Universitat Politècnica de Catalunya

Master's Degree

Automatic Control and Robotics

Thesis

Big Data and Data Analytics applied to the Monitoring of Water Distribution Networks

Mireya Muñoz Balbontín

Director: Vicenç Puig Cayuela

Course: 2014/15

June 2015

Abstract

The project associated with this master thesis was performed in collaboration with Water Technological Centre CETAQUA[1], located in Cornellà de Llobregat. The aim of this thesis is to implement the use of data analytics and machine learning models for demand monitoring in water networks using data collected from *Automatic Meter Reading* (AMR). With this, the scope of the project intends to analyze the influence of external variables on the consumption pattern.

Starting from the knowledge of the state of the art in the existing Big Data techniques and tools, the most adequate for water distribution networks monitoring will be chosen. Three different datasets were processed for the purpose of the study, two of these corresponding to the city of Tarragona, and the remainder to the city of Torremolinos; databases including the variables linked to the consumer account associated to the meters were also used. All of the datasets and databases were provided by the company.

The study consisted of two stages, in the first one the datasets were split by season and analyzed separately to evaluate the features presented in each one. To evaluate the representative behavior for each city, the clustering labels were analyzed to find the groups of sensors who share the same pattern in behavior. In the second stage three different models were applied to the data to find the relation between the demand patterns and the meter account variables.

The results reveal a symbolic number of groups of sensors that follow the same behavior in the seasonal analysis; outlier activity associated to high consumer and non-domestic use was also detected. The results obtained in the second stage suggest a forced input-output relation among the meter account variables and the clustered patterns; these results improve when combining these variables with features associated with the demand pattern. Some of the drawbacks during the execution of the project were the untrustworthiness of some predictors, as well as the loss of information due to outlier extraction or missing data.

[1] CETAQUA board members: the Spanish National Research Council (CSIC), the Universitat Politècnica de Catalunya (UPC) and the Agbar Group.

Index

ETSEIB

1. Preface

This first chapter provides the reader with a brief introduction to the project associated with this master thesis. The motivation is explained in the first part, which is mainly resulting from a previous project carried out in the same company. The second part introduces the proposed objectives, linked to the extraction of useful knowledge of consumption and user behavior from AMR data. The outline for this document is enlisted, and briefly summarized at the end of the chapter.

1.1. Motivation

The aging of water networks combined with an increase in demand, caused by the growth in population and constant edification, challenge the control of water distribution infrastructures. One of the main challenges is the inability to detect anomalies caused by bursts, leaks, water-loss, un-accounted activity, or other factors; added difficulty is presented when desiring to detect these anomalies in real time [9]. Water network monitoring not only helps to ensure the adequate distribution to the final user, but it also furthers sustainability by reducing loss at different stages in production and pumping. Given this need, Supervisory Control and Data Acquisition Systems (SCADA) have become the foundation for water utilities, allowing constant data collection from crucial points in the network.

Within the context of SCADA systems and the implementation of smart meters, AMR provides near-real-time data collection directly from the end user. Previous to the implementation of these meters only monthly or bimonthly readings were available and collected on-site, making it difficult to properly analyze customer behavior and evaluate to further extent the state of the infrastructure. Smart meter installation has become a widespread practice, with 220.000 currently installed and a projected increase of 60.000 per year [1]. Considering this, and the fact that some meters have been collecting data for more than 5 years, it is certain that the amount of data generated by these meters and other sensors available in the system surpass the limits of any conventional statistical analysis tool. This opens the possibility for the implementation of Big Data technologies.

The term Big Data has acquired a great level of acceptance over the past few years, not only for the competitive advantage this technology provides for large companies, but also from all

the profit and knowledge that can be obtained through the analysis of what companies had previously thought of as mere information. All of these companies with significant amounts of data and little knowledge of how to profit from it have adopted Big Data technologies and started building a data strategy. Even more important than obtaining precision from the data analyzed, is to strive for data relevance [2], which in the context of machine learning translates to an adequate feature extraction and the implementation of reliable predictors. The general scheme that any company must consider to start building a data strategy is shown in Figure 1.1.

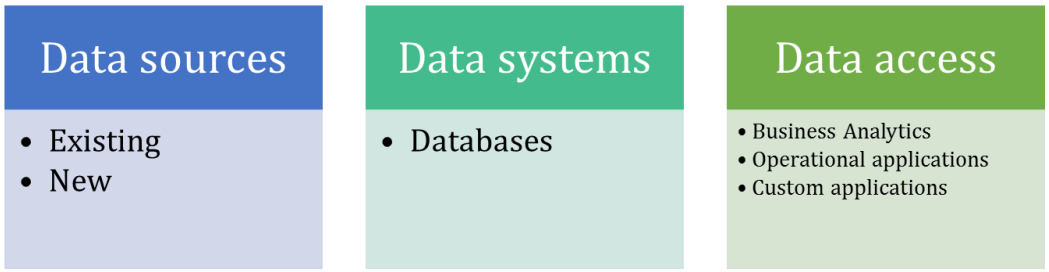


Fig. 1.1. Key elements for building a data strategy

The main goal of Big Data platforms such as Cloudera Hadoop is to help organizations profit from all their data by offering an enterprise data hub where all the elements included in Figure 1.1 can be accessed and processed. The Cloudera enterprise data hub contains many functionalities and access to apache Hadoop projects and applications both for data management and system management, Figure 1.2 is available for further information.

In the context of water distribution networks monitoring, using smart meter data, machine learning techniques have proven useful elements for analysis. Fortunately, there exists the possibility to implement these techniques in a Big Data platform. One of the many applications provided by Hadoop is known as Spark™, a fast computing engine that supports applications and libraries for machine learning such as MLlib™ and Mahout™. Much can be learned by using these applications, and benefitting from the architecture of the platform, which allows storage of all types of data from all types of sources; after all, it has already been proven that expertise can come from the most unexpected sources, such as open domain data from social networks and search engines [5].

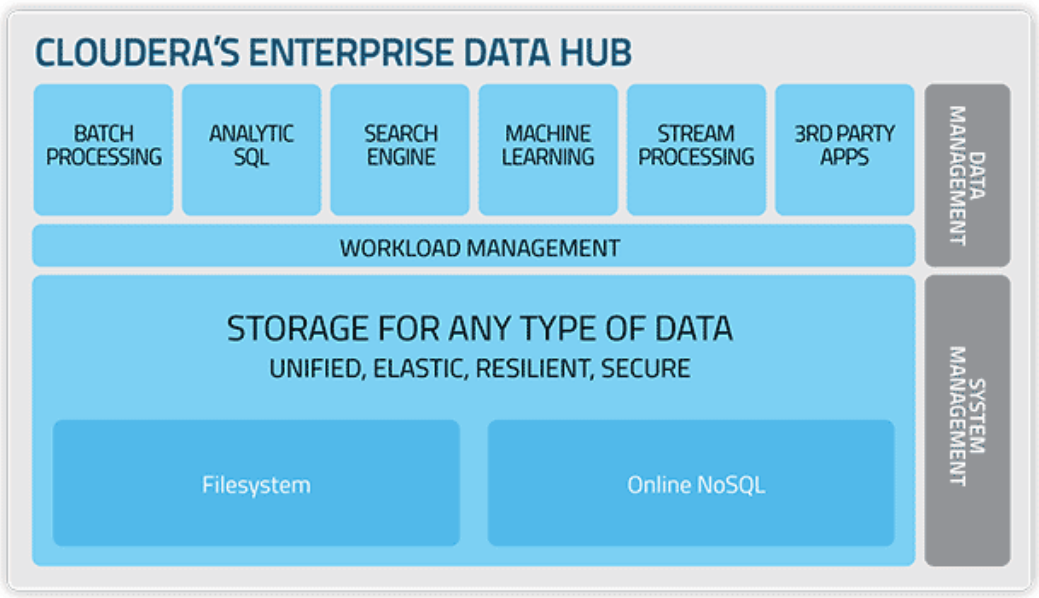


Fig. 1.2. Grand scheme of Cloudera's Enterprise Datahub

1.2. Objectives

The general objective for this project, in the framework of developing a data strategy for the company, is to extract knowledge from the data provided in order to obtain value from it. More specifically, it is desired to implement data mining and machine learning techniques to data obtained from AMR technology systems to extract useful knowledge about the variables associated to the consumption of different users and their behavior when consuming water.

The general track to be followed can be better summarized as shown in Figure. 1.3. During the pre-processing stage the data will be formatted as a time series object The timescale will be adjusted and, having these indices, outliers will be detected and removed from the datasets used for the following stage.

For the clustering stage, feature matrices according to the needs of the study will be obtained; using the timestamp indices, the data will be split into seasons for the case of the seasonal clustering.

For the knowledge extraction stage the following is expected: by applying an unsupervised clustering algorithm, as well as some summary statistics and indicators, useful conclusions about the generic behavior of consumption during the year can be extracted. The

implementation of regression and classification models will allow the identification of variables with high influence in relation to the results obtained through unsupervised clustering.

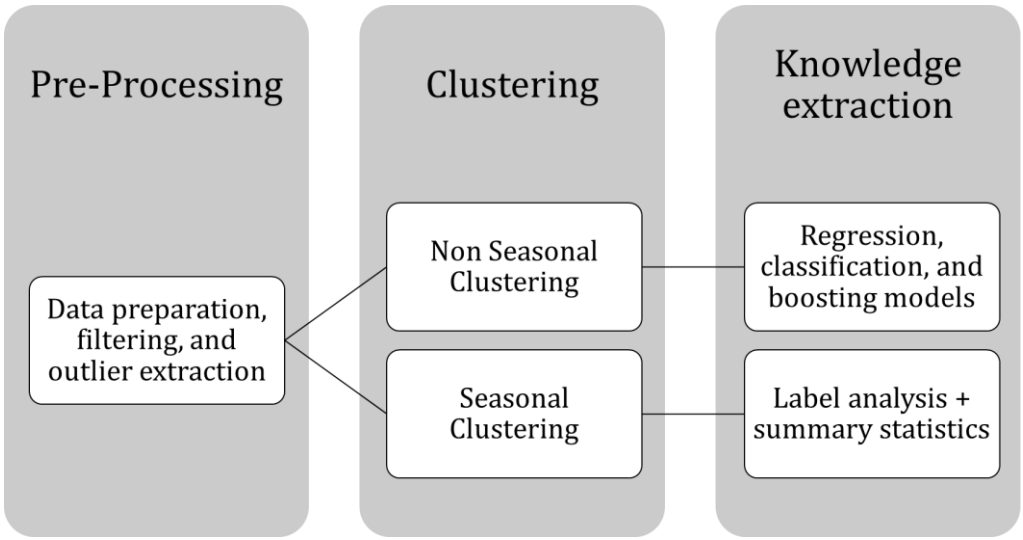


Fig. 1.3. General procedure scheme

1.3. Outline

The upcoming contents of this document are listed and briefly summarized as follows:

- **Chapter 2. State of the art in AMR:**

This chapter contains the preface and background that justifies the motivation for this project. A brief introduction to the concept of AMR is presented, followed by a brief mention of the issues that have arisen from the implementation of these SCADA systems. To finish, this chapter presents a brief summary of some previous work and applications done in the field.

- **Chapter 3. Big Data, Data Mining, and Data Analytics applied to AMR**

This chapter focuses mainly on the concepts and techniques involved in data processing. Several techniques for data mining and data analytics are discussed, explaining their advantages and disadvantages when applied to the problem in question. A listing of the selected techniques for the scope of this project is also included

- **Chapter 4. Selected Techniques**

Having previously defined the adequate techniques to be used, the sequence to apply these techniques is explained. This sequence consisting of two parts, a seasonal analysis of the datasets, and a regression and classification analysis of the general unsupervised clustering of the datasets. The first section of this chapter enlists the various clustered patterns for each dataset, including a generalized assumption of what these could represent; the second section presents all the knowledge obtained from the analyzed patterns when applying the selected techniques and cross validating the results with the rest of the variables involved. Some preliminary results and conclusions are presented in this chapter.

- **Chapter 5. Results**

All results obtained throughout the development of the previous stages, and considered relevant to the objective of the project, are presented and discussed in this chapter.

- **Chapter 6. Conclusion and future work**

This final chapter contains a summary of all the conclusions that can be drawn from the results obtained. Some suggestions to improve these are mentioned as part of future work.

2. State of the art in AMR

2.1. Problems

With all the benefits provided by AMR technologies, some of which were commented in section 1.1 of this document, also come several problems that can compromise the outcome of the data obtained during the stage of data acquisition, or during the processing of it. Not all of these are associated to the physical condition of the meter, although those problematics are harder to correct through a Big Data framework, as these could be purely physical anomalies. The main problems and their possible causes are listed below:

- **Problems with the meter:**
  - Wrongful installation
  - Buildup of dirtiness
  - Blockings
- **Problems in Data Acquisition**
  - Intermittent or missing readings
  - Large sporadic inactive periods (gaps)
  - Atypical readings (in relation to the average values)
- **Fraudulent activity**
  - Clandestine connections
  - Damaged meter
  - Magnetic add-ons
- **Others**
  - Excessively high annual or monthly consumption
  - Excessively low annual or monthly consumption

2.2. Employed analysis tools

Although a number of applications exist that allow the processing and analysis of this type of data, the chosen tool for statistical computing for the purpose of this project was R. This is a well-documented and easy to work open source tool. Aside from the basic libraries provided by R, the libraries used for the purpose of this study are enlisted as follows:

- Data formatting, and processing. Data frame handling. Subsetting and indexing
  - Xts: allows formatting of the data as an extensible time series, useful to bind all data to its specified timestamp
  - Data.table: creation and manipulation of data frames
  - Plyr: useful for indexing and partitioning datasets
- Machine learning
  - Rpart: recursive partitioning and regression trees. Includes all tree models
  - Gbm: used to fit generalized regression boosting models
- Ggmap: allows plotting of geographic data

2.3. Previous studies

The project associated to this master thesis derives from another project elaborated in the same company titled: Water demand estimation and outlier detection from smart meter data using classification and Big Data methods [1]. The methodology followed in this project can be condensed in four general steps listed below:

1. **Preprocessing**: raw SCADA readings are regularized by applying a linear interpolation, it is assumed all input data is valid.

2. **Filtering**: outliers are detected using statistical indicators and discarded for further processing.

3. **Feature space**: weekly patterns are extracted for each smart meter and considered as the input feature for the following step.

4. **Unsupervised clustering**: the *k-means* algorithm is considered to partition all the observations into a specified number of clusters.

This project was developed using a Big Data framework based on Spark™; the hardware architecture formed by two clusters, one for storage and processing, and another used to store the results. Once the hypothesis is tested, it is intended to implement the methodology

In the context of machine learning, other projects associated with the same research line, are those developed by Solanas et. al. [10][11]which use factor analysis, clustering analysis and discriminant analysis in order to analyze the variability of the data obtained through AMR. For the purpose of the study:

- Factor Analysis is used to define the categories to be used and the meaningful variables associated to these,
- Two-way Clustering Analysis is used to cluster the cases and the variables
- Discriminant Analysis is used to find linear combinations of linked variables that separate the groups of cases the best, intending to confirm or reject each associated variable.

A different approach to water demand classification is presented by McKenna S.A. in [6], where Gaussian Mixture Models represent the foundation for representing the demand patterns. Probability density functions are chosen as the function to fit the demands and assign a probability of occurrence to any demand value. The features used for multivariate clustering correspond to the estimated parameters of the mixture model. *K-means* clustering is also considered along with the GMM to examine the stability of the identified clusters which results in an adaptable approach to fitting demand.

Another previous study, in which billing variables and socio-demographic variables are taken into account was developed in Portugal by Mamade, A. 2013 [3]. Some of the methods used include Cluster analysis, PCA, and Multiple Linear Regression. For the purpose of the study, the raw SCADA readings were taken approximately every 15

minutes. The stages considered can be summarized as follows:

1. Outlier detection: performed using two robust statistics, the median and the robust standard deviations of the observations

2. Water consumption characterization: the average daily consumption per month is calculated, with the purpose of performing a cluster analysis searching for seasonality among months. A framework of consumption variables is constructed

3. Water demand profiling: given that 49 indices were calculated in the previous step, the first task will be variable reduction (PCA). A correlation matrix is used for profiling consumption patterns

3. Big data and data mining applied to AMR

3.1. Discussion on applicable techniques

When it comes to data analytics, a number of techniques become available to profit as much as can be possible from the data acquired. Given the nature of the considered data, as well as the desired use it is intended for, applicable techniques involve statistical tools, as well as machine learning models. The latter prove useful given their known reduced computation time and efficient results, as proven in the associated literature. The following list contains all possible applicable techniques, of which some will prove adequate given the scope of the project:

- Basic statistics
    - Summary statistics
    - Correlations
    - Stratified sampling
    - Hypothesis testing
    - Random data generation
- Classification and regression
    - Linear models (SVMs, logistic regression, linear regression)
    - Naive Bayes
    - Decision tres
    - Ensembles of trees (Random Forests and Gradient-Boosted Trees)
    - Collaborative filtering
    - Alternating Least Squares (ALS)
- Unsupervised Clustering
    - k-means
- Dimensionality reduction
    - Singular Value Decomposition (SVD)
    - Principal Component Analysis (PCA)
- Feature extraction and transformation
- Optimization (developer)
    - Stochastic gradient descent

- Limited-memory BFGS (L-BFGS)

3.2. Selection of adequate techniques

Below, all the techniques chosen to be applied for the scope of this project are listed. In the end, and due to time constraints, not all of these were applied to the provided data, but their further development is still being studied, and the proposed implementation for these data is further discussed in this document.

- **Basic statistics**
  - *Summary statistics:* basic summary statistics are helpful for a generalized understanding and simplification of all the variables in question. There is no specific stage at which these techniques will be needed, they will rather be used when needed and when the data calls for it. These techniques are also useful for presenting results in a concise way.
  - *Random data generation:* most of the models implemented in this project are subject to a random initialization; it was therefore useful and necessary to set a specified seed to keep results reproducible.
- **Classification and regression**
  - *Linear models (SVMs):* SVMs represent a useful and practical tool for linearly separable data, choosing the adequate variables and formatting, in order to obtain good results
  - *Decision trees:* Decision trees are simple and convenient models for predicting relation among variables even when these do not necessarily show it.
  - *Ensembles of trees (Random Forests and Gradient-Boosted Trees):* Due to their iterative nature during computation, as well as their online error minimization feature, it is expected that these techniques will compute an accurate prediction with very low error. A slight drawback of using these ensembles is overfitting, which is usually associated to the quality of the predictors.
- **Unsupervised Clustering**

- *K-means:* This algorithm aims at splitting *n* observations (sensors) into *k* clusters, each observation belonging to the cluster with the closest mean value. The use of this algorithm will possibly represent the foundation for it, as all output classification labels will then serve as input to the regression and classification models

- **Feature extraction and transformation**
  - This technique allows the summarization of the data in a separate level; whether the features obtained are for time-series clustering or in the context of classification and regression predictors.

4. Patterns

As discussed in the previous chapter, several techniques were chosen in order to process the data and extract important knowledge to be used for the company. The analysis using the proposed techniques was divided in two parts. In the first part, a seasonal analysis was computed to evaluate the different types of consumption behavior over time, i.e. detect and analyze seasonal and stationary properties. In the second part, using the partitions obtained from the general unsupervised clustering analysis, a regression and classification study was executed to find the relation between all billing variables provided and the classification given by the *k-means* clustering. These parts are further explained in the sections below.

Three datasets were provided for the completion of this project, two of them corresponding to the city of Tarragona, and the last one to the city of Torremolinos. These datasets contain the raw hourly readings per sensor. Given that each datum is bound to a timestamp, the xts() object is used to format all of the input data; this allows simplified time series data handling/formatting, which will become quite useful when proceeding to the seasonal analysis. Having this formatting, it is then possible to adjust the timestamp of the readings, assuming all data is valid, by applying a linear interpolation. Having made these adjustments, it is then applied a filtering stage in which the outliers will be extracted from the dataset. For the purpose of this study, all meters registering a constant consumption of zero, as well as those in which the missing data surpasses a factor of 10%, will be discarded.

Once the data is arranged in the manner described previously, we are left with the task of constructing the feature matrices. Each feature matrix corresponds to a season, where each row contains 168 features (i.e. columns) from every sensor in the dataset. These features correspond to the mean hourly water consumption of a weekly pattern for a sensor during the seasonal timestamp assigned, meaning the full set of features comprises the demand weekly pattern which characterizes the seasonal behavior for the sensor. A weekly pattern has been chosen rather than the widely used daily pattern [6] to include the weekend behavior.

Aside from unsupervised clustering, these features proved useful for a second stage of feature extraction when conducting the regression and classification analysis of the data,

which will be discussed in detail further. Each feature matrix will be used as the input to extract the classes through unsupervised clustering using k-means.

This algorithm requires the number of clusters as input. Hence, the estimation of the adequate number of clusters was achieved with the aid of a plot of the within groups sum of squares; one plot per feature matrix was computed. Choosing an optimal value for this specific parameter proved slightly difficult for the case of the seasonal feature matrices, given the dissimilarities among the obtained plots, but ultimately it was decided to choose the same for all four seasons, though not being the optimal number, this simplifies the analysis by keeping the results consistent, and allows room for atypical or irregular pattern extraction, that may have been disguised otherwise. The plots will be shown in each section, with the number of clusters selected marked in red.

4.1. Non- Seasonal Clustering

4.1.1. Tarragona:

As mentioned previously, two different datasets were provided for the city of Tarragona. The unsupervised clustering performed on each of these datasets is presented in the following sections. These results were helpful to determine the basic differences between these two separate groups of users

Dataset 1 – Domestic

The complete dataset contains 199 sensors, 17 of which are discarded as they do not comply with the quality requirements desired. This leaves 182 sensors which will be used to obtain the demand patterns. Figure 4.1. shows the within groups sum of squares plot for the features.

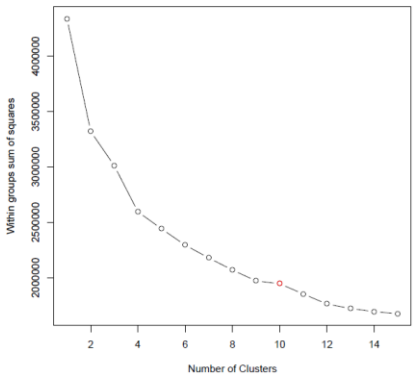


Fig. 4.1. Within groups sum of squares plot. Tarragona Dataset 1 - Domestic

Figure 4.2. shows the clustered patterns for the first dataset, with the cluster center (weekly average) drawn as a thick black line. With the exception of cluster 10, whose consumption and maximum value are the highest of all, all clusters have more than 2 members. Characteristcally to domestic use, most patterns show the standard morning/evening activity peaks, some more evident than others, which can simply be associated to the living habits and characteristics of the user, such as: showering and cooking schedule, working week schedule, activity during the weekends, and the number of inhabitants per household.

At this point no further information about the nature of the data or the sensors (e.g. activity declared by the customer, address, sensor model, brand, or caliber, among others) is included in the analysis. This will be discussed and analyzed further in this document.

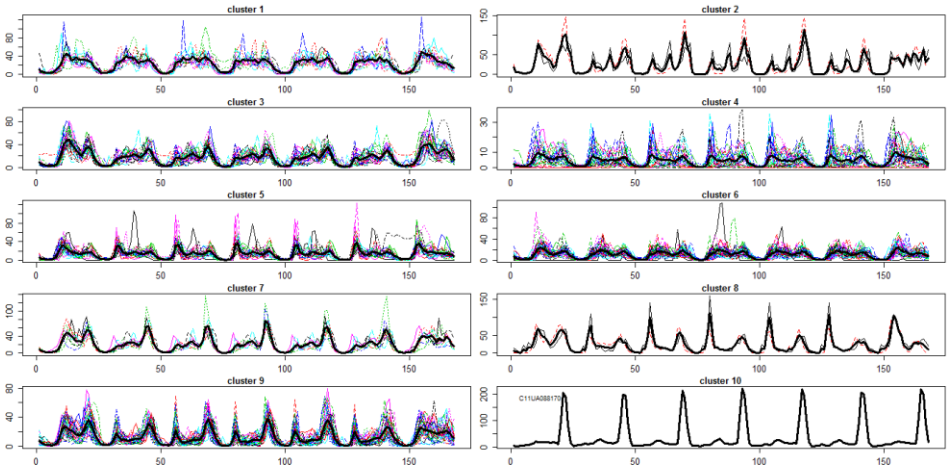


Fig. 4.2. Clustered demand paterns for Dataset 1, x axis: time as 168 hour week, y axis: mean demand in l/h.

Dataset 2 - Industrial

The complete dataset contains 110 sensors, 40 of which are discarded as they do not comply with the quality requirements desired. It was found that some of these had already been replaced due to malfunction in readings. This leaves 70 sensors which can be used to obtain the demand patterns. The within groups sum of squares plot for this dataset, shown in Figure 4.3., helps in determining that 6 clusters is more than enough for this dataset. Once the algorithm is computed, it is seen that the great majority of the sensors are classified in cluster 5, which corresponds to the one with the lowest mean weekly demand. All of the remaining clusters do not seem to contain more than 6 sensors.

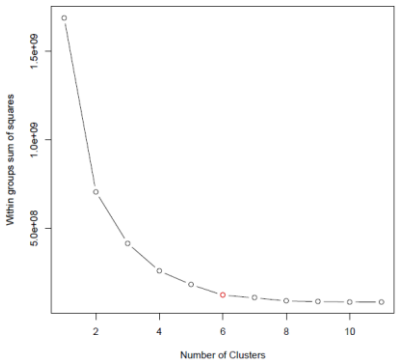


Fig. 4.3. Within groups sum of squares plot. Tarragona Dataset 2 - Industrial

All of these patterns show a vey high weekly consumption average, consistent with their registered use as industrial consumers, with the exception of cluster 2, all have day-to-day activity patterns. Some other notes on these classification can be:

- Although cluster 1 shows a day-to-day basis consumption, these appear to be consistent with part-time us, given the long inactive interval between working hours. It also shows high consumption on Sundays. Similarly, cluster 6 also shows this type of high consumption during the weekends, but with irregularities during the week.

- Cluster 2 shows uninterrupted use during the weekdays, which could correspond to a factory or business which closes almost completely during Saturday and Sunday.

- Cluster 4 shows similar behavior to that of domestic users, given the notoriety of the morning/evening activity peaks, but with a very elevated mean consumption in relation to any domestic user.

- 81% of the sensors are classified in cluster 6 with a mean weekly demand of 54,4 l/h, the mean pattern shows regular to constant use during the day, in contrast to the rest of the patterns which appear noisy. These could range from local businesses, restaurants, bars, up to domestic businesses.

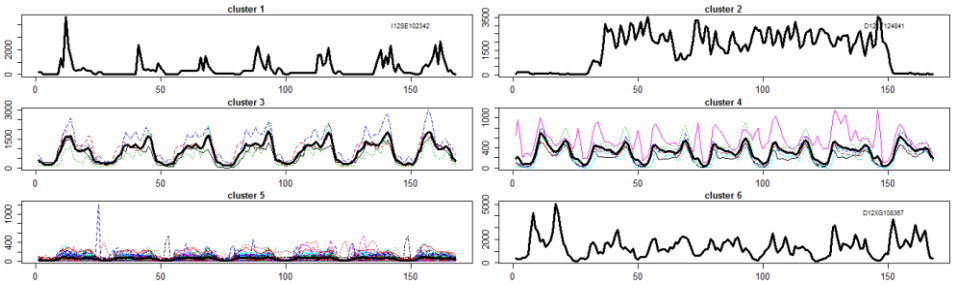


Fig. 4.4. Clustered demand paterns for Dataset 2, x axis: time as 168 hour week, y axis: mean demand in l/h.

Further analysis executed using the clustering for this dataset proved it would not be useful for the seasonal analysis, or the classification and regression analysis. Aside from the fact that not enough data is available per season, it was decided that not enough sensors were available to find a generalized outcome; some of these data did not figure in the billing information (representing more loss of information), meaning they had to be discarded, and due to the fact that 81% belong to the same class, it would prove difficult to find an underlying relationship with other variables associated to the account, or the meter.

Torremolinos

This dataset was provided once it was decided that more sensors would be needed in order to improve the feasibility of the study and improve the generalization, although during the outlier detection stage there was also a significant loss of sensor data, it is unknown whether the use of the sensors is domestic or industrial. From all the sensors available in this city, some filters were applied in order to ensure there would be sufficient data, or at least enough to also compute the seasonal analysis. These indicators are explained below and their chosen values shown in Table 4.1.

Indicator	Meaning	Value
Duration	Hours measured	> 8500
Total consumption	Total consumption of the sensor	> 0
MaxStat	Maximum value	> 10
dataRate	Number of readings/number of hours	> 0.6

Table. 4.1. Indicators used to ensure quality in readings

Once after applying these indicators on the raw dataset, 1209 sensors are found to be useful for the analysis, 503 show big gap rates when passed through the filtering process and thus will be discarded from the analysis. This leaves 706 sensors for the unsupervised clustering stage. Figure 4.5. shows the within groups sum of squares plot for the features. Although 10 clusters could have been enough, it is decided to use 12 given that this allows room for extracting irregular patterns associated with a different use of the sensor.

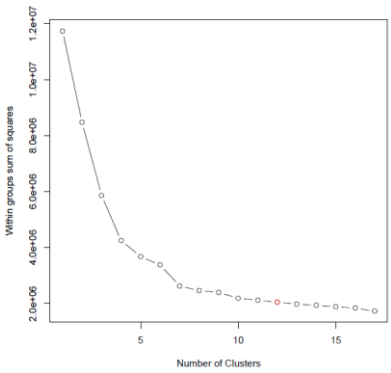


Fig. 4.5. Within groups sum of squares plot. Torremolinos Dataset

Table 4.2. summarizes the contents of the clustered patterns seen in Figure 4.4. Approximately 80% of the sensors show a mean weekly consumption ranging between 1,02 and 6,2 l/h, classifying them as low consumers. These sensors belong either to cluster 7, 8, or 10, whose patterns show consistent and reduced daily use. Since it is known that Torremolinos is a coastal city, where some of the sensors might be installed in summer homes, more information about the user behavioural consumption pattern can be obtained from the seasonal analysis. The cluster mean morphology, which does not show the pronounced two-peak consumption pattern known to domestic users, is a slight indicator of atypical domestic behaviour which could be characteristic to cities such as this one.

Clusters 3, 4, 5, and 6, on the other hand, contain only one sensor each, whose mean consumption is very high compared to the rest of the clusters. Given the irregularities in the consumption pattern as well, these can be classified as non-domestic users up to this point, which will further be validated in the regression and classification analysis.

Cluster	Mean	Sensors
1	16,3487459	29
2	30,2121162	5
3	32,8052481	1
4	41,4007941	1
5	37,3111981	1
6	27,4040985	1
7	1,02852389	252
8	3,25862098	174
9	10,0409081	65
10	6,20306417	137
11	77,4581539	2
12	14,1463158	38

Table. 4.2. Summary statistics for unsupervised clustering of consumption paterns in Torremolinos

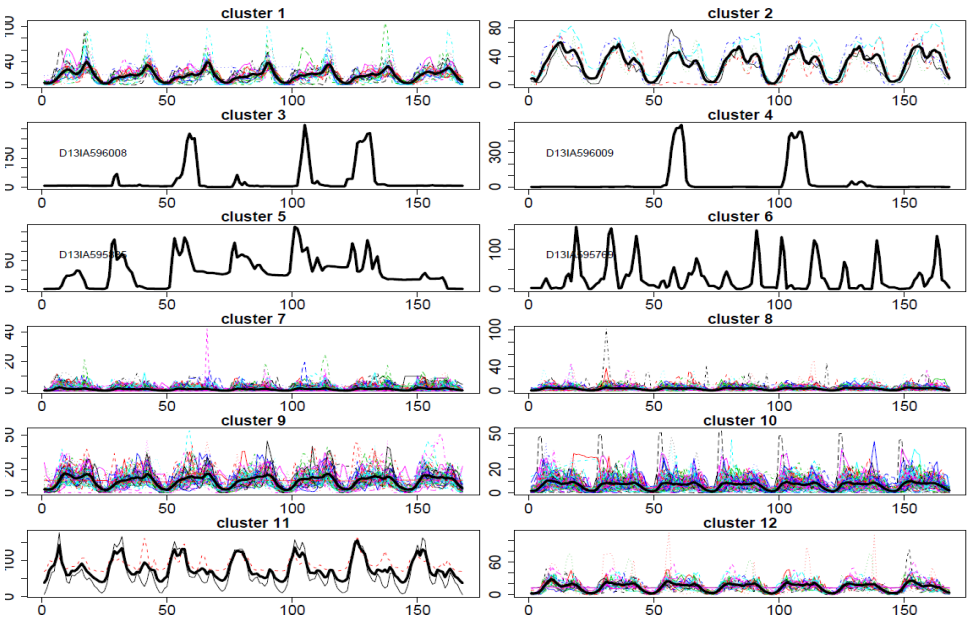


Fig. 4.6. Clustered patterns for Torremolinos. x axis: time as 168 hour week, y axis: mean demand in l/h.

4.2. Seasonal clustering

Given all the data is formatted as an xts (eXtensible Time Series) object, it is possible to use functions such as .indexmon(), .indexwday(), and .indexhour(). These allow the splitting of time series data through specified indices associated to the timestamp, facilitating the extraction of

all necessary data pertaining to each season; each one defined as shown in Table 4.3.

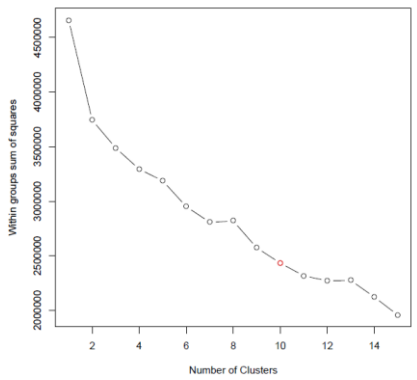
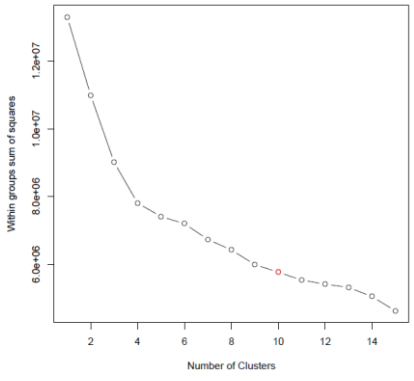
Season	Months
Spring	March, April, May
Summer	June, July, August
Fall	September, October, November
Winter	December, January, February

Table. 4.3. Definition of months considered per season, used to subset the data.

The same procedure for feature extraction and *k* parameter estimation is followed, although from now on, the clustering will have to be done for each season, which will give us four sets of classification labels to be later analyzed in the stage of knowledge extraction. The seasonal patterns for both cities are shown in the following sections, as previously decided, the industrial dataset corresponding Tarragona will not be considered for the purpose of this anlysis.

4.2.1. Tarragona

Before computing the clustering of these patterns, it must be ensured that all sensors contain enough data per season to obtain the best approximation possible. Fortunately, all 182 sensors comply with this. Table 4.4. shows the within groups sum of squares plot per season, marked in red is the chosen number of clusters which will be used in the analysis (for this particular case, it was considered adequate to use 10 clusters).

Spring	Summer
	

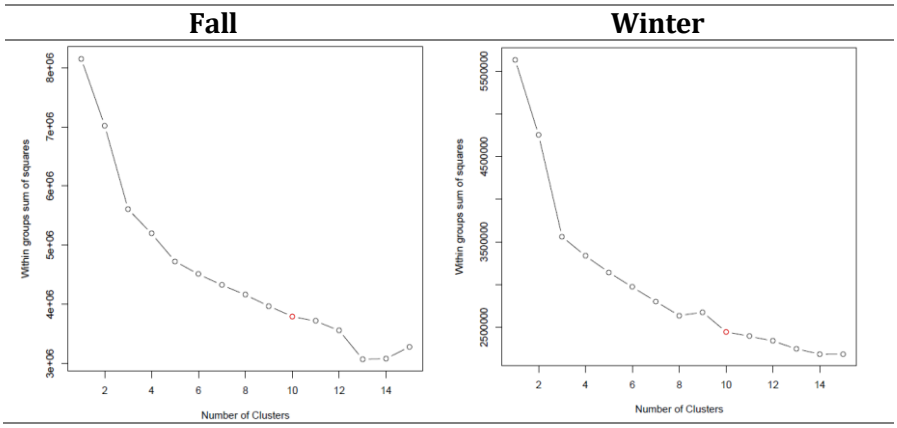


Table. 4.4. Within sum of squares plots per season. Tarragona

The observable dissimilarities among the plots represent the first hint of characteristic behavior pertinent to the season associated; this will become more evident after computing the clustering and visualizing the patterns. Having chosen the adequate number of clusters, it is then proceeded to compute the k-means clustering of the weekly demand patterns. Since all the sensors in this dataset are of domestic use it is of no surprise that all cluster mean patterns show the characteristic morning/night activity peaks known to domestic users. Some irregular activity is also present. This will further be discussed in the entry for each season.

Spring

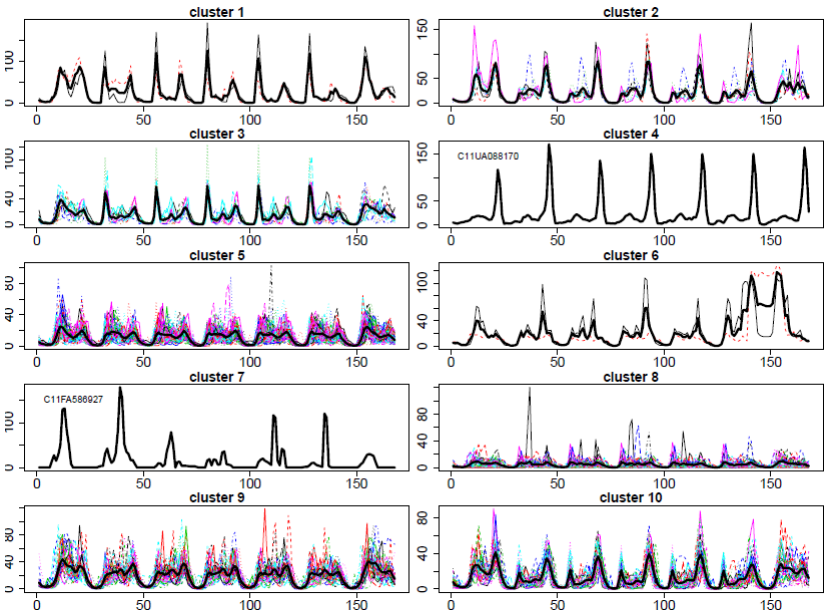


Fig. 4.7. Clustered paterns for Spring subset in Tarragona. x axis: time as 168 hour week, y axis: mean demand in l/h.

The clustered patterns for this season are shown in Fig. 4.7. Clusters 2 and 3 are consistent with the standard domestic pattern, though they show a small third peak in consumption during daily afternoon hours, which could be consistent with families that eat at home, work only half a day or activity such as this one. Clusters 4 and 7, the two outliers found in this clustering have very high consumption peaks at what appears to be the end of the day. Clusters 5, 8, and 10 also show the standard domestic characteristics with a less pronounced peaks in the morning and evening hours. Irregularities during the weekend can be seen in cluster 7.

Summer

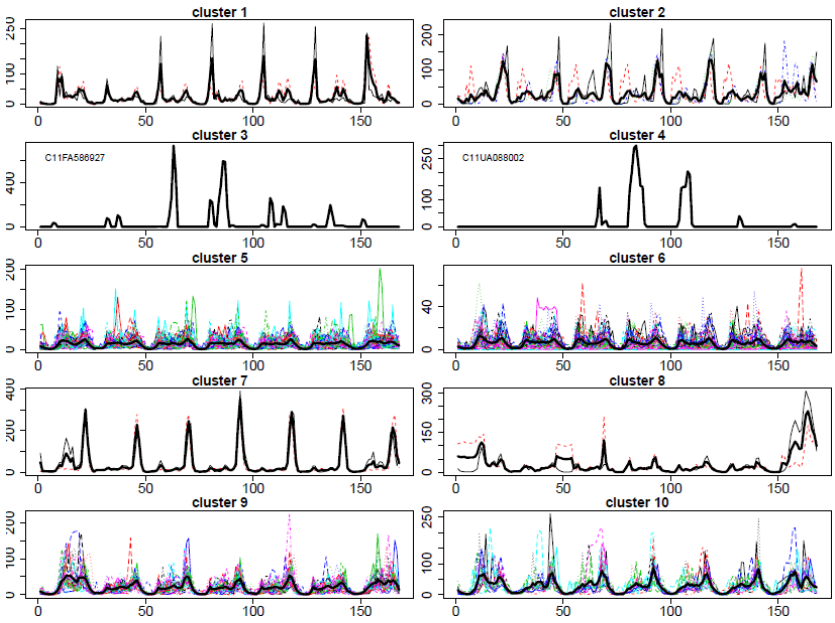


Fig. 4.8. Clustered paterns for Summer subset in Tarragona. x axis: time as 168 hour week, y axis: mean demand l/h.

As expected for this season, the mean consumption per cluster increases, which could be caused by families with small children spending the summer holidays at home or increase their consumption due to climate conditions. The patterns expressed by clusters 5, 6, 9, and 10 seem to be consistent with this assumption, meaning the domestic consumption characteristic peaks are present, only with an increase in the mean consumption. Clusters 3, 4, and 7 contain outliers with sporadic high consumption peaks throughout the week. Clusters 1, 2, and 8 show irregularities that could be consistent with vacationing families.

Fall

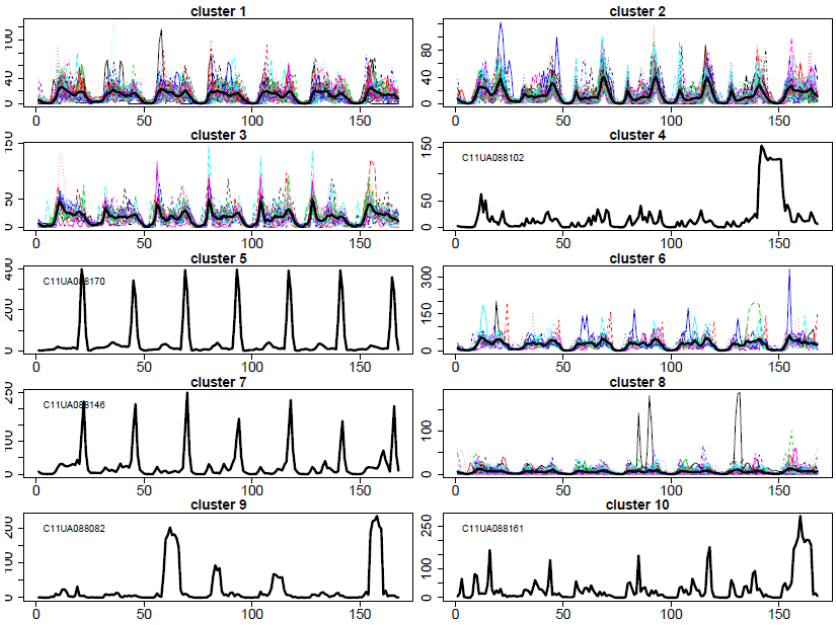


Fig. 4.9. Clustered paterns for Fall subset in Tarragona, x axis: time as 168 hour week, y axis: mean demand l/h.

Fig. 4.9. shows the clustered patterns for the fall months. It is observed that half of the clusters contain outliers, while the other half show the same patterns previously seen with the characteristic peaks for domestic users; the cluster mean decreases in relation to the activity observed during the summer. The outliers in clusters 5 and 7 show regular activity with a high weekly consumption. Irregular activity is characteristic to the remaining outliers, specially for cluster 9, where the high demand peaks are only present twice per week; clusters 4 and 6 show the highest demand peaks on Saturday.

Winter

Fig. 4.10. shows the patterns obtained for the winter months, the cluster means do not differ much from those observed in the fall, and the patterns appear similar to the patterns observed in the summer. This is consistent with the behavior expected for this time period.

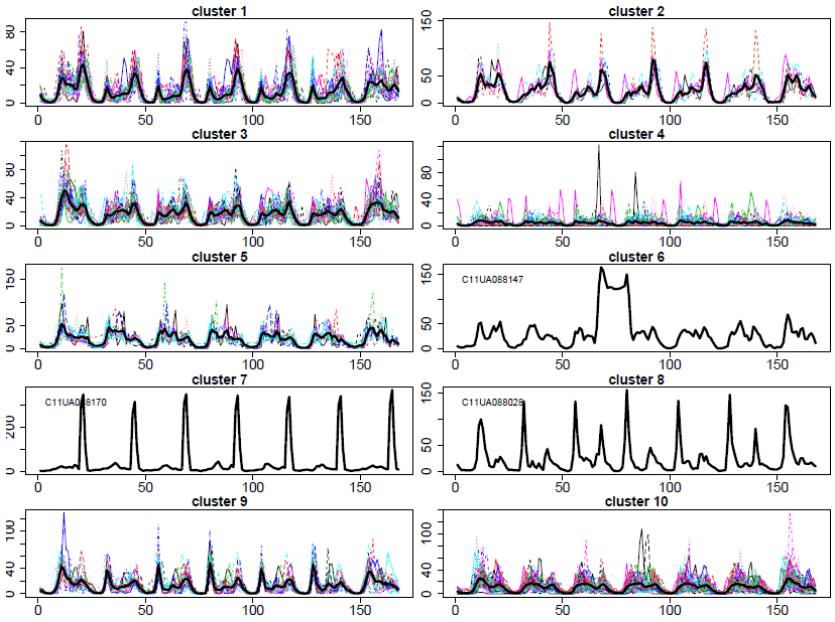
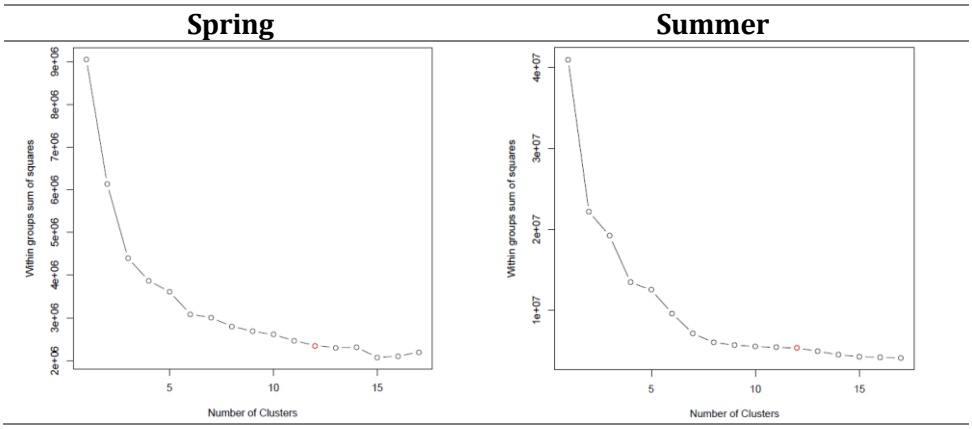


Fig. 4.10. Clustered paterns for Winter subset in Tarragona, x axis: time as 168 hour week, y axis: mean demand l/h.

4.2.2. Torremolinos

Out of the 706 sensors that were used for the non-seasonal clustering of this dataset, only 596 contain a full year (12 months) worth of data; these are the sensors that will be taken into account for the seasonal clustering and the seasonal analysis further detailed in this document.

Table 4.5. shows the within groups sum of squares plots for each seasonal feature matrix. It was decided to use 12 clusters to keep consistency with the non-seasonal clustering, and to allow room for outlier/non-domestic pattern extraction. The patterns obtained will be shown above in the same manner as those presented for the city of Tarragona.



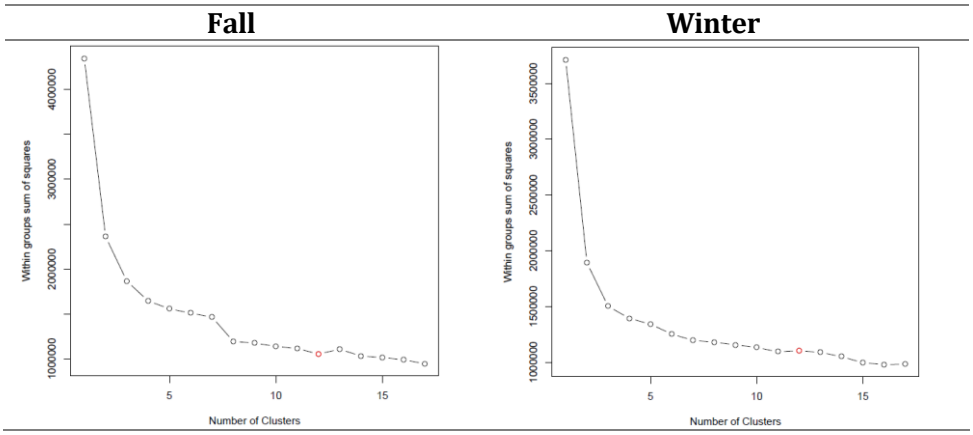


Table 4.5 Within groups sum of squares plots per season for Torremolinos Dataset

Spring

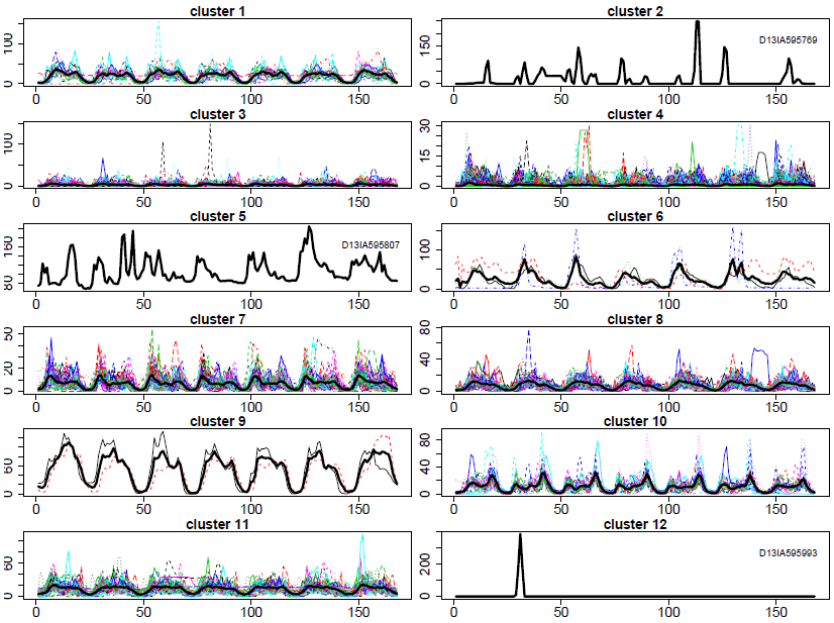


Fig. 4.11. Clustered paterns for Spring subset in Torremolinos. x axis: time as 168 hour week, y axis: mean demand l/h.

Most of the clustered patterns shown in Figure 4.11 are characterized by a low mean weekly consumption pattern. Three outliers are extracted, two of which show sporadic activity, the remainder show irregular weekly behavior with a reduced number of consecutive inactive periods, which could indicate a minor leak. These three outliers could belong to non-domestic users, although at this point this information is not considered, it will be explored in further detail later on in this document.

Summer

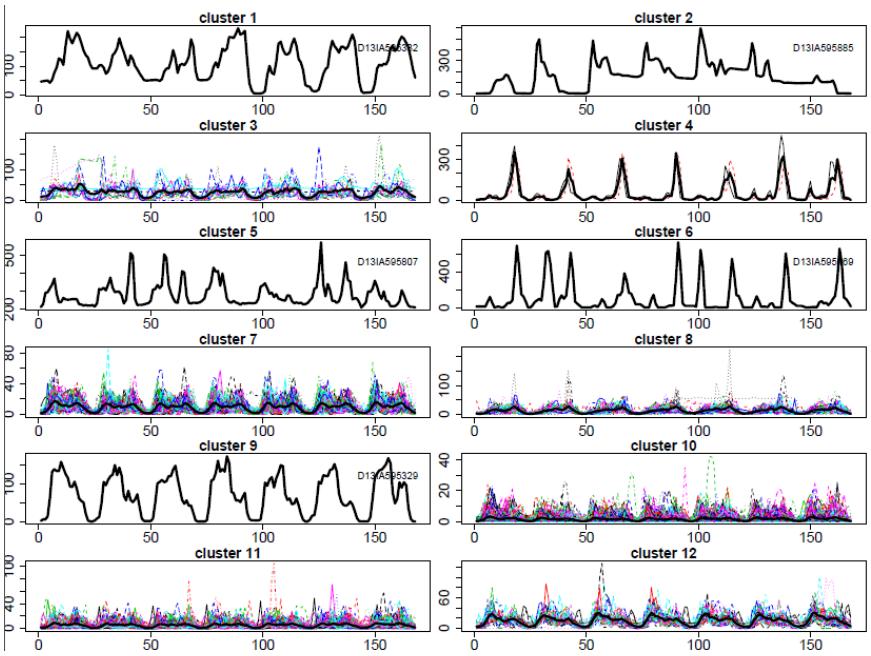


Fig. 4.12. Clustered paterns for Summer subset in Torremolinos, x axis: time as 168 hour week, y axis: mean demand l/h.

Consistent with the general behavior expected in coastal cities, Figure 4.12 shows an increase in the mean weekly consumption for most clusters. Four outliers with irregular activity and an elevated mean consumption are present, which could belong to sensors installed in irrigation sites, pools, and common areas whose use increases due to weather conditions. There are still some clusters showing a low cluster mean. These could be associated with the behavior of vacationing families.

Fall

Mean consumption drops in relation to the clustering for the summer period, as observed for most clusters in Figure 4.13. The two outliers present in clusters 3 and 4 could still belong to meters installed for pools, community use, or irrigation purposes. These uses are expected to decrease due to the end of the vocational period or climate conditions, this is the reason that explains that their mean consumption also decreases. The remainder of the clusters shows domestic patterns both of frequent low-mean-consumption daily use, and the 2-3 daily demand peaks. Further assumptions can be obtained from the billing variables provided, discussed in the following section of this chapter.

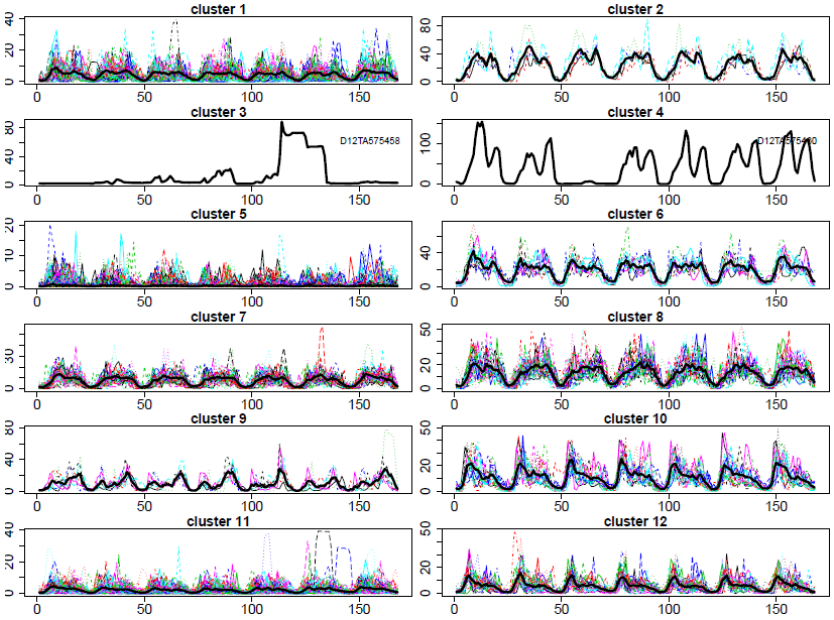


Fig. 4.13. Clustered paterns for Fall subset in Torremolinos. x axis: time as 168 hour week, y axis: mean demand l/h.

Winter

Low mean consumption domestic patterns characterize the behavior during the winter for the city, as seen in Figure 4.14. No outliers are present, which indicates the end of the vacationing period, some meters associated to uses such as irrigation might still be present, but given that their mean consumption approximates that of a domestic user, they classify as such. This same situation can be valid for meters of shared use, and/or pools.

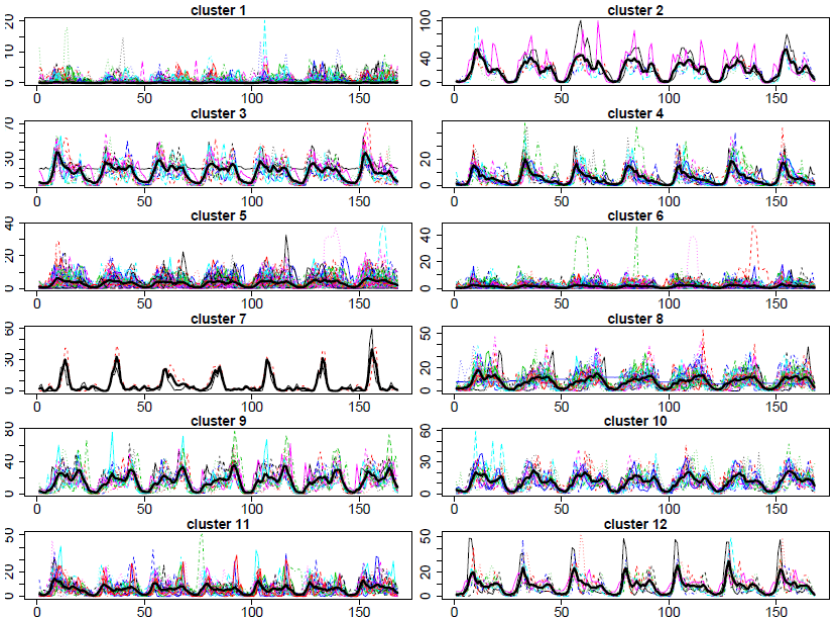


Fig. 4.14. Clustered paterns for Winter subset in Torremolinos. x axis: time as 168 hour week, y axis: mean demand l/h.

4.3. Knowledge extraction

All the clustering results enlisted in the previous section were the basis for knowledge extraction, presented in this following section. These results were matched with other variables, some obtained from the feature matrices in an additional feature extraction process, and others from datasets containing variables associated to the account of the meter, the latter provided by the company. The implementation of these new study variables allows validation of several assumptions and conclusions obtained in the previous stage of this document.

4.3.1. Seasonal Analysis

By associating the classification label to the statistical features, related to the cluster mean, it is possible to quantify the overall demand of the city. These indicators can be used to evaluate the progression of the demand throughout the year to extract the groups of sensors that behave similarly, representing the characteristic behavior of the dataset in question. A summary of statistics for the seasonal clustering of both cities are presented in the following sections.

Tarragona

Table 4.6. contains the summary of statistics for the seasonal clustered patterns corresponding to the city of Tarragona. The top three highest cluster means and the sensors associated to them are marked in boldprint. This allows to validate that outliers are responsible for the highest mean consumption in the dataset. As expected, summer is the season with the highest average, while spring represents the lowest.

	Spring		**Summer**		**Fall**		**Winter**	
Cluster	**Mean**	**Sensors**	**Mean**	**Sensors**	**Mean**	**Sensors**	**Mean**	**Sensors**
1	**26,17**	**2**	23,35	2	12,06	50	12,53	24
2	21,93	7	**33,89**	**4**	12,38	45	22,35	7
3	13,89	11	**35,58**	**1**	16,69	24	15,67	29
4	**24,05**	**1**	15,74	1	19,30	1	3,95	29
5	10,85	66	12,52	84	**47,75**	**1**	20,18	11
6	**23,42**	**2**	5,49	40	24,05	17	**32,19**	**1**
7	15,59	1	**41,14**	**2**	**27,37**	**1**	**42,00**	**1**
8	4,26	30	31,66	2	5,28	41	**23,38**	**1**
9	19,11	32	20,00	30	23,71	1	12,68	20
10	12,63	30	28,21	16	**32,12**	**1**	10,09	59

Table 4.6. Summary statistic for seasonal clustering of the Tarragona Dataset. The three clusters with highest mean demand per season are shown in boldprint.

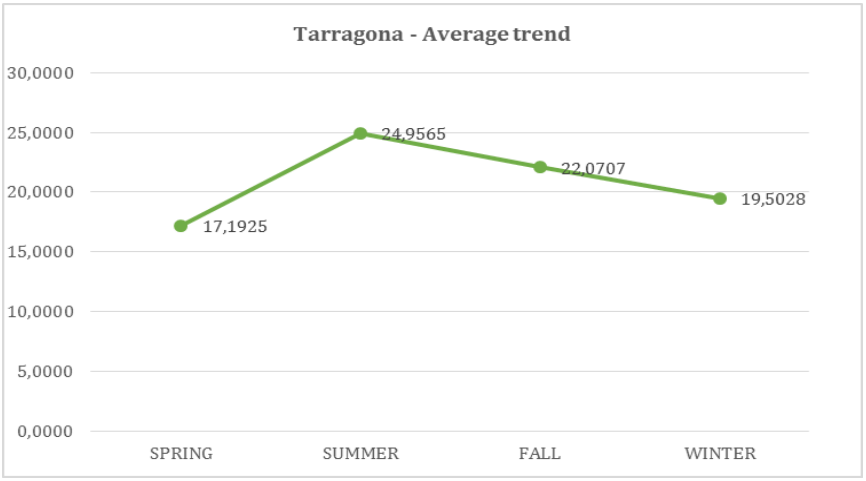


Fig. 4.15. Mean consumption (l/h) per season for Tarragona

The tendency appears to be an increase in demand during summer and fall, as observed in fig. 4.15; knowing the statistical features of each cluster and where the great majority of the sensors are condensed it is possible to validate these data. Consequently, the seasonal cluster labels are studied to extract the representative groups which fall in the same categories for all seasons. The results are available in the following chapter of this document.

Torremolinos

	Spring		Summer		Fall		Winter	
Cluster	**Mean**	**Sensors**	**Mean**	**Sensors**	**Mean**	**Sensors**	**Mean**	**Sensors**
1	18,83	21	**110,25**	**1**	4,33	51	0,11	300
2	17,86	1	**158,31**	**1**	**23,43**	**5**	**20,96**	**6**
3	3,42	127	26,63	18	12,75	1	**14,98**	**14**
4	0,59	261	55,47	2	**42,99**	**1**	5,64	16
5	**107,29**	**1**	**281,18**	**1**	0,18	333	3,08	63
6	**25,23**	**4**	106,39	1	**19,39**	**11**	1,31	100
7	6,75	51	9,15	106	7,44	33	6,23	2
8	7,25	54	13,10	49	12,89	25	7,90	29
9	**51,51**	**2**	69,90	1	9,58	7	**15,16**	**15**
10	10,75	22	1,49	187	10,79	17	11,06	16
11	12,81	51	5,09	194	2,18	84	5,75	26
12	4,41	1	15,16	35	5,72	28	8,70	9

Table 4.7. Summary of statistics for seasonal clustering of the Torremolinos Dataset. The three clusters with highest mean demand per season are shown in boldprint.

From the summary of statistics observed in Table 4.7, the mean consumption of this city is significantly higher, though most of the high consumers, once again, appear to be outliers. In all

seasons, the clusters with high number of sensors show a low mean consumption. These groups will be subsetted and further analyzed to find the most representative groups of consumers in the dataset. Figure 4.16 shows the average consumption trend for this city, where a peak in consumption during the summer can be detected.

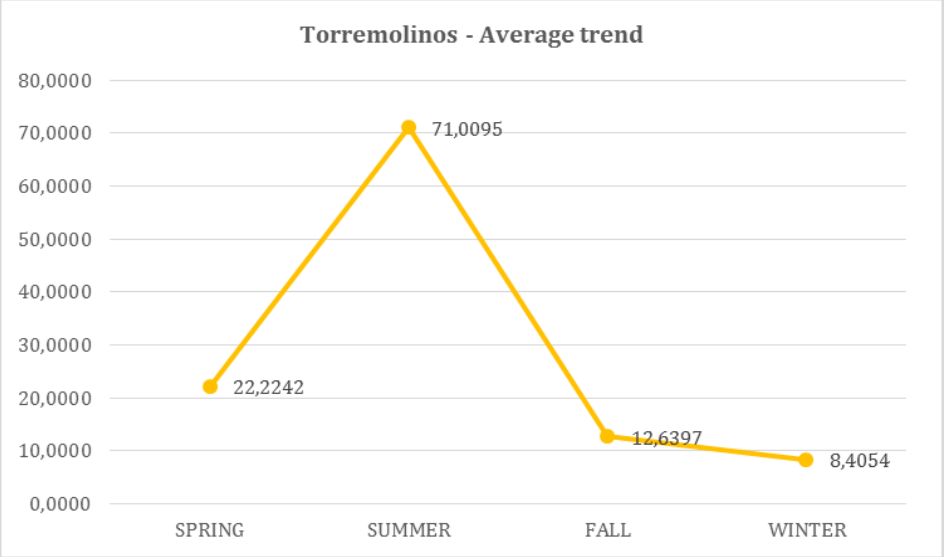


Fig. 4.16. Mean consumption (l/h) per season for Torremolinos.

The behavior of all the outliers identified using these summary of statistics was traced separately to evaluate what could be the cause for it. Since it is known that all the sensors in the dataset for Tarragona are registered as domestic, the outlying behavior could be coming from users with undeclared industrial use (fraud), some type of error in measurement associated to the meter (fault), or a fault in the distribution network (leak). In the case of Torremolinos, it must be validated that the irregularities and outlying activity is mainly caused by non-domestic sensors.

4.3.2. Classification and Regression

This section contains the extraction of knowledge from the datasets using classification and regression techniques. The industrial dataset for the city of Tarragona was not considered for this analysis, as explained previously in this chapter.

Tarragona

Using the labels obtained in a previous unsupervised clustering stage, the regression, classification, and Gradient Boosted Machines (GBM) analysis was computed. Said analysis was first computed using variables related with the account associated to the meter, and further on, using the statistics belonging to the pattern associated with the label.

The following trials were executed:

I. Using variables belonging to the account associated to the meter
   i. Regression tree
   ii. Classification tree
   iii. GBM

II. Using statistical features obtained from the pattern associated to the label.
   i. Regression tree
   ii. Classification tree
   iii. GBM

It is important to mention that for the first trial, it was possible to add an eleventh cluster containing all the outliers, including both the outliers found previous to the k-means as well as those belonging to a cluster with a single member, for this case it would be cluster 10.

As mentioned previously, it was chosen to use GBM for the purpose of this analysis. This was decided since in contrast to others, boosting is a method in which multiple models are trained in a sequence. The error function used for training a particular model depends on the performance of previous model. This will produce significant improvement in performance, which can be seen when computing relative influence of variables. For all trials, a multinomial distribution was selected, given there are more than two classes for the dataset.

Trial I: Using variables pertaining to the account associated to the meter:

For this first trial, the variables selected for the analysis were:

- Date Installed (Meter)

- Street
- Neighborhood
- Brand (Meter)
- Model (Meter)

All these variables were included in the formula that will be used to adjust the fit for the regression/classification, respectively. Table 4.8. shows part of the output from the *rpart* function; only 2 out of 5 predictors are used in the fitting of the tree: Date Installed, and Street. This is somewhat justified by the relative influence found during the computation:

Variables actually used in tree construction: **[1] Date.Installed Street** **Root node error: 1392/194 = 7.175 n= 194**

Variable importance:				
Date.Installed	**Street**	**Neighborhood**	**Brand**	**Model**
50	**35**	**13**	**1**	**1**

Table 4.8. Output to the *rpart* function showing variable importance and variables used in tree construction

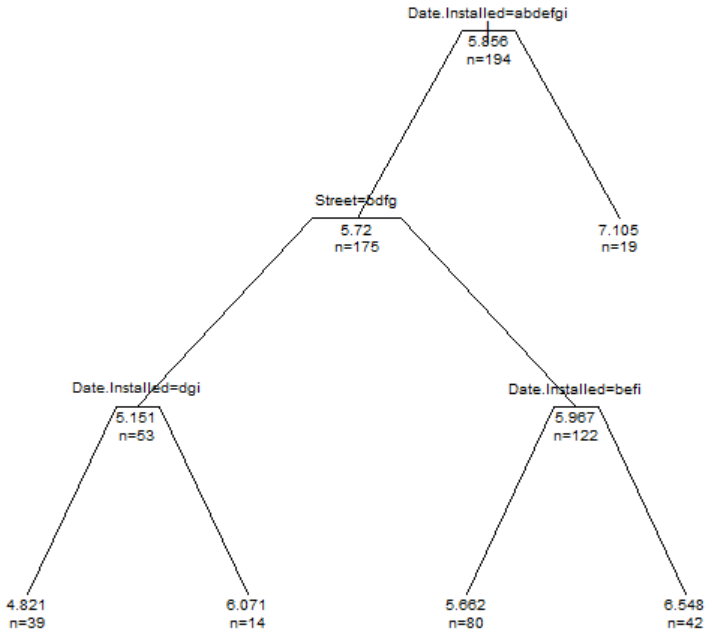


Fig. 4.17. Regression Tree for Tarragona – meter account variables

The regression tree for the domestic data is shown in Figure 4.1. Three indicators are shown at every split: text that indicates the feature chosen in each decision, the mean value for the cluster, and the value for "n" which corresponds to the number of observations; the terminal nodes only contain the last two indicators. A more extensive and detailed summary of the variables in each node is presented in Table 4.9. The 'Node' column lists all the main nodes chosen by the algorithm, 'Split' represents the point in which a decision is made and the feature taken into account for it, the cluster mean for each node is associated to the label assigned to each member. It is important to mention that the lowest deviance is presented by the terminal nodes (marked with "*")

Node	Split	n	Dev	Mean
1	Root	194	1391.95	5.855
2	Date.Installed=02/02/11,13/04/11,16/09/11,23/11/10,25/01/11,26/01/11,28/01/11	175	1197.28	5.720
3*	Date.Installed=14/02/11,27/01/11,10/09/14,29/05/14	19	161.78	7.105
4	Street=MAS D'EN GARROT,PARADA GRAN,SECA,ST.SALVADOR(S.RAMON)	53	508.79	5.150
5	Street=MAS DELS CUPS,RENGLES LLARGUES	122	663.86	5.967
8*	Date.Installed=16/09/11,26/01/11,28/01/11	39	377.74	4.820
9*	Date.Installed=02/02/11	14	114.92	6.071
10*	Date.Installed=13/04/11,23/11/10,25/01/11,28/01/11	80	271.88	5.662
11*	Date.Installed=26/01/11	42	370.40	6.547

Table 4.9. Node summary for regression tree - meter account variables

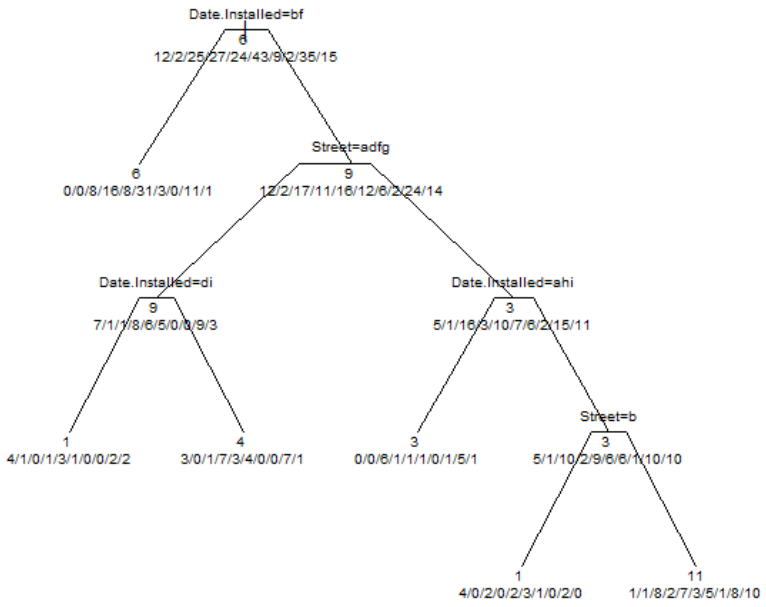


Fig. 4.18. Classification Tree for Tarragona – meter account variables

Once again, the output of the algorithm renders Date.Installed, and Street as the variables with higher importance, and will therefore use these in the construction of the tree, shown in Figure 4.18. The variables depicted in each decision are the splitting variable, label, and the member count for each cluster; meaning the terminal nodes include the members classified in that specific terminal node label.

Variables actually used in tree construction:				
[1] Date.Installed Street				
Root node error: 151/194 = 0.77835 n= 194				
Variable importance:				
Date.Installed	**Street**	**Neighborhood**	**Brand**	**Model**
51	**44**	**3**	**1**	**1**

Table 4.10. Output to the rpart function showing variable importance and variables used in tree construction

Given this is now a classification tree, the variables considered for the fitting are different. Once again the nodes shown are the most important ones chosen by the algorithm for the fitting, including the terminal ones (marked with "*"). The split includes the variables chosen for the splitting and the loss, which represents the number of members that should remain in case the splitting was done according to the Label, chosen by taking the cluster with highest probability in the variable LabelProb. Further information can be found in Table 4.11.

Node	Split	n	Loss	Pred	LabelProb
1	Root	194	151	6	(0.06 0.01 0.13 0.14 0.12 0.22 0.04 0.01 0.18 0.07)
2*	Date.Installed=13/04/11,25/01/11	78	47	6	(0 0 0.1 0.21 0.1 0.4 0.03 0 0.14 0.013)
3	Date.Installed=02/02/11,14/02/11,16/09/11,23/11/10,26/01/11,27/01/11,28/01/11,10/09/14,29/05/14	116	92	9	(0.1 0.01 0.15 0.09 0.14 0.1 0.05 0.017 0.21 0.12)
6	Street=FRANCOLI (ST.SALVADOR),PARADA GRAN,SECA,ST.SALVADOR(S.RAMON)	40	31	9	(0.17 0.02 0.02 0.2 0.15 0.12 0 0 0.22 0.07)
7	Street=MAS D'EN GARROT,MAS DELS CUPS,RENGLES LLARGUES,ALCALDE MARIAN FONTS,CASTANOS	76	60	3	(0.06 0.01 0.21 0.03 0.13 0.09 0.07 0.02 0.2 0.14)
12*	Date.Installed=16/09/11,28/01/11	14	10	1	(0.29 0.07 0 0.07 0.21 0.07 0 0 0.14 0.14)
13*	Date.Installed=14/02/11,26/01/11	26	19	4	(0.12 0 0.038 0.27 0.12 0.15 0 0 0.27 0.038)
14*	Date.Installed=02/02/11,27/01/11,28/01/11	16	10	3	(0 0 0.37 0.06 0.06 0.06 0 0.06 0.31 0.06)
15	Date.Installed=23/11/10,26/01/11,10/09/14,29/05/14	60	50	3	(0.08 0.01 0.17 0.03 0.15 0.1 0.1 0.01 0.17 0.17)
30*	Street=MAS D'EN GARROT	14	10	1	(0.29 0 0.14 0 0.14 0.21 0.071 0 0.14 0)
31*	Street=MAS DELS CUPS,RENGLES LLARGUES,ALCALDE MARIAN FONTS,CASTANOS	46	36	11	(0.02 0.02 0.17 0.04 0.15 0.06 0.11 0.02 0.17 0.2)

Table 4.9. Node summary for classification tree – meter account variables

var	rel.inf
Date.Installed	53.18514
Street	46.81486
Brand	0.00000
Model	0.00000
Neighborhood	0.00000

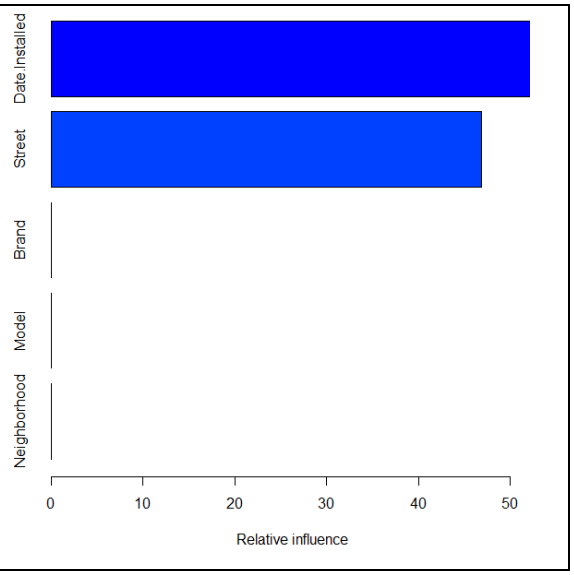


Table 4.10. GBM for domestic clustering, relative influence of meter account variables

It can be concluded that the output to this algorithm is overgeneralized, given the improvement is not significant, neglecting completely three of the predictors. This can be due to overfitting caused by the distribution of the variables considered, given that 188 of the sensors share the same brand/model. This model could be trialed for prediction, but given the forced nature of the input variables the results are not expected to be good. The relative importance values of each variable are averaged and presented in Table 4.11. These will be considered for the trials using the dataset for Torremolinos.

Variable	**Relative Importance**
Date Installed	51.3950
Street	41.9383
Brand	5.3333
Model	0.6667
Neighborhood	0.6667

Table 4.11. Average relative importance of all meter account variables

Trial II: Using statistical features of the pattern associated to the cluster:

Using the same feature matrix considered for the *k*-means algorithm five different indicators were extracted

- Mean
- Maximum
- Minimum
- Variance
- Standard Deviation

It is expected to obtain better results using these indicators since they are variables associated directly to the demand pattern, thus not representing a forced relationship among the input variables. These variable were all included in the formula that will be used to adjust the fit for the regression/classification, and *GBM* respectively.

Table 4.2. shows part of the output from the *rpart* function; three of the 5 predictors are used for the construction of the tree: Mean, Var, and Max. In contrast to the previous trial, these are not the same three with highest relative influence.

Variables actually used in tree construction:	**Variable importance:**				
[1] Max Mean Var	**Mean**	**SD**	**Var**	**Max**	**Min**
Root node error: 984.84/182 = 5.4112 n= 182	**29**	**24**	**24**	**17**	**6**

Table 4.12. Output to the rpart function showing variable importance and variables used in tree construction

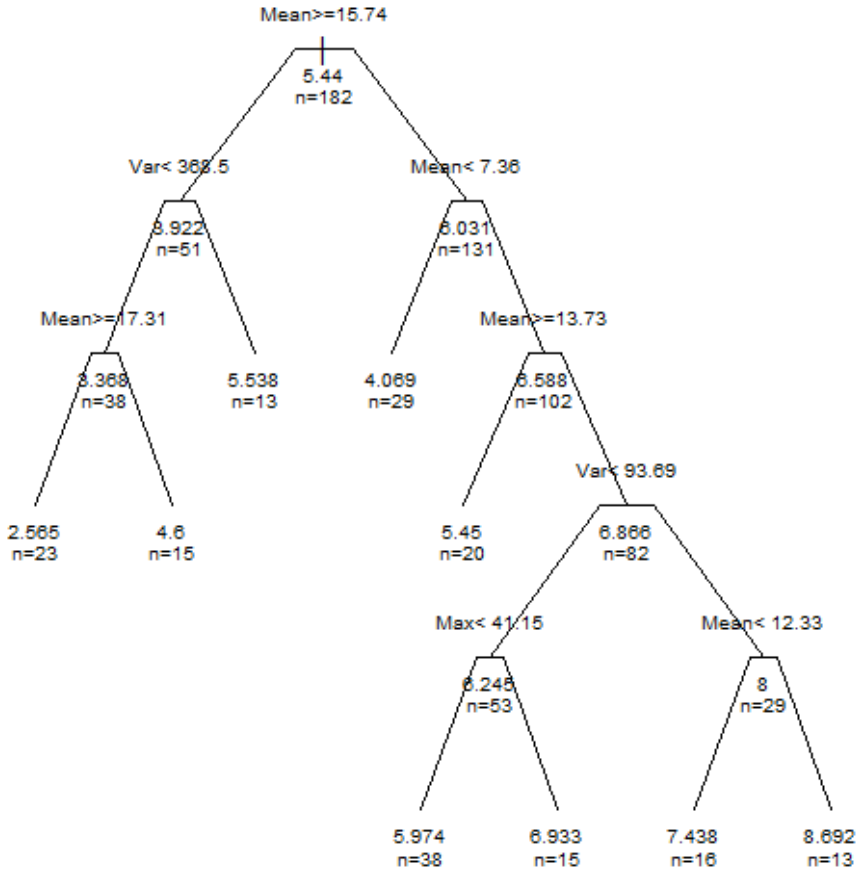


Fig. 4.19. Regression Tree for Tarragona – Statistical features

It is possible to conclude that when using numerical variables the results can be significantly improved. This is shown by the variance, which at the terminal nodes proves to be quite low. The split shows the variables used in the tree construction and the value selected for splitting. Further information is available in the chart below. The cluster mean for each node is obtained when averaging the cluster labels for all members in the node, thus being the variable of interest for the splitting.

Node	Split	Members	Deviance	Cluster Mean
1	Root	182	984.835	5.439560
2	Mean>=15.74456	51	337.686	3.921569
3	Mean< 15.74456	131	483.877	6.030534
4	Var< 368.5108	38	182.842	3.368421
5*	Var>=368.5108	13	109.230	5.538462
6*	Mean< 7.360466	29	3.862	4.068966
7	Mean>=7.360466	102	336.705	6.588235
8*	Mean>=17.30994	23	71.652	2.565217
9*	Mean< 17.30994	15	73.600	4.600000
14*	Mean>=13.72713	20	100.950	5.450000
15	Mean< 13.72713	82	203.524	6.865854
30	Var< 93.69199	53	67.811	6.245283
31	Var>=93.69199	29	78.000	8.000000
60*	Max< 41.15324	38	24.973	5.973684
61*	Max>=41.15324	15	32.933	6.933333
62*	Mean< 12.33239	16	51.937	7.437500
63*	Mean>=12.33239	13	14.769	8.692308

Table 4.13. Node summary for regression tree - Statistical features

Fig. 4.20. Classification Tree for Tarragona – Statistical variables

From the chart below, the most significant improvement can be seen in the "Label" variable; ease of classification when using statistical features can be concluded, specifically in the terminal nodes.

Node	Split	Members	Loss	Label	LabelProb
1	Root	182	139	6	(0.066 0.011 0.14 0.15 0.14 0.24 0.049 0.011 0.19 0.0055)
2*	Mean< 7.360466	29	1	4	(0 0 0 0.97 0 0.034 0 0 0 0)
3	Mean>=7.360466	153	111	6	(0.078 0.013 0.16 0 0.16 0.27 0.059 0.013 0.23 0.0065)
6	Var>=93.69199	100	72	9	(0.12 0.02 0.25 0 0.17 0.04 0.09 0.02 0.28 0.01)
7*	Var< 93.69199	53	15	6	(0 0 0 0 0.15 0.72 0 0 0.13 0)
12	Mean>=13.72713	71	46	3	(0.17 0.028 0.35 0 0.17 0.014 0.13 0.028 0.099 0.014)
13*	Mean< 13.72713	29	8	9	(0 0 0 0 0.17 0.1 0 0 0.72 0)
24*	Mean>=21.57929	19	11	1	(0.42 0.11 0 0 0 0 0.32 0.11 0 0.053)
25	Mean< 21.57929	5	27	3	(0.077 0 0.48 0 0.23 0.019 0.058 0 0.13 0)
50*	Mean>=16.13931	29	11	3	(0.14 0 0.62 0 0.069 0 0.1 0 0.069 0)
51*	Mean< 16.13931	23	13	5	(0 0 0.3 0 0.43 0.043 0 0 0.22 0)

Table 4.14. Node summary for classification tree - Statistical features

var	rel.inf
Mean	66.971119
Var	25.306322
Max	6.402886
Min	1.319673
SD	0.000000

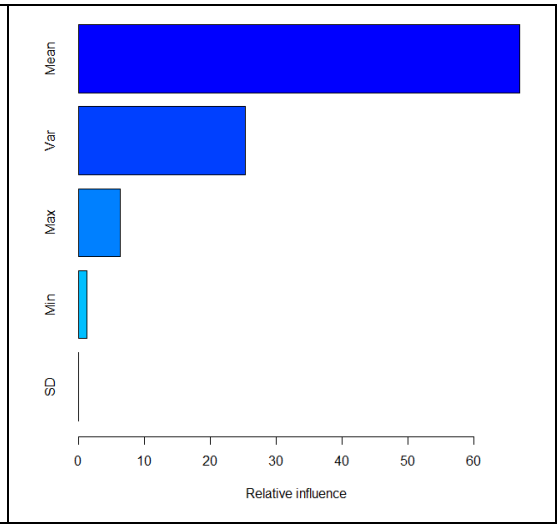


Table 4.15. GBM for domestic clustering - Relative influence of statistical variables

The output shows that out of the 5 predictors given, only 4 have influence, the mean being the most important one, which was also true for the regression and classification trees. As expected, significant improvement is shown associated to the relative influence of each variable. This becomes useful when computing predictions associated to the variables in question, thus giving a better fit for the model.

Overall, the influence of the variables chosen for this analysis is as follows:

Variable	Relative Influence
Mean	41.6570
Var	24.4354
Max	13.8009
Min	12.3196
SD	16

Table 4.16. Average relative importance of all statistical variables

The benchmarking of the regression, classification, and boosting models explained in this section was useful to become acquainted with the parameters involved, their functionalities and the ways in which these can improve the outcome of the results obtained. No further analysis was performed on this dataset, since it was chosen to focus on the dataset for Torremolinos. A few considerations were made for the tuning parameters and were implemented to the computation of the algorithms in the following stage:

- The complexity parameter (cp) computes the weight of adding another split to the tree. Therefore, it is reasonable to think that this parameter controls the size of the tree and helps choosing the optimal size. For large datasets, it is preferred to avoid pruning since the reduction of variance is not specifically useful [12]. Knowing this, it is decided to tune *cp* to obtain greater depth in the tree construction and observe if this reduces the training and testing error. The default value for *cp* is 0.01.
- Since it is now desired to know if these models can be trained for the computation of prediction, the datasets were split approximately 80% for training and 20% for testing.
- For the computation of GBM, an iterative process was followed to obtain the progression of the training error. The number of trees used was set to 5000, which increased the computation time, but achieved error minimization on the training dataset. A multinomial distribution was used for the computation given the nature of the data, which is split into 12 clusters.
- Knowing the relative influence of all the variables in each trial, it was decided to combine all of these to also reduce error and find out which of these provide the best predictions.

Torremolinos

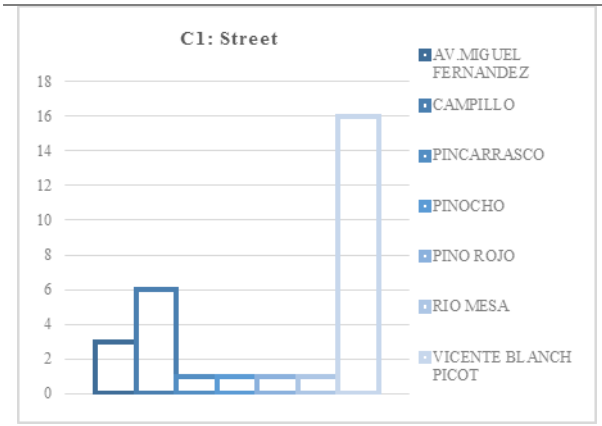
The same trials were considered for this dataset:

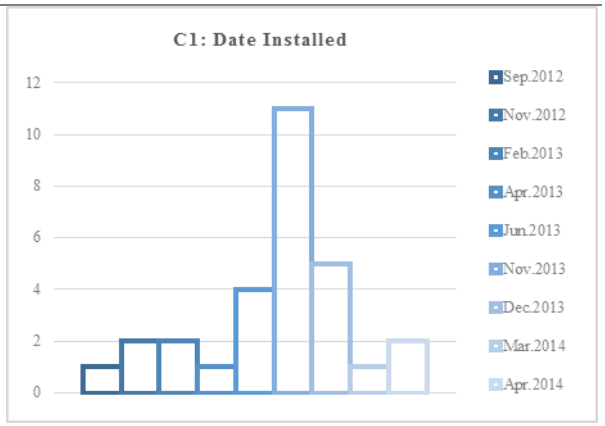
I. Using variables pertaining to the account associated to the meter
   iv. Regression tree
   v. Classification tree
   vi. GBM
II. Adding statistical features obtained from the pattern associated to the label.
   vii. Regression tree
   viii. Classification tree
   ix. GBM

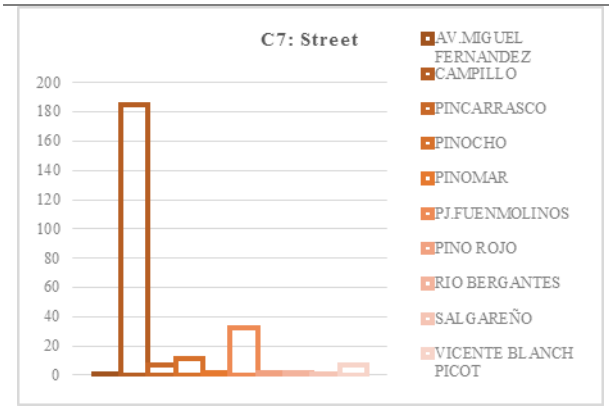
Prior to the computation of the algorithms, some statistical features from the clusters were computed, given the distribution of the variables and number of sensors in each cluster only the ones shown in Table 4.16 were considered for the regression and classification analysis. The remainder part of the analysis shows the following features:

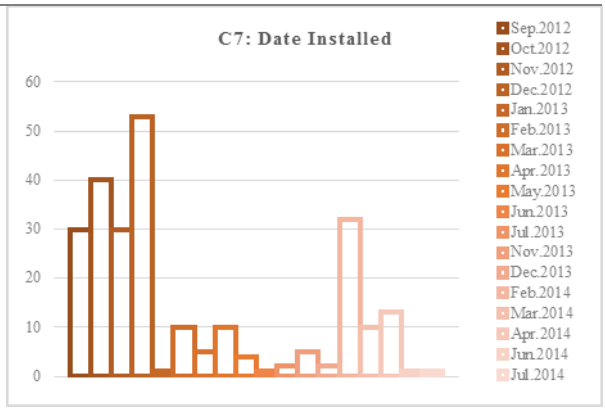
- Cluster 2: 5 sensors, 3 of them registered as industrial
- Cluster 3: 1 sensor with registered irrigation/wash use
- Cluster 4: 1 sensor with registered use for pools
- Cluster 5: 1 sensor registered as use for community owners
- Cluster 6: 1 sensor with registered use for pools
- Cluster 11: 2 sensors, one of which is the only member registered to sector 7

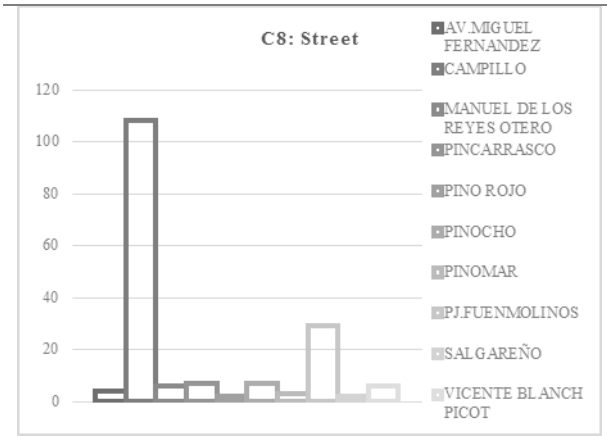
With this it is observed how setting the number of clusters to 12 extracts most of the non-domestic users. By neglecting these for the regression and classification the prediction for each class is simplified.

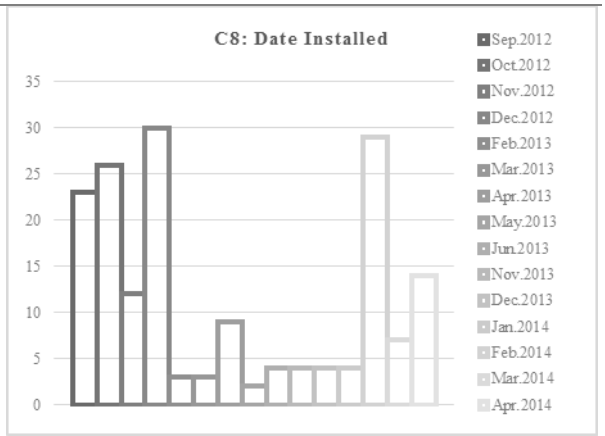

C1: Street
AV.MIGUEL FERNANDEZ
CAMPILLO
PINCARRASCO
PINOCHO
PINO ROJO
RIO MESA
VICENTE BLANCH PICOT
0
2
4
6
8
10
12
14
16
18

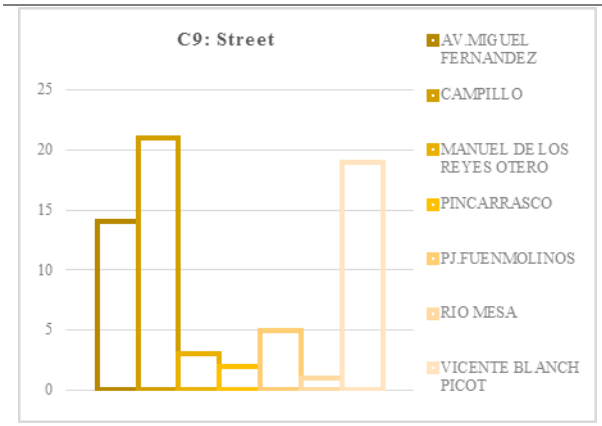

C1: Date Installed
Sep.2012
Nov.2012
Feb.2013
Apr.2013
Jun.2013
Nov.2013
Dec.2013
Mar.2014
Apr.2014
0
2
4
6
8
10
12

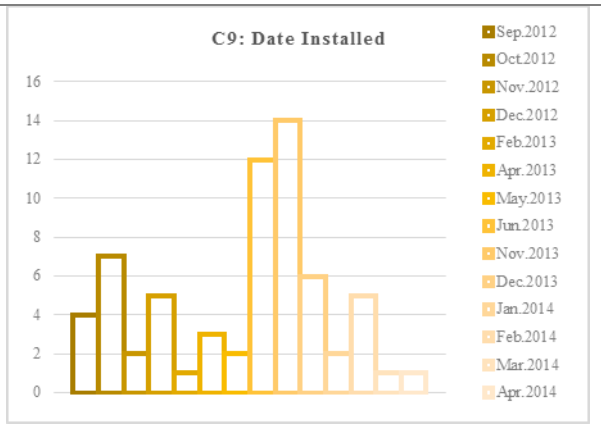

C7: Street
AV.MIGUEL FERNANDEZ
CAMPILLO
PINCARRASCO
PINOCHO
PINOMAR
PJ.FUENMOLINOS
PINO ROJO
RIO BERGANTES
SALGAREÑO
VICENTE BLANCH PICOT
0
20
40
60
80
100
120
140
160
180
200

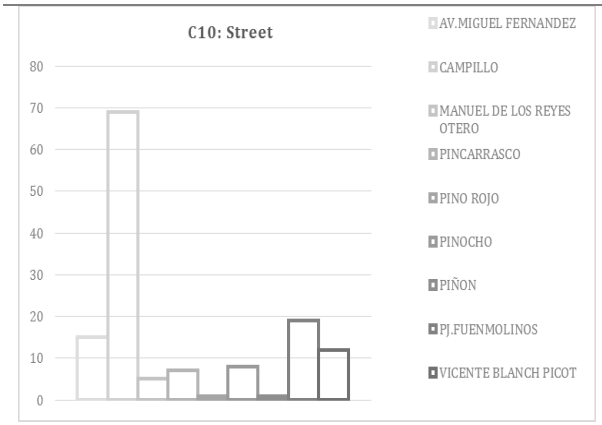

C7: Date Installed
Sep.2012
Oct.2012
Nov.2012
Dec.2012
Jan.2013
Feb.2013
Mar.2013
Apr.2013
May.2013
Jun.2013
Jul.2013
Nov.2013
Dec.2013
Feb.2014
Mar.2014
Apr.2014
Jun.2014
Jul.2014
0
10
20
30
40
50
60

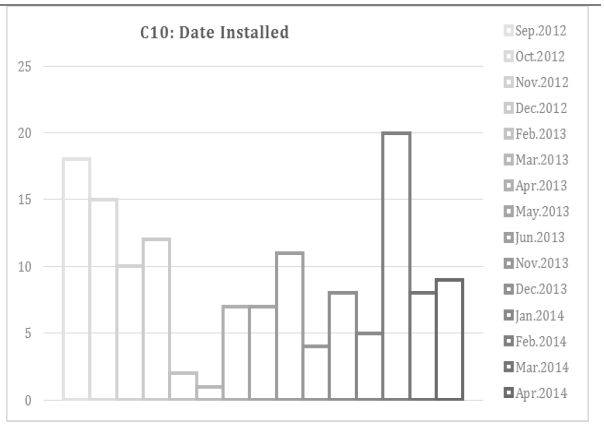

C8: Street
AV.MIGUEL FERNANDEZ
CAMPILLO
MANUEL DE LOS REYES OTERO
PINCARRASCO
PINO ROJO
PINOCHO
PINOMAR
PJ.FUENMOLINOS
SALGAREÑO
VICENTE BLANCH PICOT
0
20
40
60
80
100
120

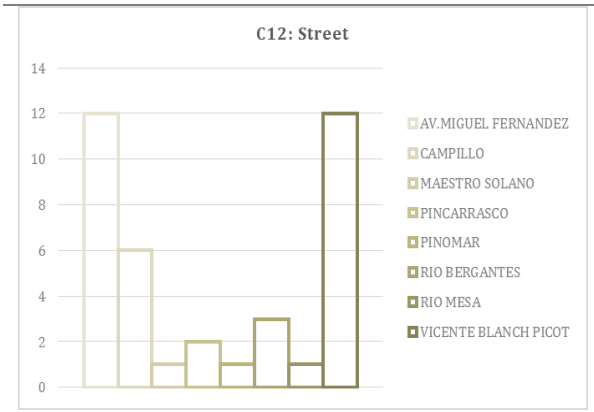

C8: Date Installed
Sep.2012
Oct.2012
Nov.2012
Dec.2012
Feb.2013
Mar.2013
Apr.2013
May.2013
Jun.2013
Nov.2013
Dec.2013
Jan.2014
Feb.2014
Mar.2014
Apr.2014
0
5
10
15
20
25
30
35











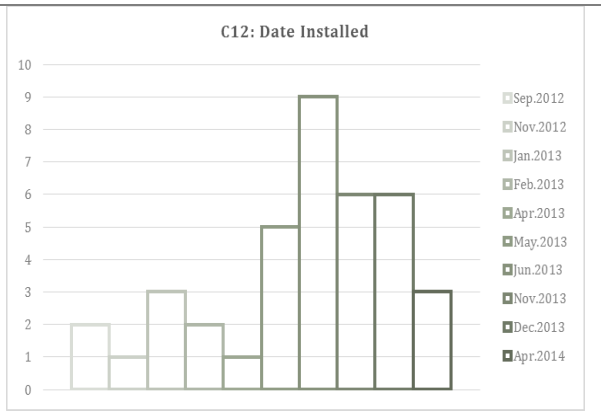


Table 4.16. Distribution of streets and installation dates per cluster

Trial I: Using variables belonging to the account associated to the meter:

For this first trial, the variables selected for the analysis were:

- Caliber: 13 for domestic users, 15 for industrial and other types of users

- Sector: the dataset contains meter from four different sectors, although two of these sectors (8 and 15) account for ~90% of the data.
- Brand*
- Model*
- Street: 16 different streets are available as could be seen in Table 4.16
- Date Installed: the initial dd-mm-yyy format resulted in 80 different streets, which represents more factors than can be processed with the software used. This variable was reformatted to include only the month and year, reducing the number of factors to 19, this allowed the use of all the models selected.
- Type: five different user types are included in the dataset, most of these being domestic, and the rest identified as atypical activity.

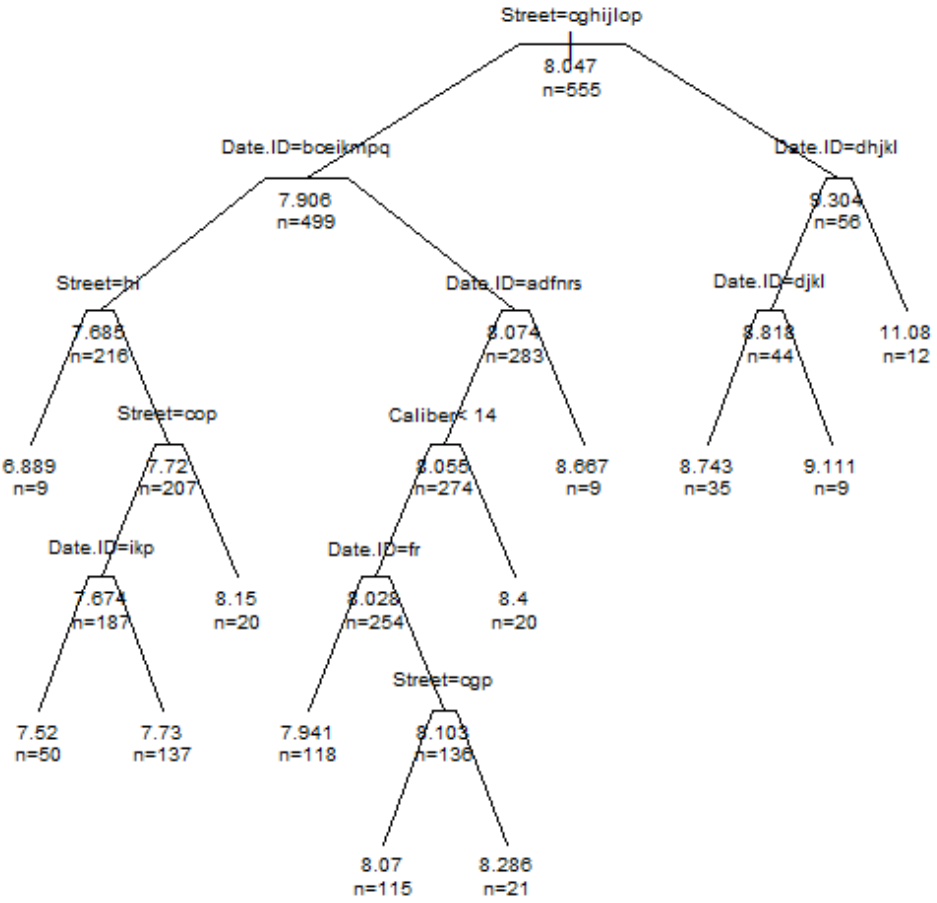


Fig. 4.21. Regression Tree for Torremolinos – meter account variables

Figure 4.2 shows the regression tree model obtained for the training dataset (n = 555). This model will later be used on the test dataset to compute the prediction. The variables taken into account for the construction of the tree are: Date.ID, Street, and Caliber. The value of *cp* was set to 0.0001, which allows greater depth for the construction of the tree without allowing overfitting.

The classification tree model is shown in Figure 4.22 where a significant training error is seen at the terminal nodes, explained by the forced nature of the input variables. The value for *cp* was set to 0.002, where it was seen that the error in the predictors was lower than the baseline error.

The prediction output to these models is shown in Table 4.17.

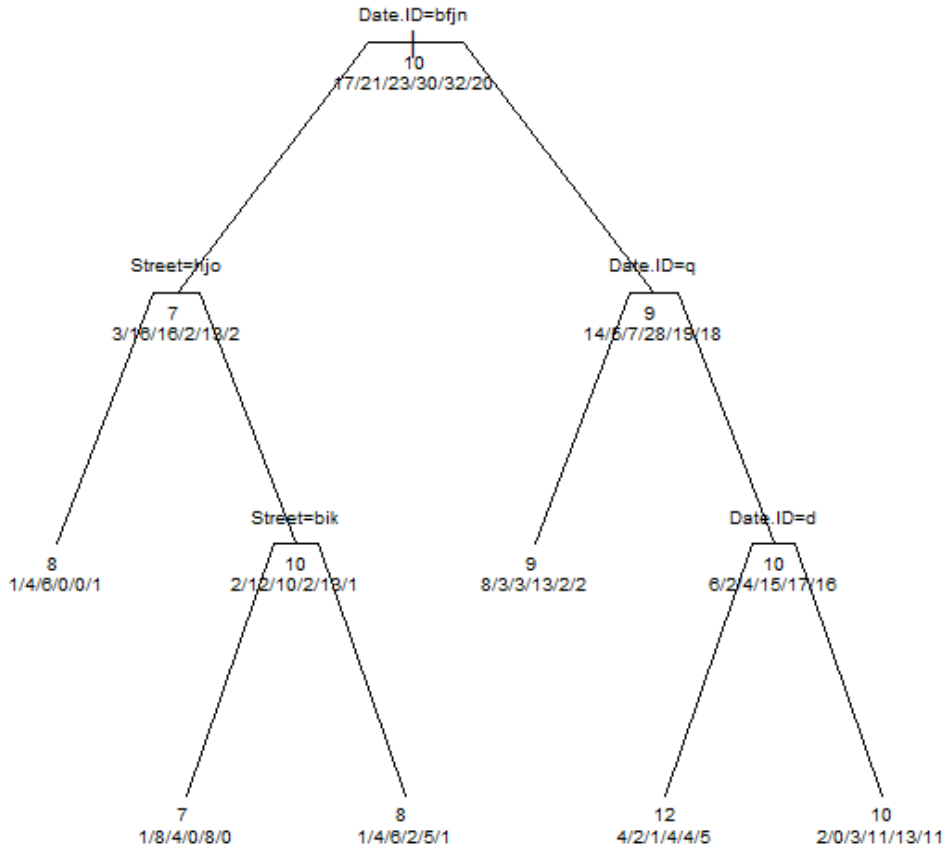


Fig. 4.22. Classification Tree for Torremolinos – meter account variables

Trial II: Adding statistical features of the pattern associated to the cluster:

Using the same feature matrix considered for the *k*-means algorithm some statistical indicators were extracted:

- Mean
- Maximum
- Minimum
- Variance

These indicators were combined with the meter account variables mentioned in the first trial in order to evaluate the performance using all the variables with high relative influence. With this it is expected to obtain error minimization and determine the most adequate variables to be studied.

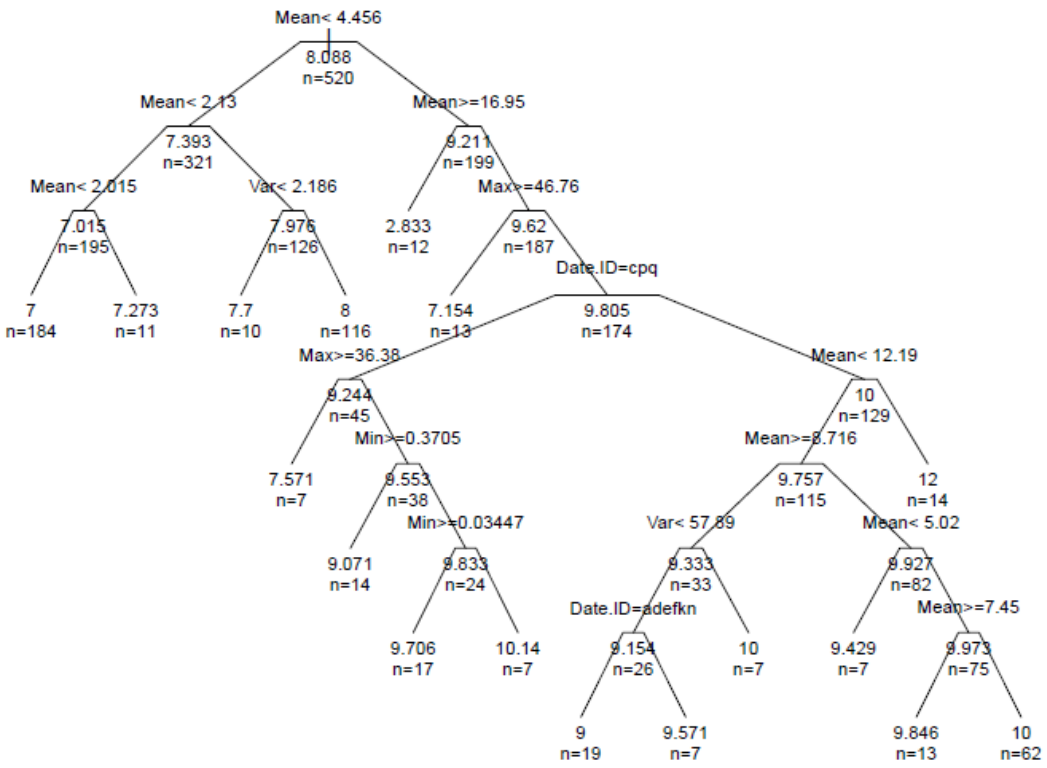


Fig. 4.23. Regression Tree for Torremolinos –statistical variables added

Although the error is minimal early on in the construction of the tree, a value of cp =0.0001 is set in order to explore all the variables that can possibly be involved in the computation. This results in the regression tree shown in Figure 4.23, where it is observed that the varables

considered are: Mean, Var, Max, Min, and Date.ID. CP is set to 0 for the computation of the classification tree, which leaves us with a tree constructed to its full depth. The mean being the only variables considered for its construction. The tree is shown in Fig. 4.24, showing a minor classification error at the terminal nodes.

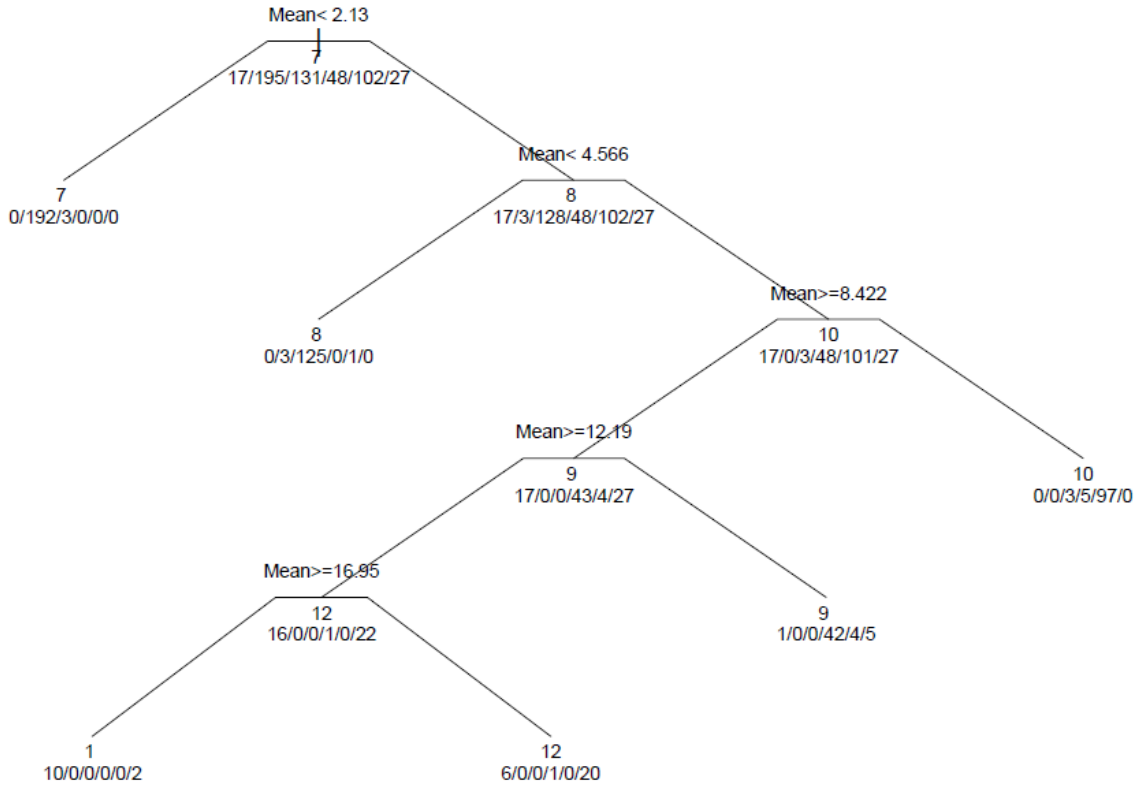


Fig. 4.24. Classification Tree for Torremolinos –statistical variables added

Table 4.17 shows the prediction output to each of the models used for both trials. The regression and classification matrices shown for trial I demonstrate misclassification, and bias for class 10. A quasi-diagonal matrix, such as the ones seen for trial II, suggest low prediction error, which means the statistical features perform much better as predictors, given these are strictly associated to the demand pattern, this result is expected. The results for GBM show a stable error for all iterations, an once again, this error appears much lower for the second trial.

In an attempt to explore further, the datasets were partitioned for sector 8 and sector 15, the prediction and traning errors are shown in Chapter 5.

Trial I

Regression

Prediction	1	7	8	9	10	12
7.52	0	8	2	0	1	1
7.72992700729927	5	16	7	3	4	0
7.94067796610169	0	12	7	2	8	0
8.0695652173913	0	6	8	4	6	1
8.15	0	0	4	0	6	0
8.28571428571429	1	0	1	0	1	0
8.4	0	2	2	1	1	0
8.66666666666667	0	0	0	0	2	0
8.74285714285714	0	0	1	2	2	4
9.11111111111111	0	0	0	2	0	0
11.0833333333333	1	1	0	2	0	1

Trial II

Regression

Prediction	1	7	8	9	10	12
2.83333333333333	4	0	0	0	0	1
7	0	53	0	0	0	0
7.15384615384615	6	0	0	0	0	0
7.27272727272727	0	1	0	0	0	0
7.57142857142857	1	0	0	0	1	1
7.7	0	0	1	0	0	0
8	0	1	36	0	0	0
9	0	0	0	3	0	0
9.07142857142857	0	0	1	4	2	1
9.42857142857143	0	0	5	0	4	0
9.57142857142857	0	0	0	3	1	0
9.70588235294118	0	0	0	0	1	0
9.84615384615385	0	0	0	2	3	0
10	0	0	0	3	22	2
10.1428571428571	0	0	0	0	1	0
12	1	0	0	2	0	6

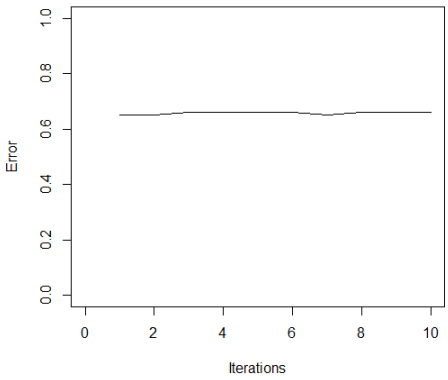
Classification

Prediction	1	7	8	9	10	12
1	0	0	0	0	0	0
7	1	7	6	1	8	0
8	0	0	5	0	5	0
9	4	1	0	0	1	0
10	2	37	20	14	16	6
12	0	0	1	1	1	1

Classification

Prediction	1	7	8	9	10	12
1	4	0	0	0	0	1
7	0	54	0	0	0	0
8	0	1	40	0	0	0
9	2	0	0	13	1	2
10	0	0	3	1	34	0
12	6	0	0	3	0	8

GBM



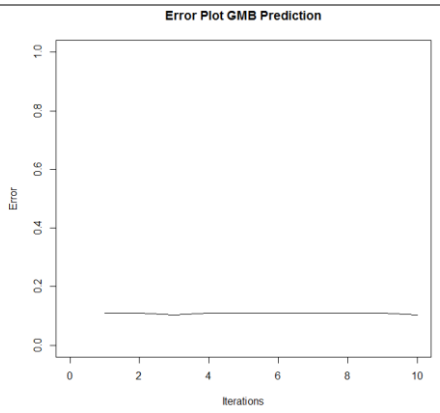


Table 4.17. Output to prediction models

5. Results

5.1. Seasonal analysis

For both cities, using the classification labels obtained from the k-means algorithm, several groups of representative behavior were extracted. Outliers and high consumers were also identified, for the city of Torremolinos it was possible to explain this behavior by cross referencing the sensor data with the meter variables

5.1.1. Tarragona

The groups were categorized according to the pattern they follow from Spring to Summer, taking into account the cluster these are moving to as a group altogether. Many more of these groups can be identified, but are not shown in the plot given the reduced number of members each of these included. All of these groups are made up of domestic users, as is the complete dataset provided for the city, which means these can be generalized as domestic behavior for the city.

Three main groups of consumers were identified, as seen in Figure 5.1. The mean consumption for these groups does not surpass 12,6 l/h. For the most part, the greatest average consumption is observed during the summer and fall months, while the lowest average is seen during the winter.

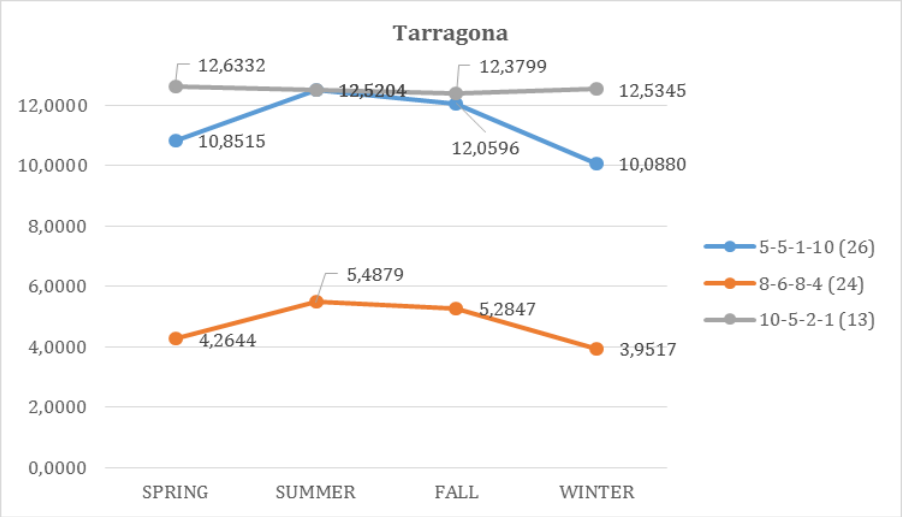


Fig. 5.1. Representative behavior groups in Tarragona.

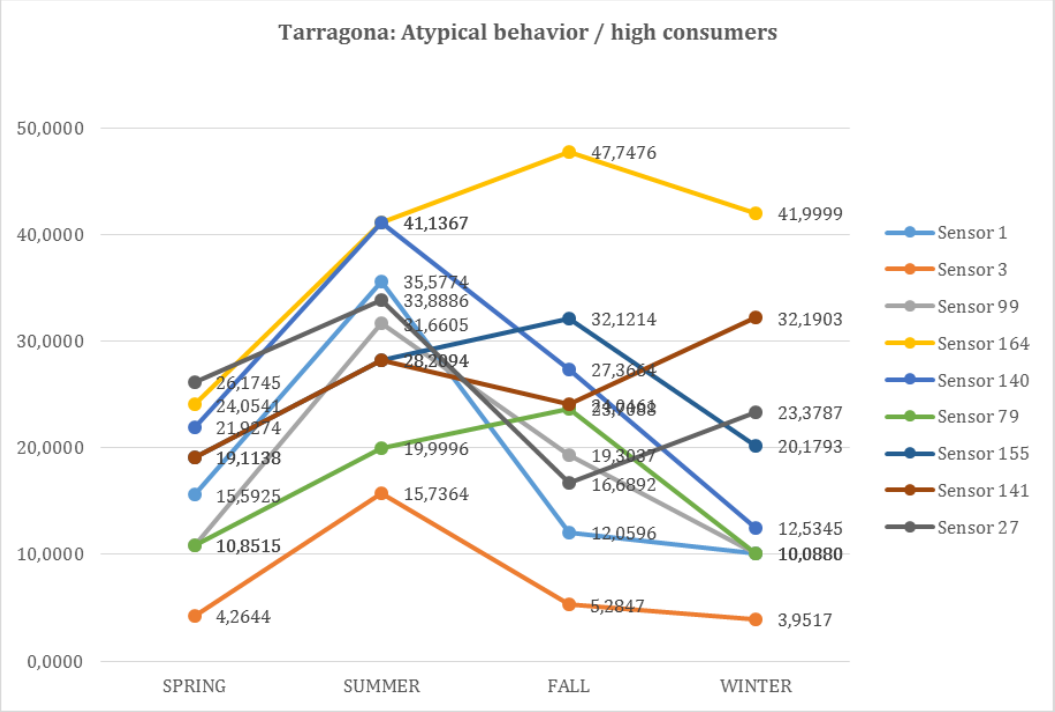


Fig. 5.2. Outlier behavior in Tarragona

The plot shown in Figure 5.2 depicts the behavior observed for all the outliers and high consumers in the clustering for each season. Given all of these are domestic users more indicators than those provided might be needed in order to draw sustained conclusions. Outliers showing a pronounced peak during the summer could be houses that contain pools, as well as summer homes, or vacation homes in general, for the case of sensors 121, and 47.

For the case of sensor 164 more information would be needed to explain the high consumption during the greater part of the year, which could be associated with the members of the family or undeclared activity. For all these matters, it would help to have more geographical information as well. This way it would be possible to plot the location for the sensors in an attempt to find features associated to the dwelling, such as the size or the existence of a pool or pond.

For the case of sensor 1, although the data exists and it is included in the dataset as a domestic user, it does not figure in the meter account variables. This meter could be registered as a separate use, or it could have been uninstalled due to failure; no indication of this was found in any of the databases, which is why it is considered as a loss.

5.1.2. Torremolinos

The behavior for these data set was traced in the same manner as those for the city of Tarragona. Three main groups identified for this city, and are shown in Figure 5.3. All of these group show a very low mean consumption during fall and winter, and a peak during the summer, meaning these meters could be installed in summer homes, or in houses with small children who remain in the city during summer vacation. Overall, these groups do not seem to differ much from the general behavior observed in Fig. 4.16, for the exception of group 3-10-5-1.

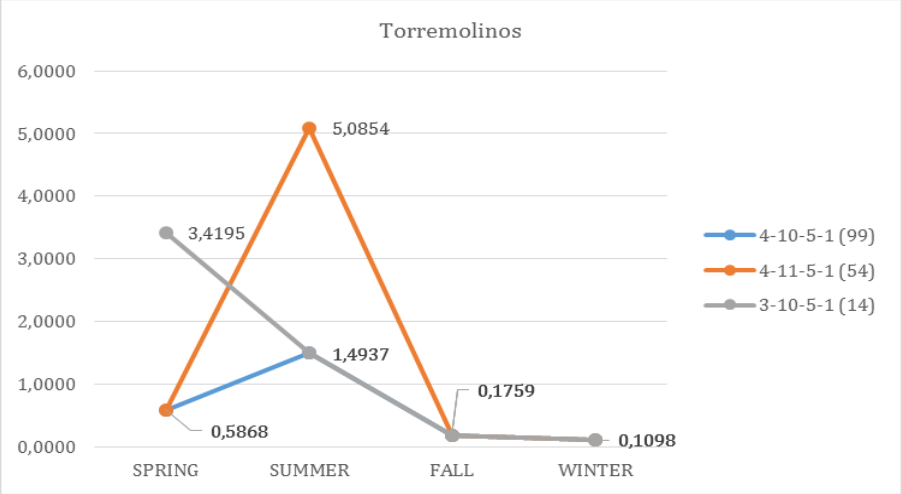


Fig. 5.3. Representative behavior groups in Torremolinos

The behavior associated to the outliers and high consumers is shown in Fig. 5.4. Although most of these shown the pattern observed for general behavior, it is evident how the mean cnsumption is significantly higher in all seasons. It was possible to match all of these sensors to their meter account, where the following cases are identified.

- Sensors 510, 588, 506, and 523 are all registered as domestic users, and although they share in the general behavior, it is alarming how much higher their mean values deviate from the expected. For the case of sensor 510, which shows the highest mean consumption for spring and summer, the pattern seen in the clustering shows irregular behavior and deviation from the baseline. This indicates that there could either be irregular activity during the peak months, a minor leak, or an additional element in the household responsible for this excess in demand, such as a pool or it being a seasonal

rental dwelling.

- Sensors 449, 452, 330, and 331 are registered as industrial use; no specific detail about this activity is known, but it justifies the increase in mean demand. This does not exclude the possibility for a fault to be present, such as the case for sensor 452, which shows a clear baseline deviation during the Summer (cluster 1).

- Sensor 560, the second highest average consumer of the dataset during the Summer, registers as community uses, justifying the increase durng these months given the touristic nature of this city.

Most of these sensors maintain the low consumption rate during fall and winter; their atypical behavior in the remaining months can therefore be attributed to wrongful use, the rest of these that do not comply with said behavior could be associated with a leak or some characteristic undeclared behavior. For those registered as domestic users, it would be recommended to analyze other intrinsic variables.

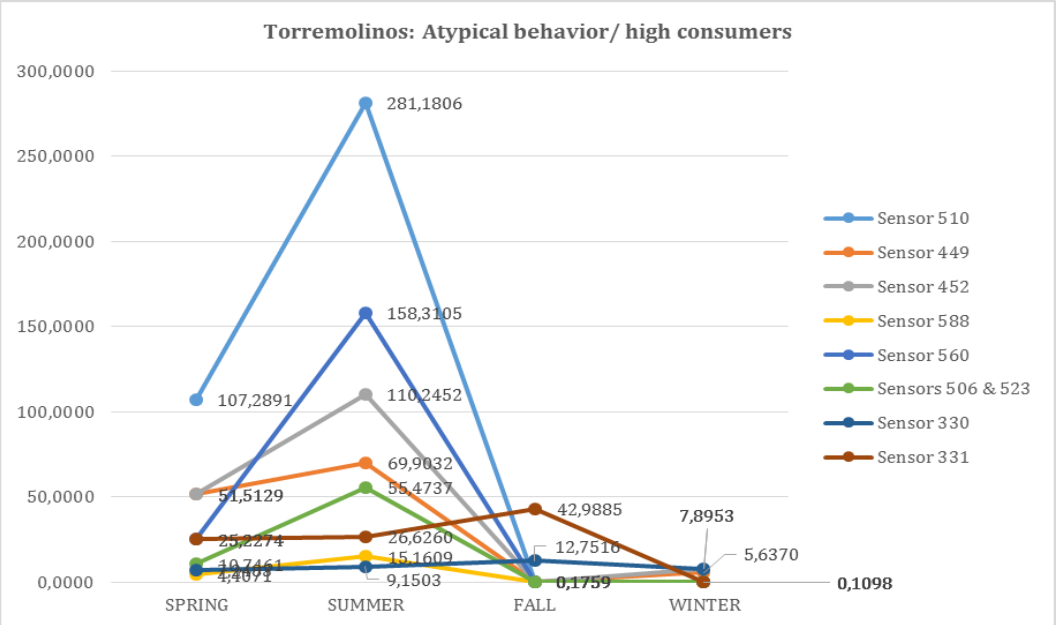


Fig. 5.4. Outlier behavior in Tarragona

5.2. Classification and regression

Only the results for the dataset provided for Torremolinos are considered for this section. All the trials considered for the purpose of this stage in analysis are mentioned in Chapter 4. The results varied according to the model and the variables selected. For the purpose of summarizing the obtained results, the training and testing errors were computed for each method. In Figure 5.5, the error plot shown corresponds to trial I, in which only the meter account variables were considered for the regression, classification, and GBM models. The highest error in training can be seen for the computation of GBM, due to the forced nature of the input variables, this high error is caused due to overfitting.

Overall, it can be concluded that these predictors alone may not represent the stringent input-output relation desired; which motivated the working out of trial II, where statistical features extracted from the pattern were also used as predictors. In an attempt to improve the results, as well as for observation, the dataset was partitioned by sector and two of these were analyzed separately. Granting the desired results were not obtained, this was considered a useful tool, provided that each sector (DMA) has separate characteristics.

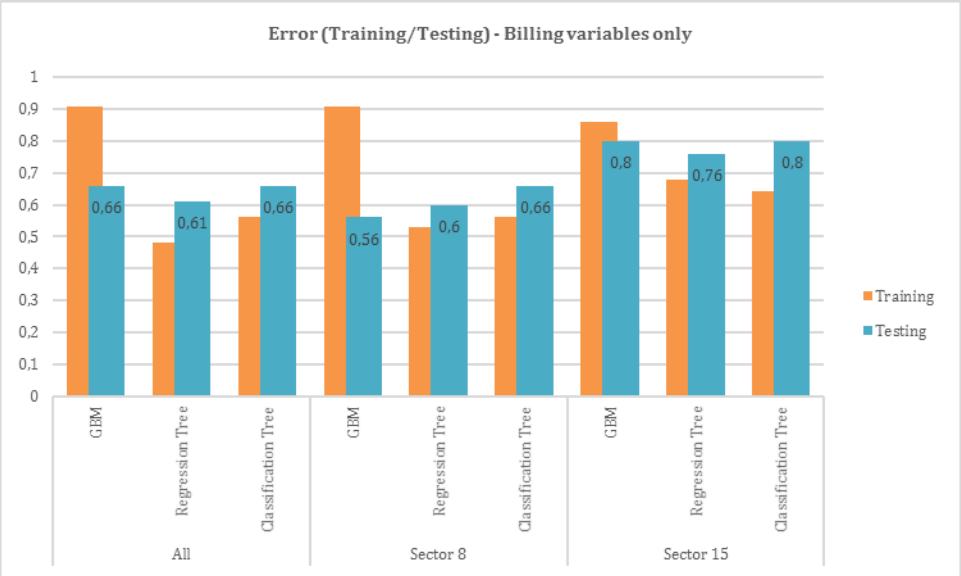


Fig. 5.5. Training and testing errors using meter account variables

The results for the second trial are shown in Figure 5.6. As previously stated, statistical features were added to the list of predictors, proving that these, along with the Date.ID form the group of predictors with the highest infuence thus far. Minimal training error is seen for all

three methods , on the whole, all models perform well, nevertheless the best performance in testing is seen when using GBM. In view of the overall good performance otained, sector partitioning was not considered for this trial; given the reliability of the predictors used, it is expected that this partitioning would improve the results even further.

Fig. 5.6. Training and testing errors using meter account variables and statistical features

Conclusions and future work

The use and implementation of these techniques has proven successful in the attainment of useful information. By partitioning the data in a temporal scale, such as the one used for the seasonal analysis, more could be understood about the behavioral aspects of the city, even more so when matching these to meter account variables; giving a justified approach to the behavior observed. By analyzing the clustering labels, it is possible to track groups with common behavior and make an educated guess as to what could be the nature of said behavior.

As far as linking the demand pattern with any kind of external variable, it allows to understand that the closer this variable relates to the features of the pattern, the easier it is to predict this sort of behavior. This assumption is valid since the demand pattern originates directly from human behavior, which could have some relation to the declared use but, for the most part, there are some other factors involved. Regrettably, some of the variables that could best indicate the desired affiliation were not available during the execution of this project. Some of the useful indicators for further study could be:

- Number of inhabitants per household
- Type of inhabitant per household
    - Single parent
    - Large/small family
    - Newlyweds
    - Retirees

The use of geographical variables could also be an important element for the purpose of this study, although the address for the registered account was provided, this did not perform well as a predictor, which could be possibly explained by the fact that not enough data was available. This is best observed using the map of Torremolinos shown in the following page, where the addresses linked to the meter accounts are marked by use, and color coded by sector. This first map contains 1215 sensors available in the area, which represent a small fraction of the living accommodations in the city.

A major drawback in attempting to use these variables is the distribution of the data. While some of it might appear sparse, there are sections in which these are tightly packed. Another significant drawback is the fact that a group of sensors with numerous types of behavior could be condensed in the same address, a clear example is the case of 'Campillo 4', which accounts for 395 sensors with 6 different types of demand patterns associated.

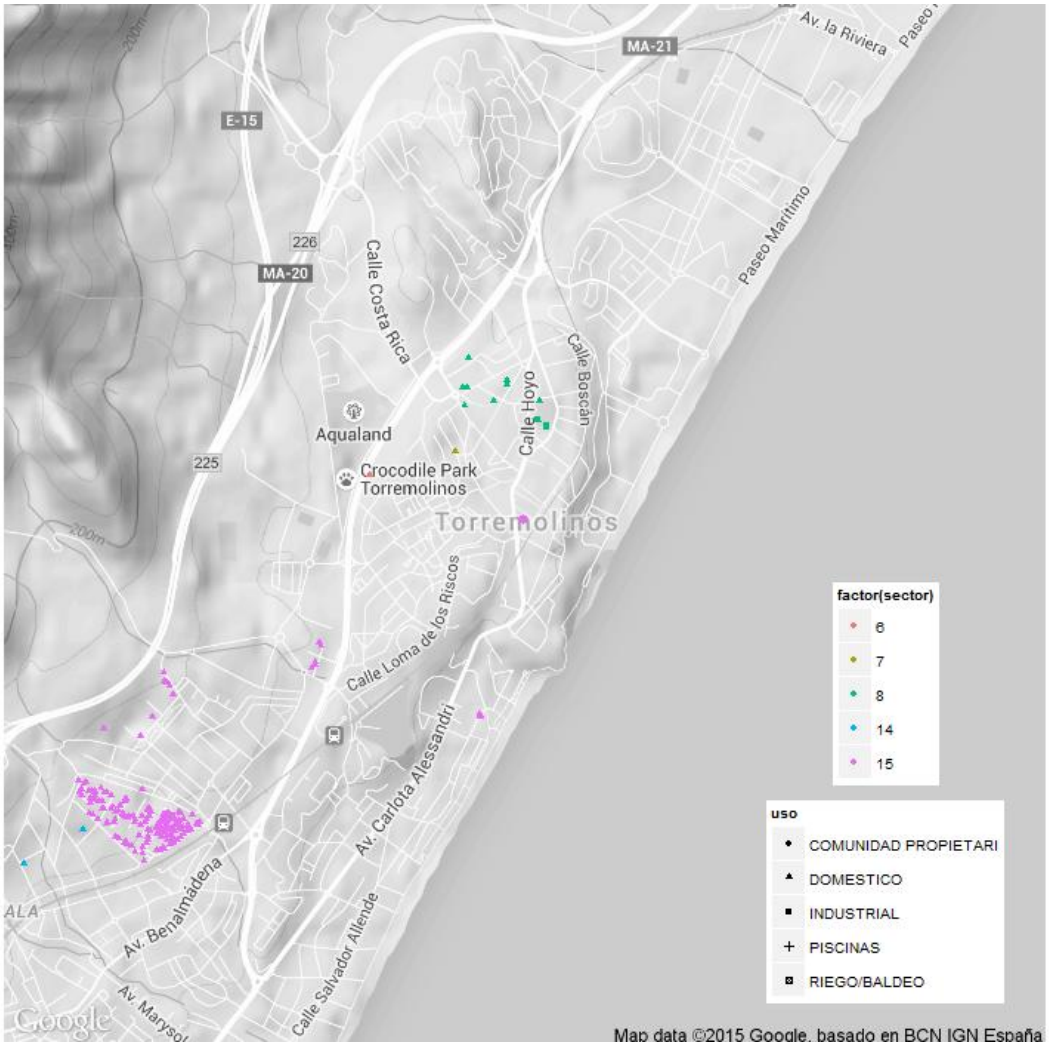


Fig. C.1. Map of torremolinos. Location of all available meter accounts

Knowing that the best ally to work with Big Data, is to allow time to collect enough data, it would be preferred to apply these geographic variables once more sensors taking data become available. This can be considered a small overreach given the significant investment it represents to the water companies dedicated to this.

Shown below is the geographic plot including the addresses of the meters considered for the

regression and classification study in this document, which represents 704 meters. The locations appear sparse with the exception of a small section in sector 15. Having all this geographic data available, it would be useful to have the household indicators, and attempt to match these by using this same plot or the techniques discussed in this document. This information is gathered by the Statistics and Cartography Institute of Andalucía.

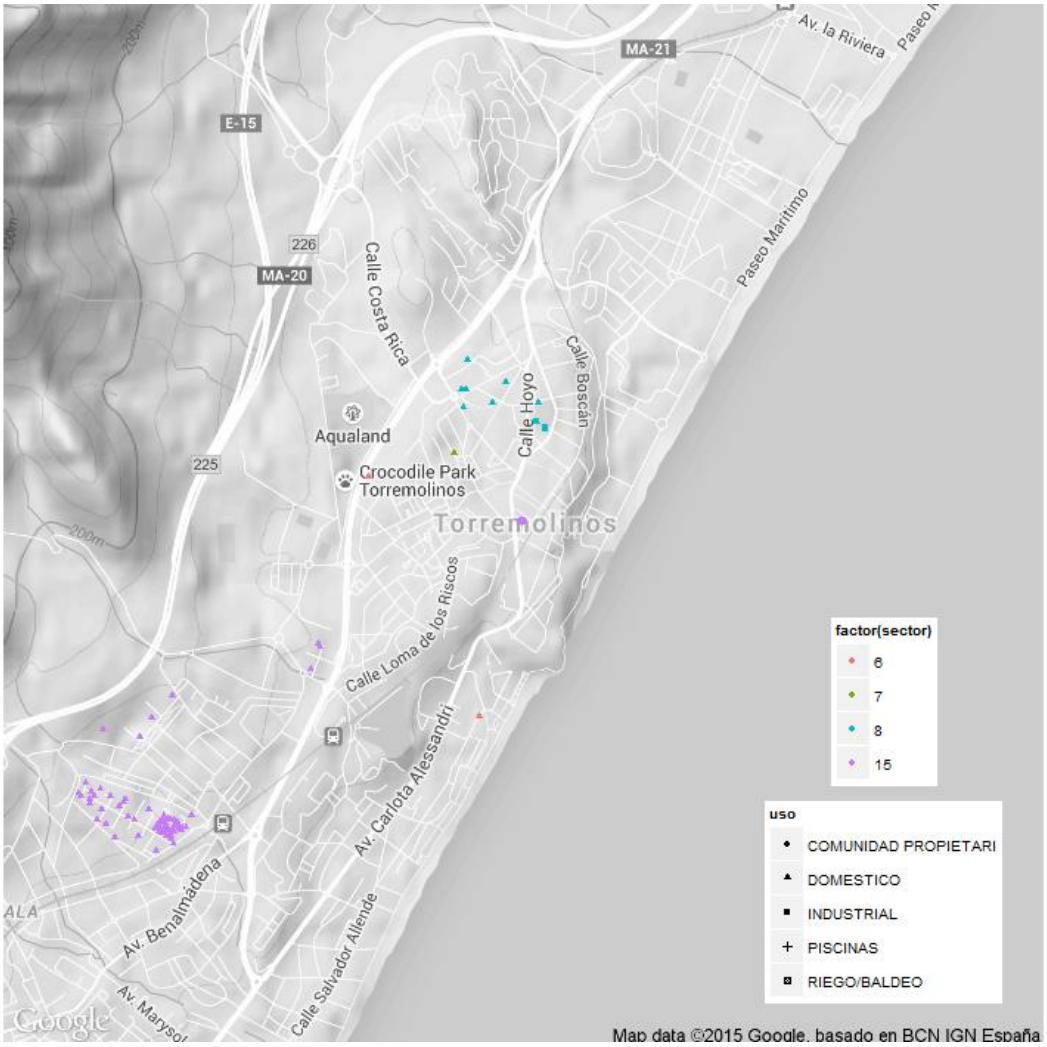


Fig. C.1. Map of torremolinos. Location of used meter accounts

Acknowledgements

First of all, I want to thank my advisor Dr. Vicenç Puig, for your undivided support and disposition since our first meeting back in 2012, when I was still figuring out what my life would be like after college. Thanks for welcoming me to the project, all your assistance, advice, availability, and for facilitating my incorporation to CETAQUA.

This being said, I want to thank Dr. Luis Eduardo Garza Castañón, for all your support during my college years, your useful advice, and for always having faith in me, even when I found it hard to have faith in myself. Most importantly I want to thank you for contacting me with Dr. Puig and motivating me to continue my education.

I also would like to thank Diego García, for your endless patience and support during my brief stay at the company, it was a pleasure working with you. On the same note, I want to thank Daniel González, Rodulfo Rodríguez, and everyone else in CETAQUA for allowing me to work in such a pleasant environment, surrounded by knowledgeable people with amazing projects and ideas.

I want to express my deepest gratitude to my parents and sisters, who have always supported every decision I make, (even when it entails being separated by the Atlantic Ocean), have never limited my desire to improve myself, and always push me to give it all.

Last but not least, I would like to thank CONACYT, for giving Mexicans such as myself the amazing opportunity to further our studies. None of this would have been possible without your support.

Bibliography

[1] GARCÍA, D., GONZALEZ, D. *Water demand estimation and outlier detection from smart meter data using classification and Big Data methods.* Barcelona: Cornellà de Llobregat, 2015.

[2] JOHNSON, J. E., *Big Data plus Big Analytics equal Big Opportunity.* 2012.

[3] MAMADE, A. Z., *Profiling consumption patterns using extensive measurements: A spatial and temporal forecasting approach for water distribution systems.* 2013.

[4] LOUREIRO, D., AMADO, C., MARTINS, A., COELHO, S. T. AND MAMADE, A. (2013) *Outlier detection in water distribution systems – a simple approach (in preparation)*

[5] MCAFEE, A., BRYNJOLFSSON, E., *Big Data: The management revolution.* 2011.

[6] MCKENNA, S. A., FUSCO, F., *Water demand pattern classification from smart meter data.* 2014.

[7] O'LEARY, D. E., UNIVERSITY OF SOUTHERN CALIFORNIA. *Artificial intelligence and Big Data.* 2013.

[8] PÉREZ, R., PUIG, V. *Methodology for leakage isolation using pressure sensitivity analysis in water distribution networks.* 2011.

[9] SCOLNICOV, H., HOROWITZ, G. *Water network monitoring: a new approach to managing and sustaining water distribution infrastructure,* 2010.

[10] SOLANAS, J. L., CUSSÓ, M.R., *Multivariate consumption profiling (MCP) for intelligent meter systems: a methodology to define categories and levels.* 2010.

[11] SOLANAS, J. L., CUSSÓ, M.R., *MCP methodology for intelligent water metering (IWM): assessment of low flow consumption.* 2012.

[12] WILLIAMS, G. *Data Mining, Desktop Survival Guide: Complexity (CP).* 2010. [http://datamining.togaware.com/survivor/Complexity_cp.html, April 2015]

Complementary Bibliography

[1] BISHOP, CHRISTOPHER M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006. Print.

[2] BUSSETI, E., OSBAND, I., AND WONG, S. *Deep Learning for time series modeling.* 2012.

[3] CHANDOLA, V., CHEBOLI, D., AND KUMAR, V., *Detecting anomalies in a Time Series Database,* Minneapolis, 2009

[4] CHANG, A., *R for machine learning,* MIT OpenCourseWare, 2012.

[5] CINAR, G., LOZA, C., AND PRINCIPE, J., "International Joint Conference on Neural Networks" *Hierarchical linear dynamical systems: A new model for clustering of time series.* Beijing, 2014

[6] DANG, T. N., WILKINSON, L., *Time explorer: Similarity Search Time Series by Their Signatures.* Chicago, 2013.

[7] DASZYKOWSKI, M., KACZMAREK, K., VANDER HEYDEN, Y. AND WALCZAK, B. (2007) *Robust statistics in data analysis — A review: Basic concepts*, Chemometrics and Intelligent Laboratory Systems, 85(2), 203-219.

[8] GARCÍA, T., WANG, T., *Analysis of Big Data Technologies and Methods.* Northridge California, 2013

[9] HASTIE, T., TIBSHIRANI, R., AND J. H. FRIEDMAN. "Boosting and Additive Trees." *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer, 2009. 359-70. Print.

[10] HO, R., *Big Data Machine Learning.* Cary NC, 2012

[11] MÜLLER, M. "Dynamic Time Warping." *Information Retrieval for Music and Motion*. New York: Springer, 2007. 69-84. Print.

[12] RADOVANOVIC, M., NANOPOULOS, A., *Hubness in the context of feature selection and Generation.* 2010.

[13] RADOVANOVIC, M., NANOPOULOS, A., *Hubs in space: popular nearest neighbors*

in high dimensional data, 2010.

[14] ZIVOT, E., *Working with financial time-series data in R.* Washington, 2014.

[15] RYAN, J., ULRICH, J., *Package 'xts'.* 2014 [http://r-forge.r-project.org/projects/xts/, March 2015].