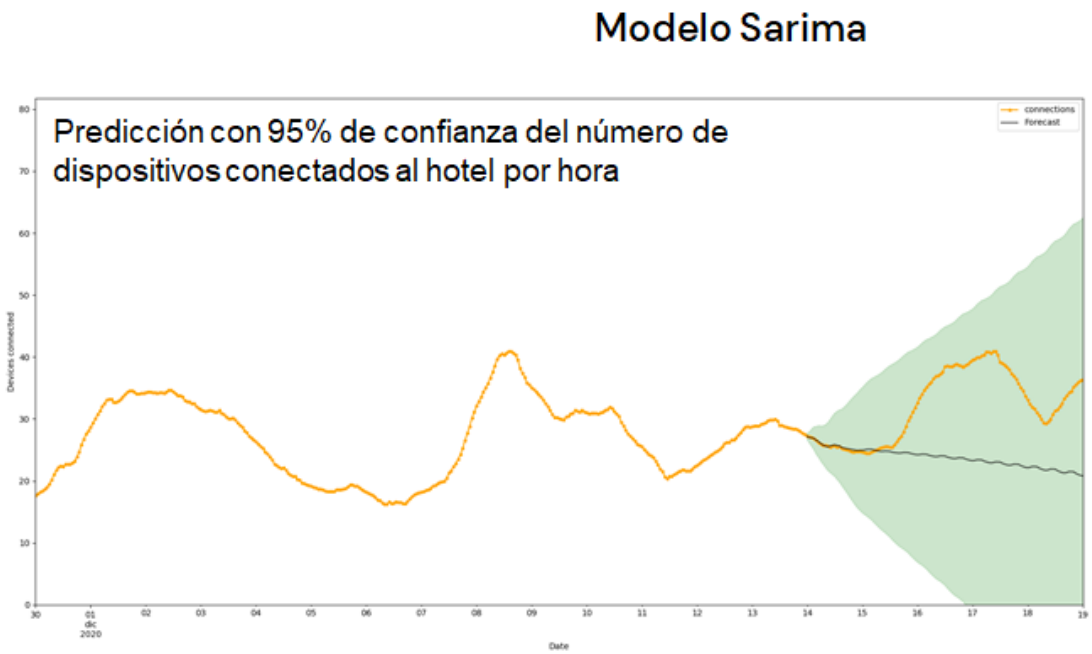
### 2.1.3.1 Model 1: Network Exploitation

#### 2.1.3.1.1 Objective

One of the objectives of Starflow has been to provide ML models or algorithms that allow obtaining relevant information for customers. Information which is useful in the business plan and management, helping in the business in general or in specific areas.

We have had the opportunity to analyze the traffic of a real case, specifically in the hotel sector. In order to provide information that is useful to this sector, we have had to study what network data or network data aggregations can provide value to the sector. Without counting the descriptive models or segmentation models that can be performed from the information of the applications of users' devices, we have found that, since most hotel customers have at least one device that connects or can connect to the Wi-Fi network of the hotel, counting the number of devices connected per hour, and more specifically predicting the number of devices connected to the Wi-Fi network, can help in research of matching supply-demand.

In general, matching supply to demand is difficult because demand can vary, predictably or unpredictably, and supply is inflexible. On average, an organization may have the right amount of resources (people, products, and / or equipment), but most organizations often find themselves in situations with resources in the wrong place, at the wrong time, and / or at the wrong time. For instance, in a hotel-restaurant / restaurant, it is important to know how to match the number of diners with the amount of resources: number of waiters, food (which will determine the menu), capacity, etc. Often the way to mitigate the supply-demand mismatch is by adjusting prices. This might influence prices of menu, drinks …




For the most part the hotel's supply will remain steady as they know how many rooms they have to sell. Also a Revenue Manager will be aware of any new supply from new hotels in their area and will often need to adjust their rates accordingly. A relatively new factor affecting supply is Airbnb, it turns out to be a strong competitor for the hotel sector.

Demand responds to motivation and desire of customers or consumers which needs a special analysis by Revenue Managers which have to consider the following factors: Events in the area, Seasonality, Midweek are quieter time, Economic conditions including foreign exchange rate, convention centre venues, tourism board activity such as advertising campaigns, …

Once supply-demand is defined, occupancy forecasting can be carried out more accurately as to how each market segment will perform each week and set rates accordingly. Where there are periods of low demand, this is where the hotel needs to create their own demand through promotions or targeting for group or conference business.

The prediction of the number of connections can be a useful tool to predict occupancy in hotels, beaches and hot spots of tourist interest.

#### 2.1.3.1.2 General Approach

Development of time series prediction by means of SARIMA model. Development and fine tuning parameters of an algorithm capable of obtaining a prediction of the number of devices connected to a point of presence (point of network traffic capture).

##### 2.1.3.1.2.1 Input Data

As has been mentioned through the deliverables, Starflow collects connectivity information from points of presents coming from devices such as mobile phones or computers. The following table summarizes the set of data obtained by processing, per packet and per flow, the network information:

| Device_id | | |
|---|---|---|
| **Field** | **Description** | **type** |
| **device_id** | *ID of the device* | *string* |
| **pop_id** | *ID of the point of presence* | *string* |
| **timestamp_zone** | *Timestamp in utc + zone of the last instant in the time interval* | *string* |
| **amount_of_connections** | *Number of connections given in the time interval* | *integer* |
| **amount_upload** | *Total upload (Egress in bytes)* | *integer* |
| **amount_download** | *Total download (Ingress in bytes)* | *integer* |
| **app_distribution** | *Counts of connections per application* | *dictionary* |
| **country_distribution** | *Counts of connections per country to which device had connected to* | *dictionary* |

Decision makers are very interested in predicting the amount of devices connected to a Point of Presence (PoP) per hour. The amount of devices connected provides a good indicator that can be used for helping to match supply-demand in areas such as hostels, catering and tourism in general.

#### 2.1.3.1.2.2 Expected Output

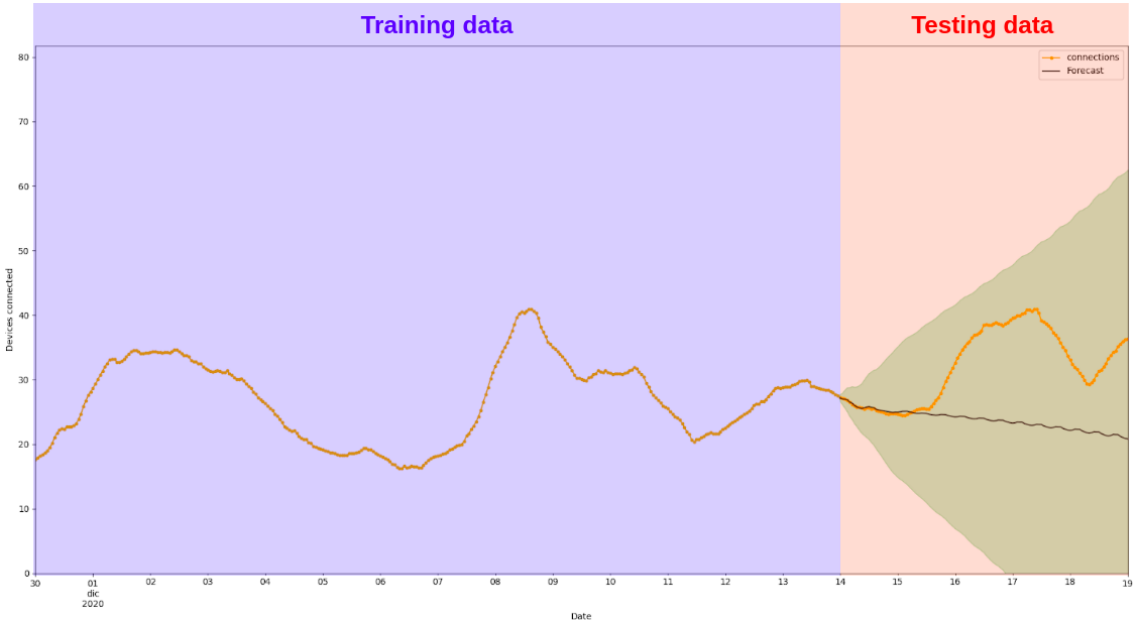
The prediction of the number of devices connected to a point of presence is expected. There can be different modes of production, among which is: 1- the prediction of the number of connected devices X hours ahead from a given moment. 2- the prediction of the number of connected devices X days ahead from a given moment.

#### 2.1.3.1.2.3 Proposed Methodology

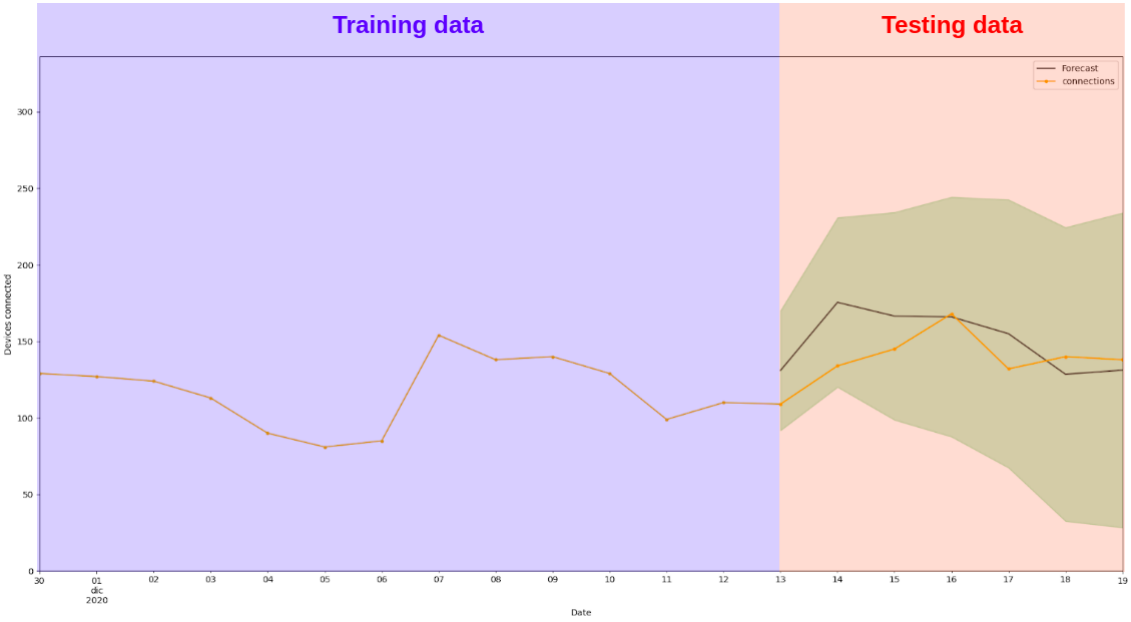
Data is collected from each of the points of presence. The number of collected devices is aggregated to obtain a time series of connected devices with a granularity of hour or day; Normally the hourly granularity of the day will be adopted. Then, once the time series has been acquired, the SARIMA model is applied to predict the connected devices. In the modeling process, the most suitable parameters for the SARIMA model are sought.

#### 2.1.3.1.2.4 Tests

An example of SARIMA time series forecasting with a very small sample size of real data obtained from a Hotel. Two figures can be seen below which represent the forecasting with different granularity of time for the same time interval. Models are trained from the training data part of the set, and prediction (in black) has more uncertainty as we go forward in time; it is therefore the cone of uncertainty in green. The first figure shows time series forecasting with a granularity of 1 hour. We obtained a good prediction, almost two days in advance. The second figure shows time series forecasting with a granularity of 1 day. We obtained a good prediction a week ahead. Prediction can be improved with more data, especially for the second case with day granularity, which is more interesting for the hotel / tourism sectors.




Devices connecting forecasting: Good prediction, hour by hour, almost 2 days in advance



Devices connecting forecasting: Good prediction, day by day, a week ahead.

#### 2.1.3.1.2.5 Conclusion

Time series forecasting with SARIMA provides a good estimation for the number of devices connected to a point of presence. Number of devices connected is a good indicator that can be used in order to make predictions.

## 2.2 City Monitoring (Monitor)

In the study, an energy optimization study was carried out for various devices in the home and building. A desired number of devices can be entered into the system in the home and building structure in order to optimize and observe the energy consumption in the study. Deep Q-learning (DQN) algorithm is used for energy optimization. As a continuation of this study, a new learning model was created. In this model, the total energy values resulting from the energy optimization were predicted. Random Forest Regressor, one of the popular machine learning models, was used for this study. At the same time, the explainable AI method was applied so that the results and outputs of the study could be interpreted and explained by the user. LIME, a popular method that is model-independent and can be applied to any machine learning model, has been used for explainable AI.

### 2.2.1 ACD

The DQN algorithm was used for the energy optimization model to be applied to the devices in the home and building. The DQN algorithm is one of the most well-known algorithms of reinforcement learning. The main purpose of the algorithm is to examine the next behavior and see the reward according to the actions to be taken, and to act accordingly by maximizing this reward. With this algorithm, a model was created that maximizes the reward and provides the best result. It has been observed that energy optimization is realized with this training model. Random Forest Regression machine learning model was used to predict the total energy values as a result of energy optimization. The created model has been made more explainable with LIME.

#### 2.2.1.1 Model 1: Model name

Home and Building Energy Optimization:

- Deep Q Learning Model (Reinforcement Learning)

Explainable Artificial İntelligence:

- Random Forest Regressor
- LIME ( LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS )

##### 2.2.1.1.1 Objective

There are many devices that consume energy in the home and building, and a large part of the energy is consumed by these devices. It is important for the smart use of energy that the amount of energy consumed by devices can be observed over time and how much energy they consume. Accordingly, it is aimed to optimize energy and reduce costs. The objective of the study is to enable the devices in the home and building to perform energy optimization with the developed model and to reduce the total cost as a result of the total energy used by the devices. At the same time, it is to make the energy optimization realized more understandable and more explainable for the user and to explain the model.

##### 2.2.1.1.2 General Approach

There are many devices and device groups that consume energy in the house and in the building, and a large part of the energy is consumed by these devices. It is important for the smart use and management of energy that the amount of energy consumed separately and in total can be observed over time. Considering such an approach, turning off the devices at certain times saves both energy and cost accordingly. In this study, the approach of optimizing energy and cost for home and building is emphasized.

#### 2.2.1.1.2.1 Input Data

There is time, total energy, generation energy, electricity cost and energy data used for 4 devices in the input data for home energy optimization. Time data are given in 5 minutes within a day. The amount of energy consumed by the devices is given in KW. In the input data, the electricity cost values are determined as three tariffs within a day and the unit is Euro. Time Of Use (Tou) enables customers to reduce their bills by using energy during off-peak times and to alleviate the pressure on the network by balancing demand.

• The energy data given as sensor0 in the data set used are the data of dishwasher device.

• The energy data given as sensor1 in the data set used are the data of heater device.

• The energy data given as sensor2 in the data set used are the data of air-conditioner device.

• The energy data given as sensor3 in the data set used are the data of electric vehicle device.

The building energy optimization data set contains data on the amount of energy consumed for 10 device groups, 7 of which are basic (Information and Communication Technology (ICT), Lighting, Space Heating, Hot Water, Other Process Heating, Other Process Cooling, Mechanical Energy). ICT is leveraged for economic, societal and interpersonal transactions and interactions. Internet access, communications technology (computer, tablet, phone etc.). Lightning group includes all the lighting systems used by the building. Space heating systems can be shown as central heating systems. A hot water boiler or a warm air heating burner. It is the domestic hot water supplied to taps. Fan heater, convector heater, Infrared heater and radiator heater devices are included in the Other Process Heating device group. Ventilation systems, air conditioning, condenser, evaporator,compressor, water tank, fan devices are included in the Other Process Cooling device group. The device types are wide in the mechanical energy device group. For example, devices such as fridge, washing machine, dishwasher, freezer, iron, hair dryer, vacuum cleaner, lawn mower, television, electric charged vehicles, small electrical appliances can be given. Device Group0, Device Group1 and Device Group2 data given in the data set are the devices that are not included in a specific group or are wanted to be monitored separately. These can be entered into the data set as many times as desired. It also includes the total energy, time and electricity cost for these device groups. Time data are given in one hourly within a week.

#### 2.2.1.1.2.2 Expected Output

In the study carried out, it is expected that an energy optimization will be realized for the device (for home energy optimization) and device groups (for building energy optimization) that consume energy in the home and building. For this reason, as a result of the study, it is expected that the device and device groups entering the training will optimize their energy, reduce the amount of energy they use and at the same time reduce their consumption costs. At the same time, it is aimed to show a more explainable result for the user by providing the explanation of the model. LIME was used to make the model explainable. The decision tree structure for energy and cost optimization is given below. Random Forest Regression, a

supervised learning algorithm, was used to create the decision tree structure and predict the model output.

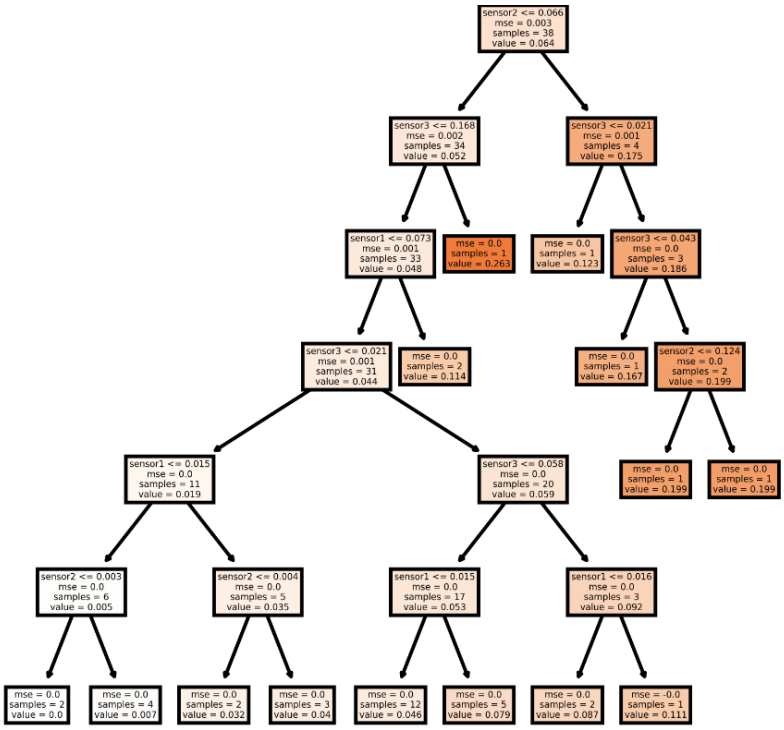



*Figure 30. Energy Optimization Decision Tree Structure*

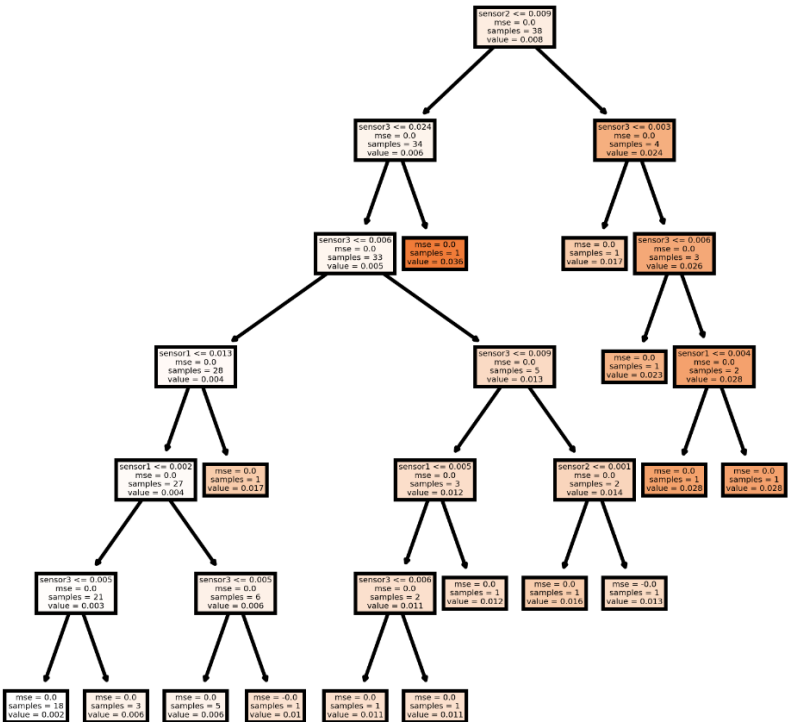



*Figure 31. Cost Optimization Decision Tree Structure*

### 2.2.1.1.2.3 Proposed Methodology

A method is proposed to optimize the energy values consumed by devices and device groups that consume energy in the home and building, and accordingly to reduce the cost. According to this proposed method, the energy consumed by the devices and the device groups separately and the total energy of the devices are given into the model as input data, and as a result, optimized energy values are obtained. The proposed model approach for the study is DQN.

#### 2.2.1.1.2.4 Tests

The energy optimization process performed with the created model can be tested with different test data to see the energy optimization process of the device and device groups.

#### 2.2.1.1.2.5 Conclusion

There are many energy-consuming devices in the home and building, and a large part of the energy is consumed by these devices. It is important for the smart use of energy that the amount of energy consumed by the devices can be observed over time and how much energy is consumed. As a result of the study, energy optimization of the devices in houses and buildings have been achieved and the total cost of the devices has been reduced as a result of the total energy use. There is a ±2% tolerance of error in the developed training model. At the same time, LIME was used to make the energy optimization realized more understandable and explainable for the user and to explain the model. Thus, the model result is shown in a more understandable way.

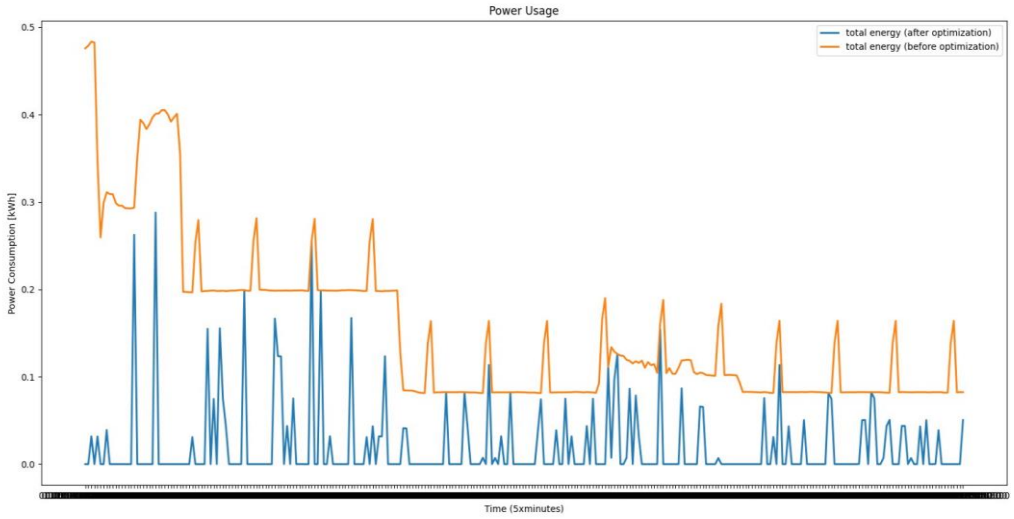



*Figure 32. Total Energy Graph Before and After Optimization (Home Optimization)*

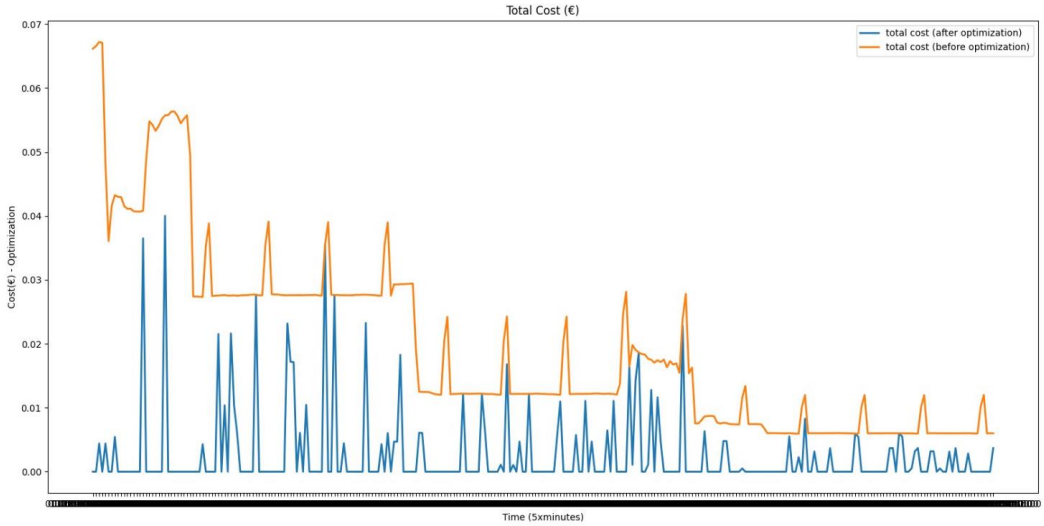



*Figure 33.Total Cost Graph Before and After Optimization (Home Optimization)*

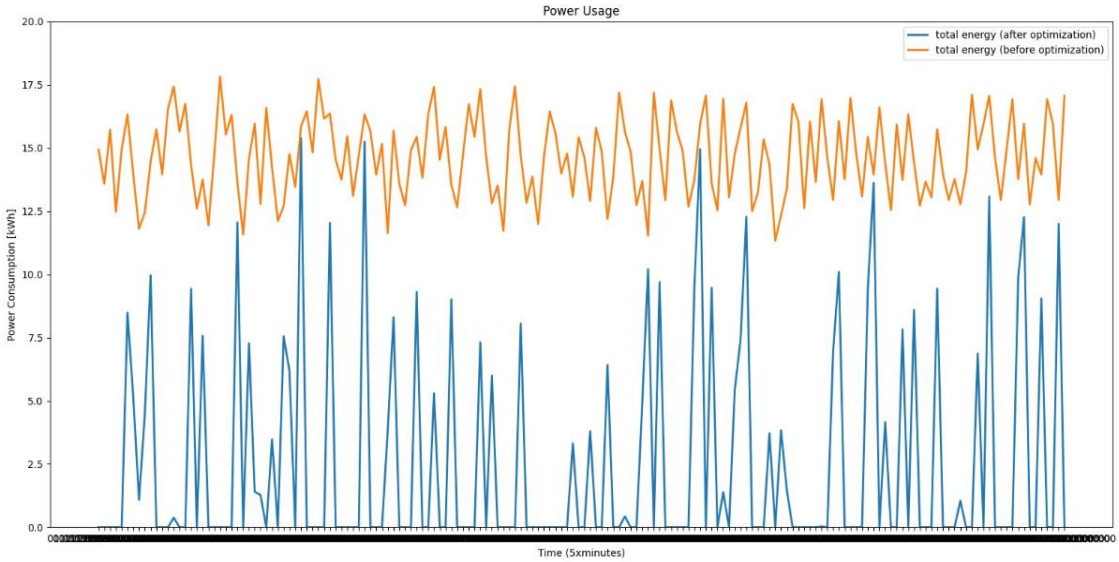



*Figure 34.Total Energy Graph Before and After Optimization (Building Optimization)*

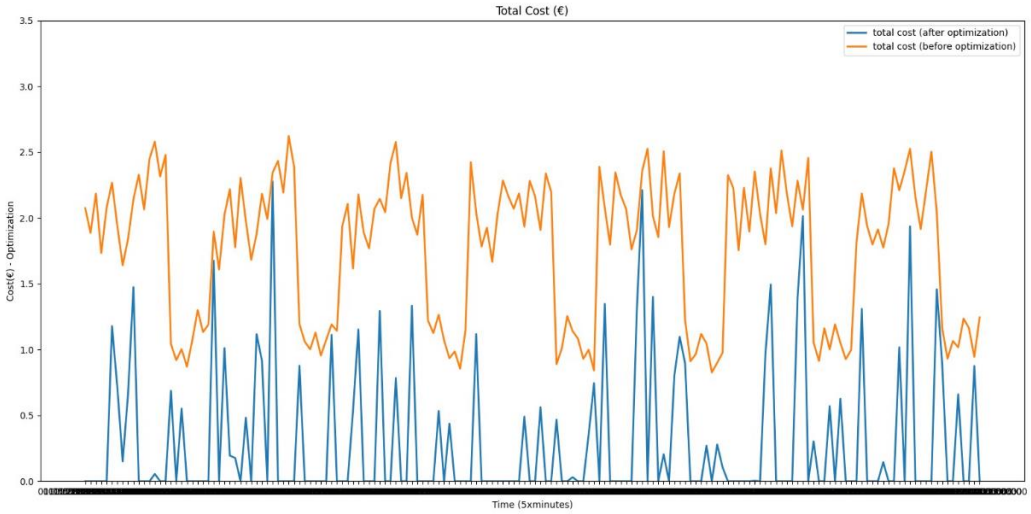



*Figure 35.Total Cost Graph Before and After Optimization (Building Optimization)*

## 2.2.2 ARD GRUP

In the platform developed for traffic monitoring, which is designed for the use of decision makers and citizens, Random Forest (RF) and Dynamic Time Warping (DTW) algorithms from machine learning models were used to predict traffic parameters.

A decision support system was developed that uses traffic flow data from multiple points together with auxiliary information obtained from sensors. The final system is expected to support decision makers in taking actions in response to changing events.

ARD considers here the problem as predicting a continuous label (regression) from multimodal time-series data. These sources may represent either point-wise traffic flows from different locations or spatial supportive data such as weather information or other IoT input. Each signal is taken as single-source time-series sequences and multiple data aligned over time is considered as a multimodal time-series. The continuous label to be predicted can define any current traffic state (local or global) which may support decision making, e.g. next flow at a specific roundabout.

### 2.2.2.1 Traffic Prediction models

The framework which has been developed is shown in Figure 36. The system takes a short multi-modal time-series as input and compares it to all other annotated records with labels in a pairwise manner. The model finds the nearest samples and applies a majority-based prediction based on the weighted averaging of known labels.

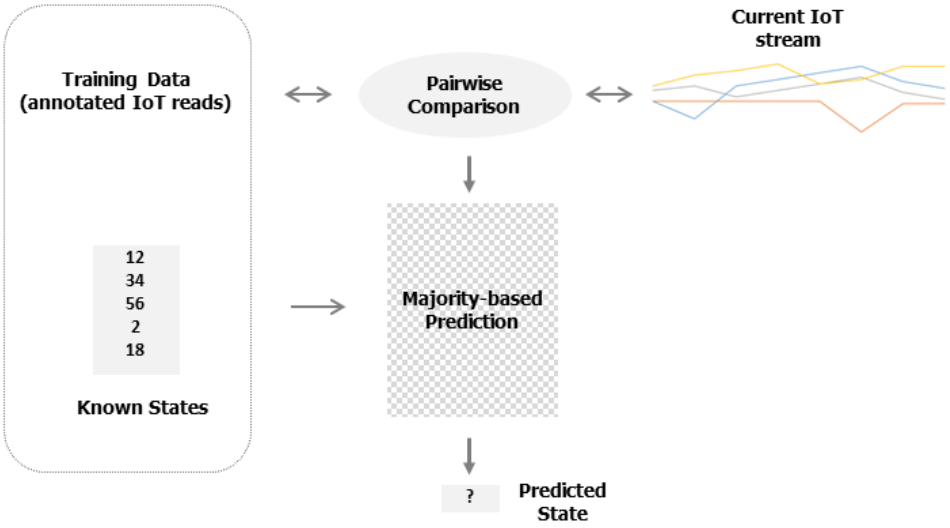



*Figure 36. Multi-modal time-series prediction framework for intelligent traffic monitoring.*

Due to the fact that the DTW algorithm is not efficient in terms of time and resource usage due to the increase in data volume on the platform used, the "constrained bending window" approach used in elastic similarity metrics has been adapted to the model. According to this approach, a sub-window is selected on the matrix used during the similarity calculation and alignment calculations are made only within the boundaries of this window (Figure 17). This solution is named Fast DTW.

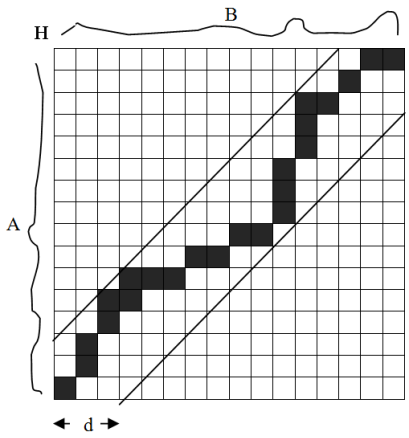



*Figure 37. Acceleration of the DTW algorithm with a bounded bending window.*

In order to perform this operation, it will be sufficient to enter an external parameter (d) that will determine the bandwidth in both rows and columns. Thus, for a fixed value of d, the algorithm complexity drops from O(n2) to O(n).

As an alternative to the DTW algorithm, the following four methods were adapted, coded and tested for our problem during this period:

1. ARIMA: It is a model that makes statistical predictions based on the past trends of the data without using training sets.
2. Long Short Term Memory (LSTM): It is an iterative deep neural network model. It is preferred for time-series and sequence data.

3. Discrete Cosine Transform - Support Vector Regression (DCT-SVR): SVR is an efficient machine learning model that works with support vectors. Since it cannot work directly on the time-series, it has to be fed with some features to be extracted from the signal. Here, DCT transform is applied for feature extraction and the obtained coefficients are used as features.
4. Discrete Cosine Transform – Random Forest Regression (DCT-RFR): The features obtained by DCT are used to train a random tree hybrid structure. Random Forest algorithm (Figure 18) is one of the most used algorithms among machine learning algorithms. Some of the reasons why it is one of the most used are that the algorithm produces large results in a wide scope, is easier to use than equivalent algorithms, and includes a more flexible working mechanism. At the same time, both Classification and Regression tasks can be used with this algorithm. Random Forest, as the name suggests, creates multiple decision trees randomly using the data given to it and uses these trees to combine them to predict the correct and sustainable path. One of its biggest advantages is that it creates a flexible working environment in Classification and Regression tasks, as mentioned above. Random Forest allows the decision trees it creates to grow by adding additional random features to the model. A reliable conclusion is that Random Forest does not search for the most important feature when splitting any tree, but rather finds the best feature among a random subset, unlike conventional decision tree algorithms. This allows it to generate a more meaningful model and a wider variety of results.

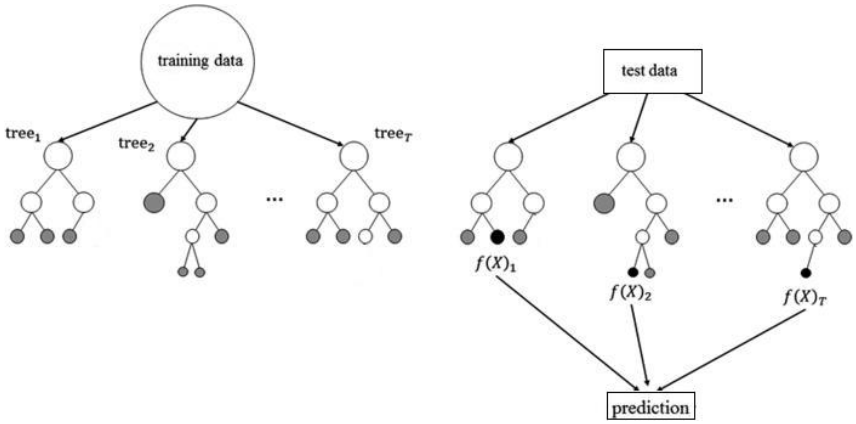



*Figure 38. Random Forest Regression*

Comparative results obtained by applying these methods in a 2-fold cross validation setup are given in the table below (Table 3). The values given show the correlation coefficient between the actual number of vehicles and the estimated number of vehicles.

*Table 6. Experimental comparison of methods used for traffic frequency estimation*

| Before (min) | After (min) | DTW | Fast DTW | Arima | LSTM | DCT - SVR | DCT - RFR |
|---|---|---|---|---|---|---|---|
| 10 | 1 | 0.547 | 0.549 | 0.000 | 0.700 | **0.728** | **0.728** |
| 15 | 1 | 0.561 | 0.536 | 0.479 | 0.736 | **0.762** | 0.746 |

| Before (min) | After (min) | DTW | Fast DTW | Arima | LSTM | DCT - SVR | DCT - RFR |
|---|---|---|---|---|---|---|---|
| 30 | 1 | 0.613 | 0.609 | 0.747 | 0.595 | **0.743** | 0.690 |
| 45 | 1 | 0.650 | 0.630 | 0.724 | 0.678 | **0.731** | 0.724 |
| 90 | 10 | 0.693 | 0.725 | 0.871 | **0.872** | **0.872** | 0.848 |
| 120 | 10 | 0.459 | 0.360 | 0.658 | 0.608 | 0.668 | **0.716** |
| 180 | 10 | 0.765 | 0.724 | 0.689 | 0.644 | **0.843** | 0.689 |

#### 2.2.2.1.1 Conclusion

According to the table, although the best results were obtained with the DCT-SVR method, DTW, Fast DTW and DCT-RFR methods, which gave close to the best results, were coded and implemented in the real system at this stage due to the limitations related to the running time of this method.

## 2.3 Trust and Citizen Acceptance

In recent years, the emergence of new information and communication technology has contributed to major changes in society. The growth of smart ICT devices changes the communication habits of people, their connection with the environment and with others, so that it creates an interconnected society. The development and structure of cities are affected by ICT and can be used to make them more sustainable. Social networks are a good example of this evolution. In general, a Social Networks Site (SNS) is defined as a service that enables users to share their personal profile and opinions regarding different topics. Social networks are also considered to be a powerful platform for acquiring knowledge of a city. They are currently one of the key points of intelligence provided to urban planners in order to figure out how public infrastructure is being used by people. Understanding the actual uses of urban public spaces plays an important role in the planning and design of smart cities. Therefore, governances take advantage of new opportunities provided by social networks to allow people to take part in the decision-making process and to obtain citizen acceptance.

### 2.3.1 Mantis

Mantis developed a sentiment classification model to recognize positive, negative, and neutral tweets with respect to the Istanbul subway so that urban planners could observe the public satisfaction rate for the Istanbul transportation system.

#### 2.3.1.1 Model 1: Neural Sentiment Analysis

##### 2.3.1.1.1 Objective

The objective of the social media analysis is to combine machine learning and semantic approaches to analyze social media to help urban planning administrations improve social sensing and urban services. In the case study we developed in the context of the POLDER project, Twitter data is used to monitor social media in relation to different aspects of the Istanbul subway, and our objective is to estimate the sentiment of these tweets using neural network models.

##### 2.3.1.1.2 General Approach

In our model, we have used a lexicon-based context dependent approach proposed in (Teng et al, 2016[1]). In this approach a Turkish sentiment lexicon proposed in (Ucan et al, 2016[2]) has been employed to calculate the sentiment polarity of tweets. This method uses a deep recurrent neural network to learn context-dependent sentiment weights to change the lexicon polarity of terms depending on the context of their usage.

---

1 Zhiyang Teng, Duy-Tin Vo, and Yue Zhang. 2016. Context-sensitive lexicon features for neural sentiment analysis. In EMNLP, pages 1629–1638.

2 Alaettin Ucan, Behzad Naderalvojoud, Ebru Akcapinar Sezer, and Hayri Sever. 2016. SentiWordNet for new language: Automatic translation approach. In Signal-Image Technology & Internet-Based Systems (SITIS), 2016 12th International Conference on, pages 308–315. IEEE.

#### 2.3.1.1.2.1 Input Data

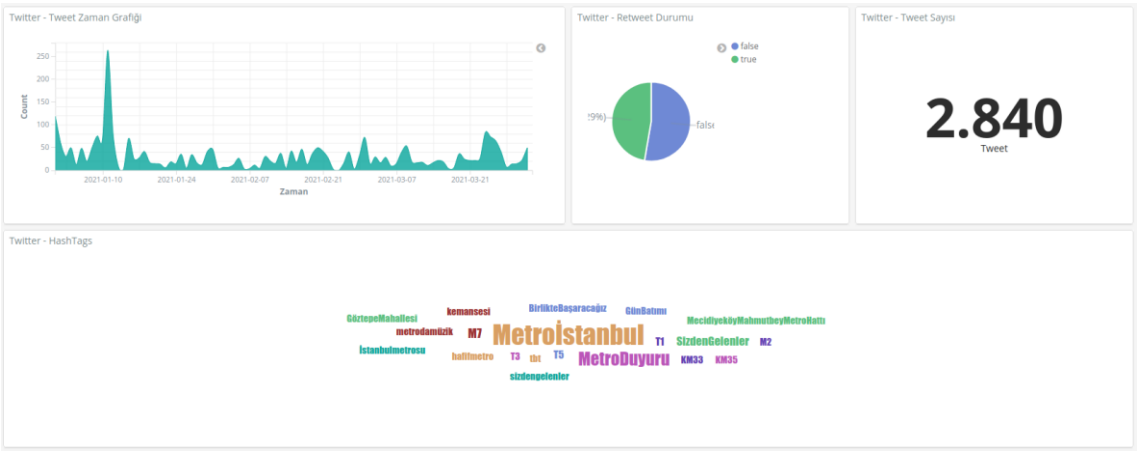
The monitored Twitter data includes message, user names , location, user description, tweet hashtags, and reply counts. This data is used to select our model's dataset that includes only text messages.

#### 2.3.1.1.2.2 Expected Output

The output of the model is sentiment tags including "positive", "negative", and "neutral".

#### 2.3.1.1.2.3 Proposed Methodology

To build our sentiment model, we have used tweets monitored in the first 3 months of 2020 with a keyword/hashtag of "MetroİstanbuI". As a result, we have achieved 2840 tweets as shown in the following fugure. In order to annotate tweets, manual and automatic approaches were used to the train and development sets, while all tweets in the test set were labeled manually. Tweets for test set were selected from April, 2021. In the automatic approach, emoji icons used in the tweets were considered for annotating. In addition, randomly 50% of annotated tweets were checked manually in the train set.




Examples of tweets for three types of classes are as follows:

Neutral tweets:

- RT @herkesicinCHP: Eminönü-Alibeyköy Hattı açılış töreni #İstanbul
- RT @: #MetroDuyuru 30-31 Ocak hafta sonu seferlerimiz 📣 🟣M1, M2, M3, M4, M5, M7, T1 ve T4 hatlarında 06.00-00.00 saatlerinde…

Positive tweets:

- RT @: Hafta sonu bakımlarımız tamam. Biz ekibi yeni haftaya hazırladık. 🚀😎 # # #hafifmetro #toplu…
- İstanbul gibi seviyorum seni, beylikdüzünden tavşantepeye kadar çok ama çok ❤ @ @istanbulbld

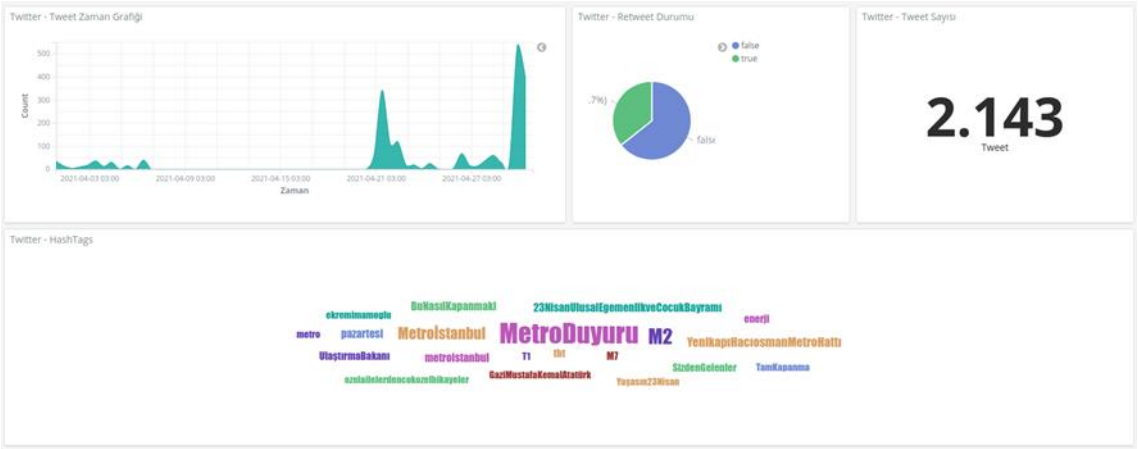
Negative tweets:

- @ Eski metrolar M1, M2 M3 M4 m6 sürücüsüz teknolojiye yıllar sonra geçecek mi?
- @ duraklarındaki ücretsiz olan tuvaletleri de ücretli yaparak nereye varmaya çalışıyorsunuz? Yap… https://t.co/0YmSfbl4gu
- @ 541 nolu araçta anons hoperlor arızası var. Bilginize arz ederim...

We employed a lexicon based deep learning method based on BiLSTM proposed by Teng et al (2016) for sentiment classification. In this approach, the sentiment score of a sentence is computed based on the weighted sum of the polarity values of the subjective words obtained from the lexicon. In fact, these weights are learned from training samples to modify the prior polarity values of words with respect to their usage context. In this model a Turkish sentiment lexicon proposed in (Naderalvojoud et al 2018[3]) is adopted. In this lexicon, the sentiment polarity of words are achieved from English SentiWordNet through a mapping process.

#### 2.3.1.1.2.4 Tests

The trained sentiment model was evaluated on the test set that we achieved from 01/04/2021 to 30/04/2021 shown in the following figure.




Out of 2143 tweets monitored in April, we selected 1000 tweets to be manually annotated as positive, negative, and neutral.

The following table report the result of sentiment classification model on the test set using BiLSTM as a baseline approach and lexicon-based BiLSTM.

| Model | Positive-F1 | Negative F1 | Neutral F1 | Avg F1 |
|---|---|---|---|---|
| BiLSTM | 45.70 | 67.49 | 80.56 | 64.58 |

#### 2.3.1.1.2.5 Conclusion

According to the results, we can observe that our sentiment model performs better on the neutral and negative tweets than the positive ones. The cause of this observation is that the majority of tweets are neutral or negative, and the model tends towards the majority classes through training. Nevertheless, negative tweets are more important in smart cities, and detecting these kinds of tweets can give meaningful feedback to policy makers and urban

3 Naderalvojoud, B., Qasemizadeh B., Kallmeyer, L., 2018. and Sezer, E.A., "A cross-lingual approach for building multilingual sentiment lexicons." In International Conference on Text, Speech, and Dialogue, pp. 259-266. Springer.

service developers to improve their performance and increase the acceptance of these services among citizens.