Department of Information Science and Technology

Flow time series clustering for demand pattern recognition in drinking water distribution systems: new insights about the most adequate methods

Pedro André Fonseca Garez Gomes

A Dissertation presented in partial fulfillment of the Requirements for the Degree of
Master in Telecommunications and Computer Engineering

Supervisor:
Prof. Dr. Ana Maria Carvalho de Almeida, Assistant Professor
ISCTE-IUL

Supervisor:
Prof. Dr. Dália Susana Santos Cruz Loureiro, Assistant Researcher
LNEC

December, 2019

Acknowledgements

I want to express my deep thanks to all those who directly or indirectly contributed to my success.

To the Professors Ana de Almeida and Dália Loureiro, supervisors of this dissertation, thank you for believing in me and for the support and encouragement shown throughout this period. Without their knowledge, dedication, sympathy and availability, the elaboration of this dissertation would not have been possible.

To my companion Ana for all the support she has given me during these years in which family life was sacrificed for the time devoted to this master's degree.

To my colleague and great friend João Brito, thank you for the moments spent throughout the course. I also appreciate all the mutual help and companionship, fundamental in the preparation of this dissertation.

To ISCTE for accepting students with different engineering backgrounds in the Master's degree in Telecommunications and Computer Engineering.

To the Hydraulics and Environment Department of LNEC, I appreciate the openness and availability demonstrated in the collaboration and development of this dissertation.

To all my family, especially my parents, Ana's parents and my brothers, thank you for all the sacrifices, for always believing in me, for the interest shown throughout my academic career and for always being present, in good times and bad.

To my friends for the support and affection shown during this period.

Resumo

Este estudo apresenta uma proposta de metodologias de *clustering* para reconhecimento de padrões de consumo usando um conjunto de dados de caudal coletados em redes de distribuição de água em Portugal. A maioria dos estudos existentes sobre *clustering* em séries temporais de caudal baseia-se em algoritmos de *clustering* hierárquicos ou de k-Means com medidas de distâncias inelásticas. Este estudo explora alternativas de algoritmos de *clustering*, medidas de distância, janelas temporais de comparação, medidas de índice interno e protótipos de *clustering*.

O desempenho das metodologias de *clustering* foi avaliado em termos de medidas de índice interno e também através da caracterização dos centroides dos clusters. As metodologias com melhor desempenho foram o Algoritmo de Partição com distância DTW, protótipo PAM e janela de temporal de 15 minutos e o Algoritmo de Partição com distância GAK, protótipo PAM e janela de temporal de 15 minutos, pois permitiram a formação três *clusters*. O primeiro método identifica um padrão de consumo noturno, um padrão típico de fim-de-semana e um padrão típico de dia útil, enquanto o segundo método destaca-se por apresentar um padrão com pequena variabilidade entre o consumo noturno e diurno.

Para melhorar a extração de conhecimento, operações adicionais de *clustering* foram realizadas ao conjunto de dados que pertence ao cluster com pequena variabilidade entre consumo noturno e diurno. Novos clusters foram identificados e caracterizados, mostrando que os padrões associados à irrigação são independentes do período do dia e da época do ano, o que indica um uso ineficiente da água.

Palavras-Chave: Aprendizagem não supervisionada; *Clustering* de series temporais; Séries temporais de caudal; Reconhecimento de padrões de consumo; Sistemas de distribuição de água.

Abstract

This study presents a proposal of clustering methodologies for demand pattern recognition using network flow data collected from a large set of drinking water distribution networks in Portugal. Most of the existing studies about clustering in flow time series rely on hierarchical or k-Means clustering algorithms with inelastic measures distances. This study explores alternative clustering algorithms, distance measures, comparison time windows, internal index metrics and clustering prototypes. The performance of the alternative clustering methodology was assessed in terms of multiple internal index metrics and the characterization of the cluster centroids.

The methods with the best performance were Partition Algorithm with DTW distance, PAM prototype with 15 minutes time window and the Partition Algorithm with GAK distance, PAM prototype and 15 minutes time window because they allow a clear partition of flow time series in three clusters. The first method identifies a night consumption pattern, a typical weekend pattern and a typical working day pattern, whereas the second one identifies a pattern with small variability between night and daily consumption.

To improve knowledge extraction, in terms of typical and anomalous existing patterns, additional clustering operations were performed with the flow data set that belongs to the cluster with small variability between night and daily consumption. New clusters were identified and characterized regarding weekday, geographical location, and dry months and wet months, showing that patterns associated with garden irrigation are independent of the period of the day and season of the year, which indicates an inefficient water use.

Keywords: Unsupervised learning; Time series clustering; Flow time series; Demand pattern recognition; Water distribution systems.

Table of Contents

List of Tables

List of Figures

List of Acronyms

Acronym	Term
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
CART	Classification And Regression Trees
DBSCAN	Density-based spatial clustering of applications with noise
DICTCRVB	Dissimilarity Index Combining Temporal Correlation and Raw Values Behaviours
DBA	Dynamic Time Warping Barycenter Averaging
DMA	District Metered Area
DTW	Dynamic Time Warping
ERSAR	Portuguese Water and Waste Services Regulation Authority
FFT	Fast Fourier Transform
GAK	Global Alignment kernels
GAM	Generalized Additive Models
HMM	Hidden Markov Models
ISCTE-IUL	ISCTE - University Institute of Lisbon
K	Kwon index
KNN	K-nearest Neighbors
LCSS	Longest Common Subsequence
LCM	Loscal Cost Matrix
LNEC	Portuguese National Laboratory for Civil Engineering
MVM	Minimum Variance Matching
MPC	Modified Partition Coefficient index
NCC_c	Cross-correlation with Coefficients Normalization
PAM	Partition Around Medoids
PBMF	PBMF index
PCA	Principal Component Analysis
PDB	Priodogram Based Dissimilarity
PNUE	Portuguese National Program for the Efficient Use of Water
SAX	Symbolic Aggregate ApproXimation
SC	Validity Function
SBD	Shape-based Distance Measurement
STING	Statistical Information Grid-based Algorithm
SVD	Singular Value Decomposition
T	Improved Validation index

Acronym	**Term**
TBATS	Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality
TGAK	Triangular Global Alignment Kernel
TLK	Tringular Local Kernel

Chapter 1

Introduction

1.1 Overview

The present chapter is organized as follows:

- **1.2 Motivation and framework**: introduces the context of the domain of water supply systems and water use efficiency;
- **1.3 Research questions**: presents the research questions;
- **1.4 Objectives**: describes the dissertation objectives;
- **1.5 Thesis outline**: describes the structure of the dissertation.

1.2 Motivation and framework

Water supply systems are infrastructures of great importance for the proper functioning of any urban agglomerate. Currently, water supply systems present problems related to water losses. These losses are associated with aging infrastructures, implying: system failures, high pumping energy costs and network rehabilitation needs. These factors, coupled with budgetary constraints, result in the need to improve the processes of loss control in water supply networks (Candelieri et al. 2014).

The most efficient loss control processes allow less need for extraction of water for human consumption, generate financial savings by delaying the need for investment in the construction of new water supply infrastructures and reduce the energy needs of treatment processes and systems associated with water supply systems (Loureiro et al. 2016a).

According to the Portuguese Water and Waste Services Regulation Authority (ERSAR), Portugal presents on average 35% of losses (apparent and real) of the total water produced. This percentage includes unbilled consumption, apparent losses and actual losses. Being that 24% correspond to actual losses and 11% to apparent losses and unbilled consumption (ERSAR 2013).

Globally more than 48 billion cubic meters per year of treated water are identified as losses, of which 66% are due to actual losses of water distribution systems, representing a significant

economic impact: over 14 billion dollars per year are lost by the managing entities (Kingdom et al. 2006).

Water losses in water supply systems are comprised of the following components (Alegre et al. 2006; Loureiro et al. 2016a; Sela et al. 2015):

1. **Apparent loss due to unauthorized consumption**: consumption through unauthorized connections to the water distribution system;
2. **Apparent loss by unmeasured consumption**: municipal fountains and irrigation systems without flow metering systems for billing control;
3. **Apparent loss due to measurement errors or communication errors**: errors in measuring devices or errors in communication with telemetry or remote management centers;
4. **Actual losses**: leaks or ruptures of conduits and leakage or overflow in cells of the reservoirs represent physical losses of water resources.

The typical method of flow analysis for loss control is characterized by sectioning a water distribution network in small sections, consisting of about 500 to 3000 household connections. These zones are referred to as District Metered Areas (DMAs). At the boundaries that separate the zones, flowmeters are installed that measure what goes in and what leaves each section of the distribution network. Figure 1.1 typifies a layout of DMAs:

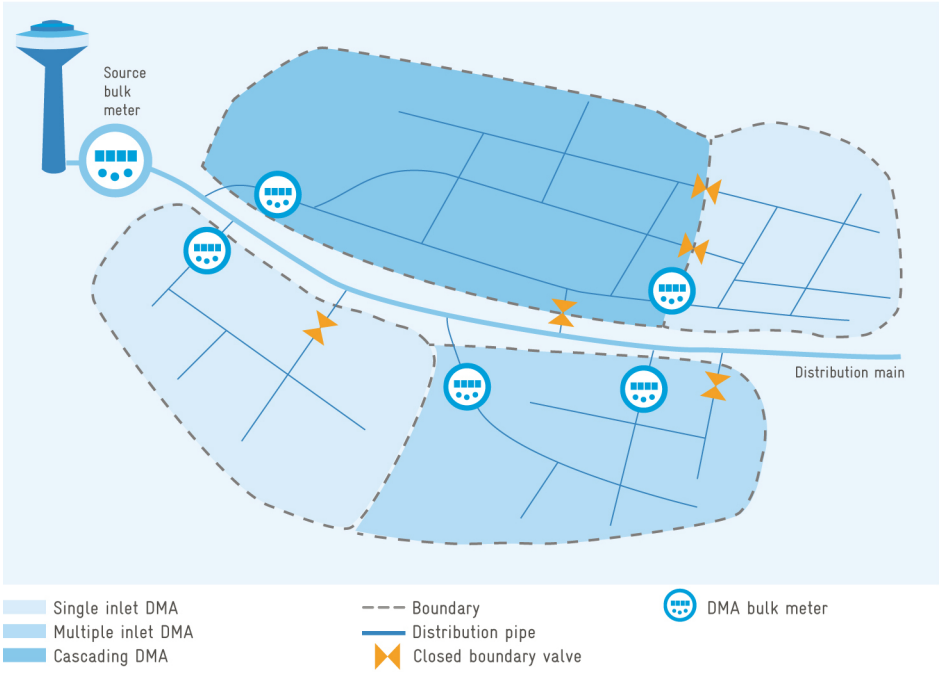

Figure 1.1: DMAs layout (Farley (2001) and Strategic Alliance for Water Loss Reduction (2017)).

Through the flowmeters installed in the DMAs it is possible to obtain Time Series of Average Flow. The data from each of the DMAs is transmitted to the remote control systems, allowing a reactive action in an event of abrupt consumption (e.g., rupture of a conduit) and a more

precise identification of the zone where this event may be occurring. However, the analysis of a history of these flow time series of flow can allow for preventive actions (e.g., identification of leakage in its initial stages), consuption beahavior analysis (e.g.,using statistical analysis and unsupervised learning clustering techniques) and predictive action (e.g., using supervised learning techniques for classifying real-time events).

From the point of view of the preventive and predictive actions at the level of the operation of the systems, it is essential the identification and evaluation of outliers in historical data of time series of medium flow (Loureiro et al. 2016a).

Outliers in water distribution systems can be a consequence of:

1. Periods of abnormal consumption (domestic or non-domestic);
2. Changing in the operation of the system;
3. System failures: rupture of conduits or in fittings of connections;
4. Problems with instrumentation, communication and data storage.

Outliers for instrumentation, communication or storage read errors may reflect an arbitrary change in the flow time series. Changes in the operation of the system by opening valves or starting pumps, can also lead to the appearance of outliers (Loureiro et al. 2016a).

In the case of leaks and ruptures of conduits, this type of events causes increases in measured flow (contributing to actual losses and service interruptions). In these cases, the analysis of the time series of the nocturnal period between midnight and five in the morning becomes essential to evaluate the real losses. During this periods since the water consumption is low and consequently the flow associated with leakage tends to be a significant part of the consumption identified (Loureiro et al. 2016a).

Outliers related to periods of abnormal consumption may reflect a significant change in consumption habits due to fluctuating population and changes in water use (Loureiro et al. 2016a).

The use of DMAs has become common practice for water utilities. Continuous collection of flow data in DMAs generates large volumes of historical data. Typically, this data is only used for online operation and control of water distribution systems. Historical data is discarded over time or stored in aggregate form (e.g., dailly mean value). Preservation of historical data by water utilities to extract knowledge about consumption behaviors has yet to be seen as an important practice (Loureiro et al. 2016a). In this case, we intend to show the added value of employing clustering techniques over the historical data of DMAs for knowledge extraction and decision support.

According to the Portuguese National Program for the Efficient Use of Water (PNUE), Portugal started the 21^{st} century with an annual demand for water in the continental territory estimated at about 7500 million m^3 in all three sectors: urban, agricultural and industrial. The agricultural sector is, in volume terms, the largest consumer ($> 80\%$). In terms of supply costs, the urban sector is the most significant, since water for human consumption requires treatment. Total water demand declined significantly between 2000 and 2009 (around 43%). Several factors contributed to this reduction. Several water supply management entities

(urban sector) have made a considerable effort to reduce losses in the transportation and distribution systems (APA 2012).

The most significant reduction in consumption was in the agricultural sector, the largest water consumer. This reduction was due to a combination of factors related, on one hand, to the national situation, which led to a reduction of irrigated areas in the first decade of the century, mainly in the north and center of the country and, on the other hand, water use efficiency related to management of losses associated with the storage, transportation and distribution systems and also to the application of more efficient water irrigation systems in parcels. The drought that occurred between 2004 and 2006 also contributed to a temporary reduction of irrigated areas (APA 2012).

The application of some measures in the various sectors provided for the improvement of water use efficiency. The inefficiency associated with losses in the adduction and distribution system was more significant in the urban sector (APA 2012). Figure 1.2 indicates the reduction of inefficiency between 2000 and 2009 for the various sectors.

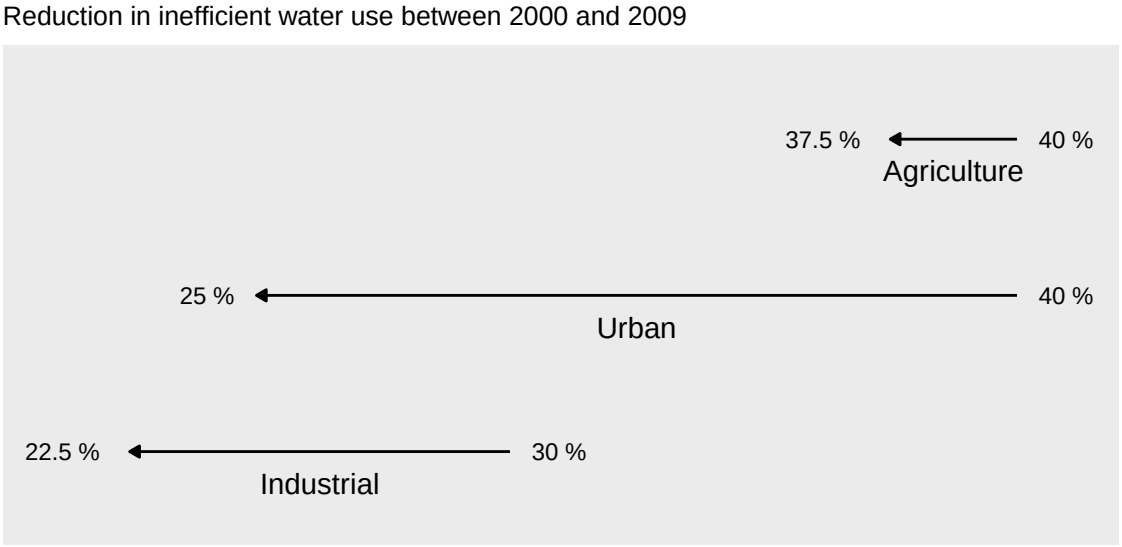

Figure 1.2: Variation of inefficiency in water use between 2000 and 2009 in Portugal by sector, based on APA (2012).

The inefficiency of water use is especially burdensome in periods of water scarcity. Portugal has already experienced several periods of drought, the most recent one being in 2004/2005. In addition to the social dimension inherent to the drought experienced by the directly affected populations and productive sectors, a drought can have a strong economic impact (APA 2012). Figure 1.3 shows the economic impact of 2005 drought by sector.

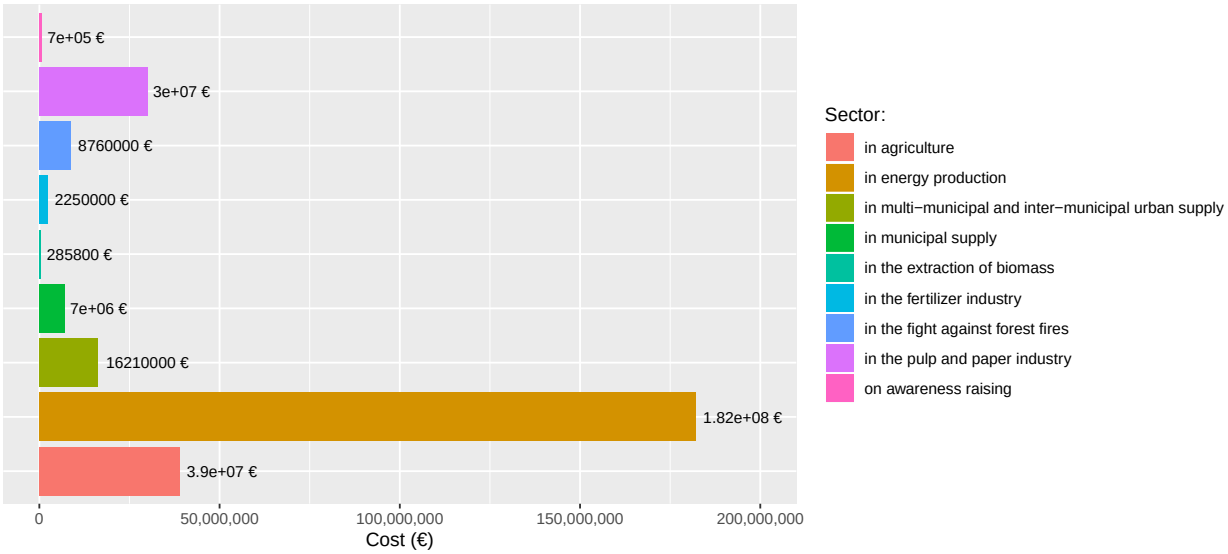

Figure 1.3: Sectoral costs of the 2005 drought, based on APA (2012).

Taking into account the efforts that have been made in Portugal, there is still an important component of wastage associated with losses and the inefficient use of water for the intended purposes. The inefficiency of water use leads to high environmental, social and economic damages. These impacts are more severe in areas that are more sensitive to dry periods. It is therefore important to characterize consumption behaviors throughout the regions of Portugal in order to be able to develop specific actions taking into account the specificity of each region.

1.3 Research questions

The initial problem identified was finding a process to organize of the time series of average daily flow, wiched allowed to characterize them in several representative groups.

Starting from this problem, this work raises a set of questions like:

1. What are the best clustering approaches to apply to daily time series of mean flow?
2. What are the appropriate distance measures for characterization of similarity in daily time series of mean flow?
3. What are the appropriate prototypes to be used in clustering methods for daily time series of mean flow?
4. How to evaluate the performance of the clustering algorithms and the proposed solutions?
5. Will it be possible to detect the occurrence of anomalous behaviors (e.g., unsustainable consumption habits, ruptures or water leaks), based on historical data and through the approaches of unsupervised learning proposed in this dissertation?

1.4 Objectives

Flow measurement in urban water supply systems is fundamental for improving knowledge about domestic and non-domestic (public, commercial and industrial) urban consumption components and water losses. The clustering of time series allows to classify zones with similar behavior in terms of consumption (e.g., seasonality), and to identify zones whose consumption behavior can be considered inefficient (e.g., ruptures, anomalous consumption), whose study is important for loss control and efficient management of water distribution systems.

The present dissertation intends to contribute to the study of the temporal series of mean flow timeseries focusing on the following objectives:

1. Organization of daily time series of mean flow in similar groups, through approaches of unsupervised learning, to understand the behaviors and patterns present in the dataset;
2. Comparison of results obtained through the various approaches of unsupervised learning, taking into account the empirical knowledge about the data domain and the accuracy of the solutions obtained;
3. Characterization of groups formed with the following parameters:
   - Working days vs. weekend / holiday;
   - Geographical region;
   - Dry months vs. months.
4. Identification of groups with anomalous water consumption behaviors that may lead to inefficient water use.

1.5 Thesis outline

This dissertation is divided into 5 chapters.

- **Chapter 1**: presents a context of the domain of water supply systems and water use efficiency, the research questions, the objectives and the structure of the dissertation;
- **Chapter 2**: review of the state of the art concerning unsupervised learning techniques applied to time series and also review of research work in the domain of water demand management with clustering techniques;
- **Chapter 3**: methodology and theoretical description of the techniques used in preprocessing, clustering operations, cluster evaluation and cluster visualization;
- **Chapter 4**: the techniques and methodology described in Chapter 3 will be applied to the dataset to characterize it and extract knowledge about the various identifiable behaviors in daily mean flow patterns.
- **Chapter 5**: dedicated to the conclusions on the analyzes carried out and the proposed future works.

Chapter 2

State-of-the-art

2.1 Overview

This chapter will present a general review on unsupervised learning applied to time series and also a review on research work in the field of water demand management that have applied unsupervised learning techniques.

The present chapter is organized as follows:

- **2.2 Unsupervised learning of time series**: an overview of time series clustering techniques;
- **2.3 Unsupervised learning in water demand management domain**: a review of research work in water management that used unsupervised learning techniques.

2.2 Unsupervised learning of time series

One of the branches of Machine Learning is Unsupervised Learning. This branch encompasses a set of techniques that group data homogeneously, without prior knowledge of the definition of groups (Aghabozorgi et al. 2015; Rai and Singh 2010; Warren Liao 2005). This set of techniques is useful in the exploratory analysis of data because they identify structures in non-categorized data sets by organizing similar data into groups. The groups are formed taking into account the maximization of similarity of the objects belonging to the same group and minimization of similarity between objects belonging to different groups (Aghabozorgi et al. 2015).

The clustering of time series data is a particular case for the application of unsupervised learning techniques, since the time series present dynamic characteristics (the value of the characteristics analyzed is time dependent). As such, each point in the time series is a chronological observation. This typology is typically composed of a large number of data of various dimensions (Aghabozorgi et al. 2015; Keogh and Kasetty 2002; Lin et al. 2004; Rani and Sikka 2012).

The use of this type of techniques allows the detection of patterns in the time series data and

the reduction of dimensionality through the grouping in clusters. The subsequent graphical visualization of the behavior of each of the groups allows the user a better understanding of the data structure, anomalies and other irregularities. On the other hand, this technique can be used as a subroutine for more complex machine learning algorithms such as rule discovery, indexing, classification and detection of anomalies (supervised learning) (Aghabozorgi et al. 2015).

2.2.1 Time series clustering approaches

Time series clustering can be performed according to one of the following approaches (Aghabozorgi et al. 2015):

1. Adaptation of conventional non-supervised learning algorithms (applied to static data) in order to be compatible with the dynamic nature of the time series, typically by changing the distance measure;
2. Transformation of the time series into static objects and later application of conventional algorithms of unsupervised learning;
3. Multi-phase approaches that use different temporal resolutions as input.

In addition to the described approaches, the methods for grouping time-series data can be classified as (Aghabozorgi et al. 2015; Warren Liao 2005):

1. **Form-based approaches:** time series are best adjusted by contraction or extension of the time axis. In this approach, conventional clustering algorithms are applied, but distance measures are adapted to the time series typology;
2. **Characteristic-based approaches:** time series are converted into feature vectors of smaller size. Subsequently, grouping operations are performed;
3. **Model-based approaches:** each time series is transformed into a set of parameters of a model. Each model generated for each series is later grouped taking into account a distance measure and a grouping algorithm. This type of approach usually presents scalability problems and its performance is reduced when the groups formed are close (Mahalakshmi et al. 2016; Vlachos et al. 2004).

Figure 2.1 outlines the different approaches of time series clustering:

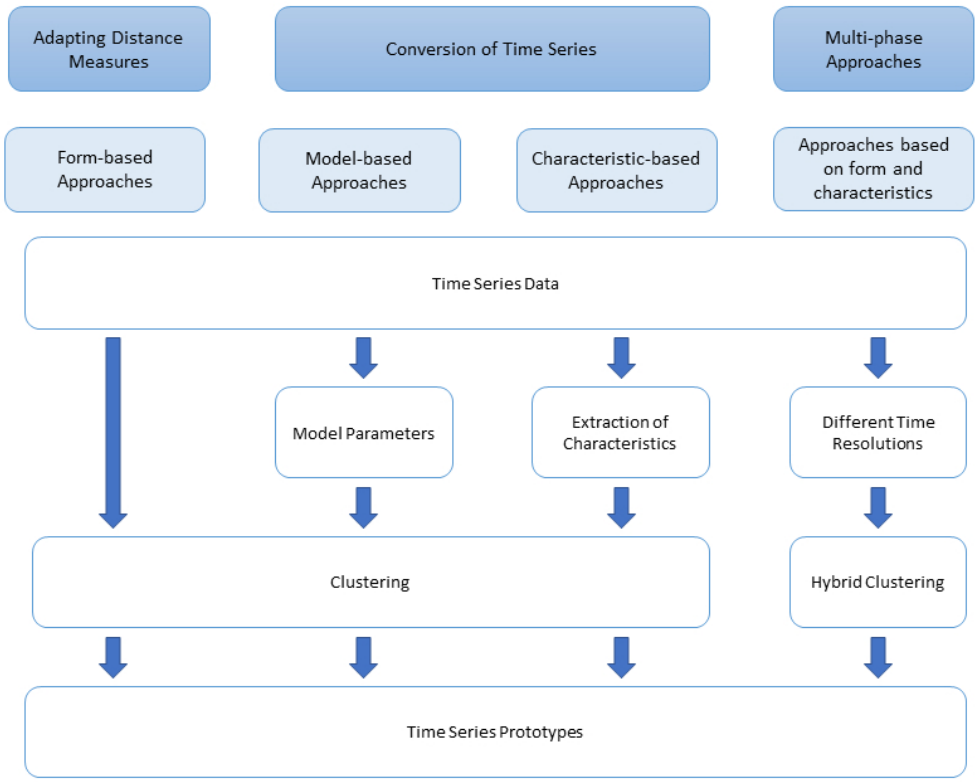

Figure 2.1: Time series clustering approaches, based on Aghabozorgi et al. (2015) and Vlachos et al. (2004).

2.2.2 Components of time series clustering

Time series clustering can be defined according to the following five components:

1. Representation of the time series;
2. Measures of distance or similarity to be applied;
3. Clustering prototypes;
4. Clustering algorithms to be used;
5. Measures of performance evaluation.

Time series representation

The appropriate choice for time representation is an important component as it directly affects the execution efficiency and the end result of the clustering approach used. Given that the time series are typically composed of a large number of data of several dimensions, size reduction methods are usually used in order to improve the performance of the algorithms (Keogh and Kasetty 2002).

These methods of time representation can be grouped into the following typologies (Aghabozorgi et al. 2015):

1. **Adaptive to data:** they are applied to all dataset of time series and try to minimize

the overall reconstruction error using arbitrary length segments. Example of this type of algorithm are the reductions through Singular Value Decomposition (SVD) matrices or with Principal Component Analysis (PCA). This methods can represent better each original series, but the comparison between different time series may be more complex;
2. **Non-adaptive to data:** this typology of representation is suitable for time series of equal length and the comparison between several time series is direct;
3. **Model-based approach:** based on stochastic representations such as Markov Models or Hidden Markov Models (HMM) (Minnen et al. 2006, 2007; Panuccio et al. 2002). In this approach, as in the previous ones, the level of data compression can be adapted according to the type of application;
4. **Data-dependent compression:** in this approach, unlike those presented above, the compression level is automatically set based on the original time series.

Distance measures

Distance measures are an important parameter in the definition of time series clustering methods. Typically the time series present different time spacings which makes the comparison more complex. In temporal series problems there are four typologies of similarity distances between series (Aghabozorgi et al. 2015):

1. **Similarity based on form:** Euclidean distances, Dynamic Time Warping (DTW), Longest Common Subsequence (LCSS), Minimum Variance Matching (MVM) can be applied in these cases. They are suitable for short time series;
2. **Similarity based on compression:** these are distances suitable for both short and long time series. Included in this category are distance measures such as Pearson's correlation coefficient, Cepstrum and Cousine Wavelets;
3. **Similarity based on the characteristics:** the long time series are appropriate in these cases, since the time series have undergone a process of extraction of characteristics allowing to decrease the dimensionality;
4. **Similarity based on models:** these cases are also appropriate for long time series. In this group, models are created using HMM and Autoregressive-moving-average (ARMA).

Clustering prototypes

Proper choice of a representative of a group is an essential subroutine of some time-series clustering algorithms. In algorithms such as k-Means, k-Medoids, Fuzzy C-Means, the choice of prototypes has direct implications on the quality of the formed groups (Aghabozorgi et al. 2015; Bagnall and Janacek 2005; Chu et al. 2013; Corradini 2001; Keogh and Pazzani 1998; Rabiner et al. 1979).

Generally there are three approaches to defining prototypes:

1. **Use of medoid as prototype:** in this method the distance of all pairs of time series belonging to the same group, using distances such as Euclidean, DTW or LCSS, is calculated. Subsequently, the time series that has the smallest sum of the quadratic error is chosen as the prototype of the group (Vuori and Laaksonen 2002);

2. **Use of an average prototype:** applies to time series of equal size in which a measure of rigid distance was used in the grouping process (e.g., Euclidean distance). The process of choosing the prototype is performed by calculating the average or the median of all time series at each point. However, in case the time series have different lengths, it is not possible to apply this method directly, and in a first phase the average prototypes of the series with the same time length must be calculated, and later use these prototypes to calculate a prototype medoid through distance measures such as the DTW or LCSS (Banerjee and Ghosh 2001; Sakoe and Chiba 1971, 1978; Vlachos et al. 2002);
3. **Use of local search prototype:** this method calculates as a first approach a medoid to represent the group. Later, using a method of calculation of average prototype, through the warpping path of the previous calculated distance matrix, it is calculated a new prototype. Finally, to this new prototype it is applied again the process of calculation of new warping paths. The type of approach to obtain prototypes by local search is used instead of medoids to overcome the poor quality in time-series clustering in Euclidean space (Hautamaki et al. 2009).

Time series clustering algorithms

Generally the time series clustering algorithms can be classified into six groups: partition clustering, hierarchical clustering, density-based clustering, model-based clustering, grid-based clustering and hybrid clustering (Aghabozorgi et al. 2015; Warren Liao 2005).

Partition clustering The grouping through these methods resorts to k pre-defined or random groups and to n elements to be categorized, so that each group contains at least one element (Aghabozorgi et al. 2015). One of the algorithms belonging to this family is k-Means, where each group contains a representative prototype of the group that was constructed based on the average value of all the objects belonging to that group (Macqueen 1967). Another possible approach is the k-Medoids algorithm, where the prototype of each group is the closest element to the cluster center (Gentle et al. 2006). These approaches are more efficient when compared with hierarchical algorithms (Bradley et al. 1998; Macqueen 1967).

K-Means and k-Medoids are algorithms whose clusters are constructed so that an element can only belong to a cluster, known as a strict clustering rule. Other approaches such as FCM (Fuzzy c-Means) or Fuzzy c-Medoids, allow an element to have a degree of belonging to each cluster (Bezdek 1981; C. Dunn 1973; Krishnapuram et al. 2001; Warren Liao 2005).

These algorithms require the definition or delivery of initial prototypes, which implies that the precision of the algorithms depends directly on the definition of these prototypes and the methods of updating them. This family of algorithms has better performances when the time series are of equal size, because it is not clear how to define the centroid of the cluster when they are of different size (Aghabozorgi et al. 2015; Warren Liao 2005).

Hierarchical clustering This algorithm typology creates a hierarchy of groups that can be additive or divisive. The additives initially consider each cluster item as a group

and gradually agglomerate groups according to the proximity of the measure similarity (bottom-up approach). The divisive algorithms begin with all items in a single group, being later and sequentially divided into smaller groups, until each group consists of only one item (top-down approach) (Gentle et al. 2006).

The similarity in hierarchical groupings of time series is evaluated based on the generation of a time series distance matrix (Vlachos et al. 2003).

This algorithm typology has the advantage of visualizing the structure of the data, namely the visualization of dendrogram. Another aspect that is important to note is that these algorithms do not require the definition of the number of groups to be created as the initial parameter. In the case of time series this characteristic is very important due to the difficulty in defining which number of initial clusters to use (Aghabozorgi et al. 2015).

It is also important to note that this type of algorithm allows the grouping of time series with different lengths through distance measurements such as DTW or LCSS (Banerjee and Ghosh 2001; Sakoe and Chiba 1971, 1978; Vlachos et al. 2002).

The rigidity imposed by the divisive hierarchical algorithms does not allow adjustment after the separation of groups, which implies that there is no reversibility of the process after an element is associated with a lower level group. In the case of the additives, the same happens when the union of two groups occurs. This aspect may be detrimental to the quality of the groups formed (Aghabozorgi et al. 2015). In addition, these algorithms are not able to deal effectively with long time series because they are of quadratic computational complexity. Consequently, the use of this algorithm typology is suitable for small data sets because of its poor scalability (Wang et al. 2006).

Density-based clustering In this category of algorithms the groups are dense subspaces composed of objects and are separated by subspaces where the density of objects is low. One of the most used algorithms is the Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al. 1996). The application of this algorithm to temporal series is not common due to its high degree of computational complexity (Aghabozorgi et al. 2015).

Model-based clustering In these algorithms a model is assumed for each cluster and the best fit of data is set to each one. In general, this typology needs a set of parameters and is based on user premises, which may be false and, consequently, result in inappropriate groups. On the other hand, they have slow processing times in large datasets (Aghabozorgi et al. 2015; Andreopoulos et al. 2009; Warren Liao 2005).

Grid-based clustering In these algorithms the space of observations is represented by a grid with a finite number of cells. Subsequently the objects are grouped based on the existing cells. Statistical Information Grid-based Algorithm (STING) and Wave Cluster algorithms are examples of this typology (Sheikholeslami et al. 1998; Wang et al. 1997).

It is not common to apply this typology to time series, since it implies a brute-force approach, which in the case of real problems with large data groups generates problems of efficiency and precision (Aghabozorgi et al. 2015).

Hybrid clustering These approaches are typically combinations of algorithms from other families, where grouping operations are performed by levels. Taking advantage of at each level the groups formed in the previous level to apply a new algorithm and thus to produce new groups (Aghabozorgi et al. 2015).

Measures of performance evaluation

The evaluation measures are an important topic in the characterization of the quality of the groups produced by unsupervised learning. Typically, the data set does not have a category that classifies them, so visualization of groups or group prototypes is an important method for assessing the appropriateness of the method used to the nature of the data, or even whether the method needs parameters calibration.

In addition to the visualization, scalar measures can be applied for the evaluation of the precision of the clustering operation. These measures can be classified into two categories:

1. **Internal index:** these measures allow to evaluate the quality of adjustment of the groups formed to the data. One of the means applied in this category is the sum of the quadratic errors related to the measure of similarity used (Han et al. 2011). These measures should only be used for the comparison between different grouping approaches that were generated using the same model and metrics (Aghabozorgi et al. 2015).
2. **External index:** these measures assume that the data are categorized (ground truth) and, in a generic way, evaluate the accuracy with which the formed groups represent the true categories, verifying the percentage of True positive (correct) and False positive (wrong category) in each group. Purity, F-measure, Entropy and Jaccard are examples of these type of measures, they are typically used in supervised learning analysis (Aghabozorgi et al. 2015).

Figure 2.2 outlines the different performance evaluation approaches for time series clustering:

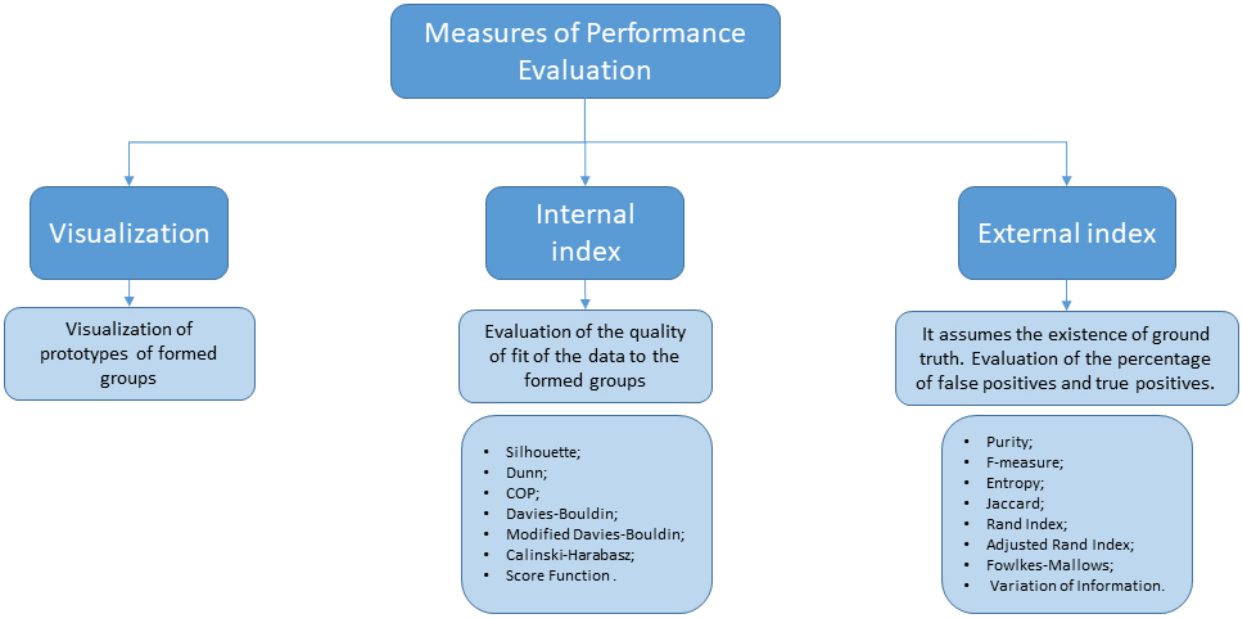

Figure 2.2: Time series performance evaluation approaches, based on Aghabozorgi et al. (2015); Sarda-Espinosa (2019) and Han et al. (2011).

2.3 Unsupervised learning in water demand management domain

This section provides an overview of water demand management studies that use unsupervised learning techniques in their methodology as an important step in resolving a problem in this domain. In order to manage water supply systems more effectively, studies have been carried out on the following topics:

1. Water demand profiling;
2. Identification of outliers;
3. Disaggregation of consumption taking into account its use;
4. Data reconstruction of flow time series.

2.3.1 Water demand profiling

Characterization models of water that demand profiling are quite important in the management, operation and planning of water distribution systems. The design of more rigorous models to characterize the water needs of a region throughout the day, taking into account the climate, seasonality (weekly, monthly and annual), water uses and socio-demographic characteristics. These models provide valuable insight that will allow water utilities to more effectively manage water distribution systems.

A multiple linear regression model for water profiling was proposed by Loureiro (2010) and later improved by Mamade (2013). This model takes into account socio-demographic

characteristics that allows improving the understanding about spatial demand distribution within the water distribution network, which is fundamental to reduce the uncertainty in network operation and planning and to identify clients with a large potential to improve water efficient use. One of the steps in this methodology is to apply a hierarchical clustering with a Euclidean distance measure, taking as parameters the flow series. Through the dendrogram formed, cutoffs are performed to form clusters. It is considered a mean prototype and typically the clusters formed allow to identify the monthly seasonality (e.g., summer and winter pattern) and weekly seasonality (e.g., weekday and weekends pattern). These formed clusters allow the construction of scenarios that come to define the consumption variables to be used as dependent variables in the construction of multiple linear regression models that will have as independent variables socio-demographic indices, infrastructure characteristics and billing.

Figure 2.3 shows the methodology for the construction of the scenarios and the expected results for the scenarios as an example.

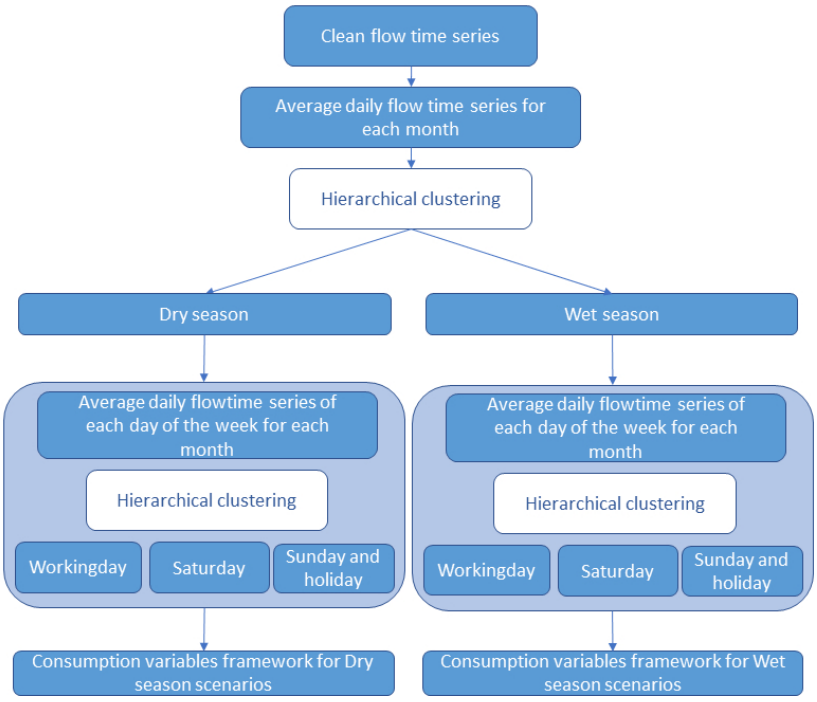

Figure 2.3: Methodology for consuption scenarios, based on Loureiro (2010) and Mamade (2013).

Another study of Loureiro et al. (2015) was developed with the aim of identifying through a Spearman correlation matrix the most important socio-demographic factors that can influence household consumption. The application of this methodology in this study was possible since it was available a dataset with consumptions at the level of the network as well as data of household consumption. In this study the author performed clustering operations using the K-means algorithm with a Euclidean distance on each of the following sets of consuption variables:

1. Average daily consumption at the level of the statistical sections;
2. Average daily consumption for each client;
3. Dimensionless daily average patterns for working days for each client;
4. Dimensionless daily average patterns for weekends for each client.

The clusters formed in each of the clustering operations allowed to create a set of new variables that correspond to the distribution of the clients by each one of the clusters. These new consumption variables were correlated with socio-demographic variables through a Spearman correlation matrix in order to validate the most important socio-demographic factors in household comption.

Another study was conducted by Loureiro et al. (2016b) with the objective of presenting a comprehensive approach for spatial and temporal demand profiling in network areas, focusing on domestic consuption. This study presents, in the first phase, a multiple linear regression classifier (supervised learning) similar to that presented in the Loureiro (2010) and Mamade (2013) studies. This model uses a Euclidean distance clustering to obtain consumption variables to serve as dependent variables of model . In a second phase, this study presents another supervised learning model based on Classification And Regression Trees (CART) algorithm with Gini impurity to classify consumption patterns on working days based on the variables public billed consumption and individuals mobility. In this second model a hierarchical clustering with Euclidean distance is also used to group the daily demand patterns.

Cheifetz et al. (2017) study was conducted with the objective of characterizing the consumption profiles existing in DMAs installed in the largest water distribution network in France. This study consisted of an application of the Fourier-based time series decomposition method to extract seasonal components from time series. Then, two clustering methods were applied to the extracted seasonal components: k-Means with PCA and Fourier regression model. To evaluate the best number of clusters to obtain was used as Bayesian Information Criterion evaluation method. In this study both models formed 8 clusters and through the prototype analysis it was possible to characterize the profiles in the following categories: residential use, commercial use, industrial use and noise cluster.

Cominola et al. (2016) study characterizes the water consumption behavior of 175 households in the municipality of Tegna in Switzerland. The methodology used in this study consisted of categorizing the flow values recorded by 4 categories (no consuption, low consuption, medium consuption and high consuption). Subsequently, a PCA was applied to reduce dimensionality. After dimensional reduction, the first main component was used to apply the clustering method k-Means with formation of 3 clusters. These formed clusters were characterized at prototype level to describe the consumption behaviors present in the dataset.

Another study using the k-Means clustering method was conducted by Mounce et al. (2016) to characterize the consumption behaviors present in a dataset composed of 3428 counters installed in the cities of Reading, Swindon and London (United Kingdom). The series featured a 15-minute time step and a duration of 3 years. The clustering operation performed in this study allowed to obtain 3 clusters that were subsequently correlated with the types of activity present in each cluster, allowing to verify that there were clusters associated

with a residential consumer profile and others with a commercial consumer profile. After this characterization and given that the data were previously categorized as residential or commercial use, k-Neirest Neighbor and Decision Trees classifiers (supervised learning) were trained to classify the flow series as consumption for commercial or residential use.

2.3.2 Identification of outliers

Outliers detection methods allow the identification of anomalous events in flow time series. More specifically, the detection of real losses associated with leakage are important for the management of water distribution networks in a more efficient way, since they allow the identification of more degraded sectors of the network that need a primary intervention.

In this domain, Loureiro (2010) and Mamade (2013) present the "Symmetrical method". This method consists of the algorithm described in Equation (2.1) which uses two robust statistics: median (MED) and Q_n which is a robust standard deviation measure of Q_n observations (Rousseeuw and Croux 1993). In this approach no clustering operation is applied previously.

$$OTL \geq MED + c \times Q_n \quad \vee \quad OTL \geq MED - c \times Q_n \tag{2.1}$$

In which OTL [m^3/h] and MED [m^3/h] are the outlier value in the data series and the median of a set of previous observations defined by the user, respectevly. The variables c [-] and Q_n [m^3/h] are the threshold value to be defined by the user (with $c > 0$) and the robust standard deviation of the observations based on the Q_n scale, respectively.

Given that the time series of flow have seasonal, weekly and in some cases monthly seasonality (Silva 2016). Another approach was proposed by Silva (2016) using outliers detection models such as TBATS (Trigonometric Seasonal, Box-Cox Transformation, ARMA residuals, Trend and Seasonality), Symbolic Aggregate approXimation (SAX), Twitter method and Tukey method.

Prior to the application of Outliers detection models, this study proposes a Z-normalization of the annual series and a hierarchical clustering operation with DTW distance measurement to the daily median flow values of each normalized annual series. Clusters are then obtained, which differ from each other at the level of weekly and monthly seasonality. To each of the clusters formed the outliers detection models are applied and the most efficient model in each cluster is chosen. Figure 2.4 describes the methodology applied by Silva (2016) for the application of outliers detection models taking into account the seasonality of the series.

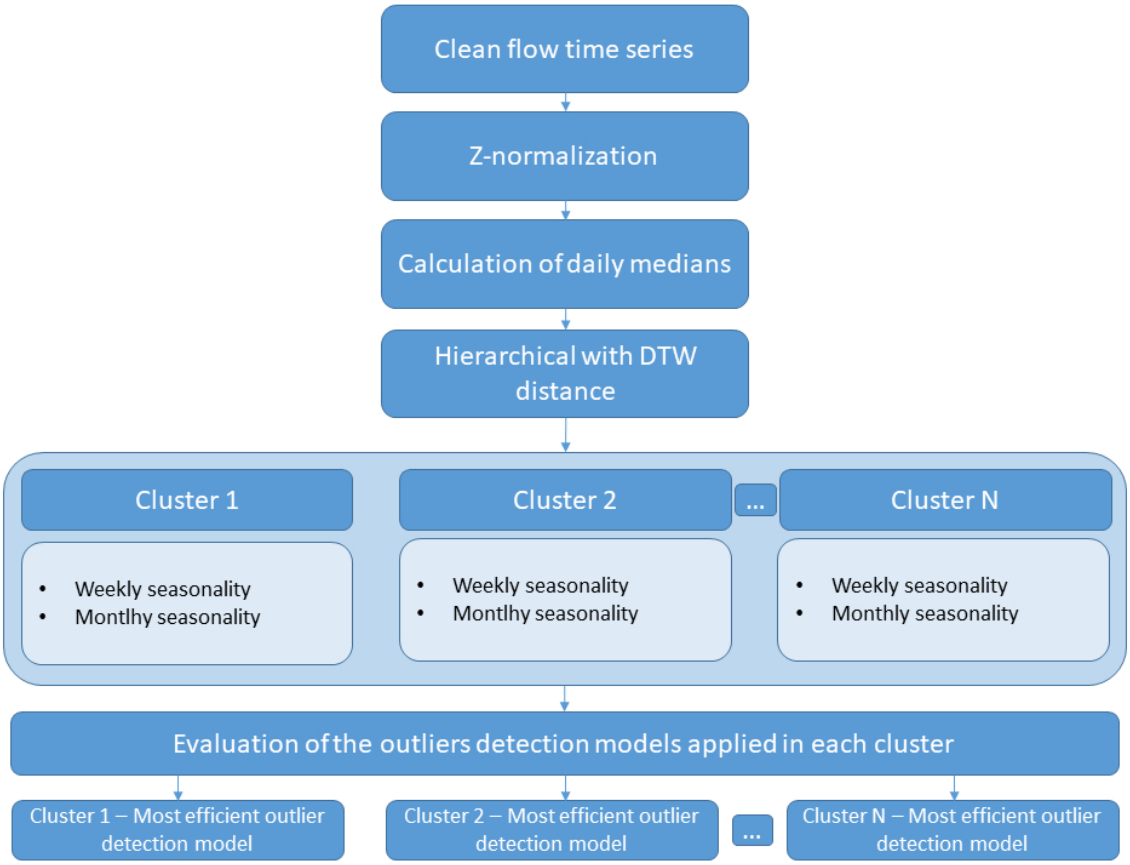

Figure 2.4: Methodology for outliers detection models, based on Silva (2016).

The process of validating outliers in a new series involves a K-nearest neighbors (KNN) classifier to verify which cluster is most similar to the evaluated series and then apply the corresponding outliers detection method associated with the cluster. This way it is guaranteed that the method of detection of outliers is more appropriate to the seasonality of the series to be evaluated.

This study also evaluates methods of prediction of uncertainty in the calculation of water balance, such as: Delta method, Confidence intervals and Monte Carlo method. But in these methods no clustering operations were applied.

2.3.3 Disaggregation of consumption taking into account its use

The increase of tourism that has been verified, as well as the problem of the saline intrusion that has led to the closure of boreholes used in irrigations of gardens. These events have led to a greater demand for water distribution systems in areas of the coast of Portugal and also in tourist areas in the south of the country (Marques 2018). Studies related to the characterization of the external water consumption are relevant for the management entities assigned to these regions.

The study conducted by Marques (2018) aims to characterize the consumption of water for outdoor use. In this study it is proposed a set of predictive models of water consumption based on Generalized Additive Models (GAM).

First step of the construction of the predictive models, is to normalize the series that has internal and external water consumption measurement, and then perform a hierarchical clustering operation with median type prototype for clusters representation. For this clutering operation three distance measures were tested: DTW, Dissimilarity Index Combining Temporal Correlation and Raw Values Behaviours, and Priodogram Based Dissimilarity. The construction of each of the GAM models will be based on the formed clusters.

In addition to the predictive models, a monthly weighting of the consumption of outdoor use against the total water consumption is also calculated for each cluster.

Another important contribution of this study was the development of a methodology for the disaggregation of consumption between uses (indoor vs. outdoor). In this methodology, in a first phase, customer consumption series that only have a single flowmeter are normalized and later a hierarchical clustering is performed, similar to what was done for the predictive models.

The next phase of the methodology is characterized by the use of a KNN classifier (supervised learning) to identify which cluster of predictive models is most similar with each of the clusters formed in the first phase of this methodology. This correspondence allows us to identify which predictive model to use to estimate total consumption and then apply the monthly weighting of external consumption to estimate this type of consumption.

2.3.4 Data reconstruction of flow time series

The existence of reliable and complete information on historical consumption data is extremely important for analysis of water demand profiling, water loss management and real-time management of the systems through telemanagement or telemetry systems.

In the case of DMAs, the flow and pressure series data can often be incorrect (e.g., missing, duplicate or off-scale values). These situations can be the result of a set of problems that occur in the sensors and dataloggers or in the infrastructures of communication and data storage:

1. Power problems in equipment;
2. Poorly calibrated flow meters or pressure sensors;
3. Communication errors between the sensors and the datalogger;
4. Communication errors between the datalogger and the control center;
5. Data storage and processing errors.

The development of data reconstruction models that allow estimation of the missing values based on the values before and after the occurrence of the loss of information are important for an efficient management of the water distribution infrastructures.

In this context, Barrela (2015) presents a study of reconstruction of data of series of

instantaneous flow. The study is based on only three annual series of instantaneous flow with a time step of 15 minutes. Each of these series refers to a different DMA. In this study the author proposes TBATS models with several variants:

1. Forecast method;
2. Backast method;
3. Combined method (forecast and backast).

The author also proposes the JQ method that is based on an Autoregressive integrated moving average (ARIMA) model to reconstruct the daily flow data. It is then combined with the prototype of the daily flow series referring to the week day to which the data is to be reconstructed. In this approach, the author proposes to construct for each month of the annual series a set of average or median prototypes of daily flow series representative of each day of the week.

In this study, no clustering operations were performed prior to the application of data reconstruction methods. This option is understandable since only three annual series of instantaneous flow were available for this study. However, in situations where larger datasets are available, it may be important to perform clustering operations to estimate the prototypes and also to aggregate series with identical seasonal behaviors, with the aim of arranging models that have a good compromise between generalization and overfitting.

2.3.5 Summary and conclusions

Table 2.1 presents the various unsupervised learning methodologies applied in the studies mentioned in previous sections:

Table 2.1: Summary of clustering methods applied

Study	Purpose of the clustering operation	Seasonality	Flow time series type for clustering operation input	Time step	Clustering algorithm	Distance measure	Prototype
Loureiro 2010	Address seasonality in water demand profiling	Monthly	Average daily flow time series for each month	60 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Mean
	Address seasonality in water demand profiling	Daily	Average daily flow time series of each day of the week for each month	60 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Mean
Mamade 2013	Address seasonality in water demand profiling	Monthly	Average daily flow time series for each month	15 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Mean
	Address seasonality in water demand profiling	Daily	Average daily flow time series of each day of the week for each month	15 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Mean
Loureiro et al. 2015	Address seasonality in water demand profiling	Monthly	Average daily flow time series for each month	15 minutes	K-means	Euclidean	Mean
	Address seasonality in water demand profiling	Daily	Average daily flow time series of each day of the week for each month	15 minutes	K-means	Euclidean	Mean
Loureiro et al. 2016	Address seasonality in water demand profiling	Monthly	Average daily flow time series for each month	15 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Median
	Address seasonality in water demand profiling	Daily	Average daily flow time series of each day of the week for each month	15 minutes	Hierarchical Agglomerative (Ward)	Euclidean	Median
Cominola et al. 2016	Characterize flow profiles in water demand profiling	Daily	Daily flow time series	PCA principal components	K-means	Euclidean	Mean
Mounce et al. 2016	Characterize flow profiles in water demand profiling	Daily	Daily flow time series	15 minutes	K-means	Correlation distance	Mean
Cheifetz et al. 2017	Characterize flow profiles in water demand profiling	Weekly	Weekly flow time series	60 minutes for Fourier Regression Model and PCA principal components for K-means model	K-means and Fourier regression mixture model	Euclidean and Fourier regression mixture model distance	Mean and Fourier regression mixture model prototype
Silva 2016	Address seasonality in outlier detection	Montlhy	Daily median flow	Day	Hierarchical Agglomerative (Ward)	DTW*	Median
Marques 2018	Address seasonality in disaggregation of consumption	Montlhy	Daily median flow	Day	Hierarchical Agglomerative (Ward, Complete linkage, Single linkage and Average likage)	DTW* DICTCRVB† and PBD‡	Median
Barrela 2015	No clustering operation was performed	-	-		-	-	-

* Distance Time Warping
† Dissimilarity Index Combining Temporal Correlation and Raw Values Behaviours
‡ Priodogram Based Dissimilarity

As can be seen from Table 2.1, most of the studies presented incorporate a method of clustering in their methodology. Typically the application of a clustering method has the objective of grouping flow time series with similar seasonalities to later train specific supervised learning models for each cluster.

Since the clustering process is an important component in the methodologies, it is verified that in these studies the choice is mainly between hierarchical or k-Means clustering algorithms. The present dissertation intends to explore and evaluate other families of clustering algorithms applicable to the time series domain and consequently applicable to the studies presented in this chapter.

It was also verified that in most of the studies the Euclidean distance was chosen as measure of similarity. The present study will also explore alternative distance measures that allow for some temporal flexibility in order to better capture the forms and seasonality of the time series. This is a path that has been explored in the studies of Silva (2016) and Marques (2018) and that the present dissertation also intends to contribute.

Another important aspect that was verified in the previous studies was the choice of the representative prototype of the clusters to be the average or the median of the series present in the cluster. Since in most of these studies hierarchical clustering was used, the choice of the prototype is not relevant for the formation of the clusters and was only considered as a method of visualization of the characteristics of the series present in the clusters. In this dissertation partition algorithms were analyzed where the choice of prototype is relevant for the formation and quality of the formed clusters.

Through the column "Flow time series type for clustering operation input" of Table 2.1, it was verified that in the presented studies flow data was used in a grouped way as input for the clustering operations. This option allows for less complex clustering operations since the distance matrices will have smaller dimensions. However, with the previous data grouping, useful information can be hidden in the formation of the groups by the clustering operation. Since the focus of the present study is on clustering operations, it was decided not to aggregate information prior to the application of clustering methods. This choice naturally introduces larger distance matrices and a greater number of comparisons to verify similarity, requiring the planning of a system with greater computational capacity.

It is also verified that, in the majority of the studies presented, there is a greater focus on the Monthly Seasonality. The present dissertation will focus on clustering normalized daily time series of medium flow rate (15 minutes timestep), intending to explore more the weekly seasonality.

Chapter 3

Methodology

3.1 Overview

Section 2.2 gives an overview of the various clustering techniques applicable to time series. This chapter aims to further describe the methods and the various components that make up the clustering models to be analyzed in chapter 4.

In addition to the components of clustering models, this chapter describes the Boxplot method that will be used for data preprocessing and dataset description in chapter 4.

This chapter also describes the methods used for performance evaluation of clustering models, namely the internal index measures and the Principal Component Analysis method used to view the clusters formed by the clustering models.

The present chapter is organized as follows:

- **3.2 General methodology**: an overview of the methodology to be applied in this dissertation;
- **3.3 Boxplot method**: technique used in preprocessing for dataset characterization and outlier analysis;
- **3.4 Data normalization method**: method used for series normalization to compare series that come from different DMAs (with different flow amplitudes);
- **3.5 Timeseries clustering Algorithms**: definition of clustering algorithms to be used in clustering models;
- **3.6 Distance measures**: definition of distance measures that will be used in clustering models;
- **3.7 Prototype Methods**: definition of prototypes that will be used in clustering models;
- **3.8 Internal Indexes methods**: definition of the internal index measures that will be used to evaluate clustering models;
- **3.9 Principapal Component Analysis**: definition of the method used to visualize clusters formed by clustering models;
- **3.10 Definition of clustering models and methodology application**: presents

the models according to their model components in the context of the analysis methodology to be applied in chapter 4.

3.2 General methodology

Figure 3.1 shows the sequence of steps of the general methodology.

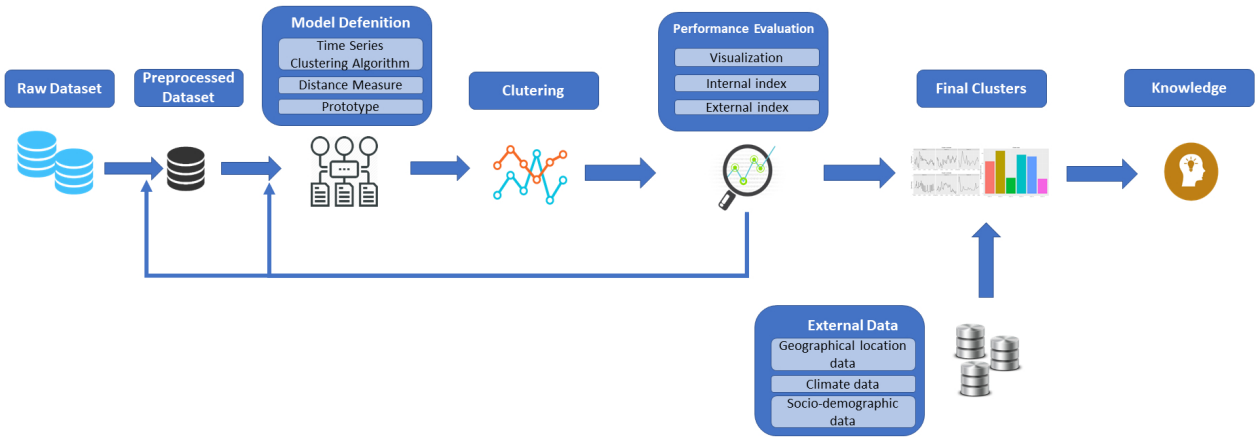

Figure 3.1: General methodology.

As can be seen from the sequence shown in Figure 3.1, the general methodology for clustering methods evaluation is an iterative process that groups the following steps:

1. **Raw Data**: represents flow time series of medium flow rate with 15 minutes step collected over 1 year from 52 DMAs distributed across Portugal;
2. **Data Pre-Processing**: in this stage, characterization of the flow series will be performed through the graphic visualization of descriptive statistics such as the median, average, 1^{st} quartile, 3^{rd} quartile and outliers through bar graphs, boxplots and violin plots. These analyzes will allow the evaluation of the amplitude of the flow rates present in each series as well as to identify abnormal flow records present in the series (e.g., negative flow rates and extremely high flow rates). Days with anomalous flow values will be removed as they occur due to errors in the data recording process and don't represent anomalous events occurring in the infrastructure. Afterwards, the annual series will be split into daily series and these will be normalized so that in clustering operations the daily series can be compared regardless of the DMA they come from;
3. **Model Defenition**: as already mentioned in section 2.2.1 the definition of a model involves the selection of time series algorithms, distance measurements and prototypes. In the present chapter the methods used in these 3 components of model definition will be described in detail;
4. **Clustering Operation**: this phase describes the execution of the defined model;
5. **Clustering Performance Evaluation**: after the clustering model implementation it is necessary to evaluate the formed groups. At this stage internal index measures, as

well as visual methods, are used by projecting the points of the various groups according to the first 3 components of the Principal Component Analysis (PCA). At the end of this iterative process, a set of models characterized by the size of the formed clusters and also by the ability of the models to identify clusters with different day typologies (e.g., weekend vs. working day) will be defined and the most performing models are selected;

6. **Final Cluster Models Analysis**: the prototypes of the clusters of the most performing models of the previous stage will be analyzed according to geographical location and also by the predominance of the season of the year in order to identify distinctive behaviors according to these dimensions;
7. **Knowledge Extraction**: in this step we intend to create a combined model that tries to group the characteristics of the most performing models and thus to characterize more completely the behaviors present in the daily patterns throughout the dataset.

3.3 Boxplot method

The boxplot method allows the visualization of the main statistical characteristics of the data in summary form (Benjamini 1988; Kampstra 2007; Patil et al. 2018; Williamson et al. 1989). Figure 3.2 represents a boxplot and a probability density function applied to a Normal population (N(0,1σ^2)).

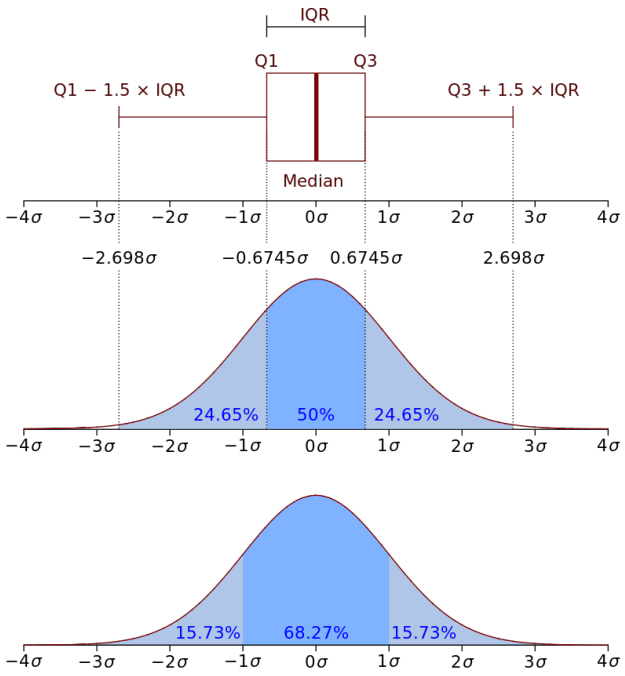

Figure 3.2: Boxplot method (Chen-Pan (2012)).

Figure 3.2 shows the central rectangle that gives the name of this chart typology. The third quartile (Q3) is represented by the vertical line that makes up the right side of the rectangle.

In contrast the first quartile (Q1) is represented by the vertical line that makes up the left side of the rectangle. The median is a vertical line drawn inside the rectangle. The step corresponds to 1.5 times the interquartile range (Q3-Q1). From the center of the right side of the rectangle a horizontal line with the step range is constructed. Similarly, the same procedure is performed from the center of the left side of the rectangle. The dataset points that are beyond the step range can be considered outliers or noise points according to the characteristics of the dataset domain (Benjamini 1988; Frigge et al. 1989; Patil et al. 2018).

This method will be used in chapter 4 for the initial characterization of data and detection of flow values that may be considered Outliers.

3.4 Data normalization method

In order to clustering algorithms focus on structural similarities / dissimilarities (shape) and not on amplitude driven ones in the formation of clusters. A process of normalization of time series is performed previously. The method used in Chapter 4 is Z-score (Z-normalization or Standard score).

Equation (3.1) presents this normalization method (MathWorks 2019; Senin 2016).

$$z_i = \frac{x_i - \mu}{\sigma}, \text{where i} \in \mathbb{N} \tag{3.1}$$

Where μ and σ are the population's mean and standard deviation, respectively.

3.5 Time series clustering algorithms

In this section the clustering algorithms used in the analyzes performed in chapter 4 will be presented. These algorithms in combination with distance measurements (section 3.6) and prototype functions (section 3.7) will allow to form clusters that characterize the dataset under study.

3.5.1 Hierarchical clustering

Hierarchical clustering performs a grouping as the hierarchical level increases. Clusters are formed by joining clusters from the level immediately below, thus obtaining an orderly sequence of clusters (Hastie et al. 2009; Sarda-Espinosa 2017). This family of algorithms can be subdivided into bottom-up or top-down approaches as described in Chapter 2. But the buttom-up strategy is more common (Hastie et al. 2009) and will be the strategy applied in the hierarchical clustering analysis in Chapter 4.

For hirarchical clustering it is not necessary a priori to specify the number of clusters to form, on the other hand this typology of clustering algorithm is deterministic which means it will always give the same result for a specified distance measurement. Both approaches to hierarchical custering have little flexibility, which means that every time a top-down or buttom-up split occurs, no adjustments can be made (Sarda-Espinosa 2017).

Bottom-up hierarchical clustering typically follows the following steps (Matteucci 2019b):

1. Each object is assigned to a cluster, so that having N objects has N clusters each with an object;
2. The distances between the various clusters are calculated;
3. The most similar (shorter distance) pair of clusters is joined into a single cluster so as to have one less cluster;
4. The distances between the newly formed cluster and the other clusters that did not undergo a merge operation in the previous step are calculated;
5. Steps three and four are repeated until all objects belong to a single cluster.

Step four distances can be calculated using a linkage criterion of the following typologies (Matteucci 2019b):

- **Complete-linkage**: the distance between two clusters is the maximum distance between a member of one cluster and a member of another cluster;
- **Single-linkage**: the distance between two clusters is the minimum distance between a member of one cluster and a member of another cluster;
- **Centroid-linkage**: the distance between two clusters corresponds to the distance between a centroid of one cluster and the centroid of another cluster;
- **Average-linkage**: the distance between two clusters is the average of the distances between members of one cluster and members of the other cluster.

The methods of single-linkage and centroid-linkage will not be used. The former will not be used because it typically forms unbalanced dendrograms and the latter typically presents inversions in the dendrograms (Legendre 2012).

The average-linkage method will also not be used because although it presents balanced dendograms it does not allow to visualize the intra-cluster distance in the dendogram for each cluster formed as the complete-linkage method allows. In Chapter 4 Euclidean distance will be used in complete-linkage criterion.

3.5.2 Partiton clustering

Partition algorithms identify as methods that minimizes intracluster distance and maximizes intercluster distance. To achieve this goal, this family of algorithms uses iterative greedy descent strategies that scan a portion of the search space until they find convergence. However, through this strategy one can converge to a local minimum instead of an absolute minimum (Sarda-Espinosa 2017).

This family of clustering algorithms typically uses the following steps (Matteucci 2019c):

1. K centroids are initialized randomly (usually k randomly chosen dataset objects);
2. A distance measurement (section 3.6) calculates the distances of all objects to the centroids and all objects are subsequently assigned to the nearest centroid;
3. A prototype function (section 3.7) is applied to each cluster to calculate a new representative cluster centroid;

4. Steps two and three are iteratively repeated until a maximum number of iterations have been reached or there are no cluster-changing objects.

This methodology always attempts to maintain the number of initially assigned clusters, which may result in instability or divergence in some cases. In these situations a different distance measurement is used or the number of clusters k to be formed is decreased.

This family of clustering algorithms requires that the cluster number to be formed (k) be assigned initially, but generally the ideal number of clusters to form is not known. To get around this, the algorithm is run with different numbers of clusters to form and use cluster validation indices (section 3.8) to evaluate which number of clusters best fits the dataset to be studied.

3.5.3 k-Shape clustering

This algorithm is a particular case of partition algorithms because it uses the custom distance measurement Shape based distance (section 3.6.4) and Shape extraction prototype function (section 3.7.4) (Paparrizos and Gravano 2015).

3.5.4 Fuzzy clustering

This method belongs to the partition algorithm family but allows you to make a soft or fuzzy partition so that each member belongs to each cluster to a certain degree. In contrast, traditional partition algorithms and hierarchical algorithms that make hard-type partitions where each member belongs exclusively to one cluster and the clusters are mutually exclusive (Matteucci 2019a; Sarda-Espinosa 2017).

Equation (3.2) defines the minimization function:

$$min \sum_{i=1}^{N} \sum_{j=1}^{k} u_{i,j}^{m} \| x_i - c_j \|^2 \quad , \quad 1 \leq m \leq \infty \tag{3.2}$$

In which:

$$\sum_{j=1}^{k} u_{i,j} = 1 \quad \text{and} \quad u_{i,j} \geq 0 \tag{3.3}$$

Where x_i and c_j are the ith of d-dimensional measured data and the d-dimension center of the cluster, respectively. The variables $u_{i,j}$ and m are the degree of membership of x_i and the fuzziness exponent with a common value of 2, respectively. The variable $\| * \|$ is the norm expressing the similarity between any measured data and the centroid.

The minimization function for Fuzzy partitioning can be performed using an iterative update process of membership u_{ij} and centroids c_j until the stop criterion $\| U^{k+1} - U^k \| \leq \varepsilon$ is met. The iterative process is described according to the following steps (Matteucci 2019a):

1. Matrix $U^0 = [u_{ij}]$ initialization;
2. At each iteration k, calculate the centroids $C^k = [c_j]$ with U^k using the equation (3.19).
3. At each k iteration, update the matrix from U^k to U^{k+1} through the equation:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3.4}$$

4. Verify that the $\| U^{k+1} - U^k \| \leq \varepsilon$ stop criterion is met. If not fulfilled return to step 2.

3.6 Distance measures

Distance measurements are one of the important elements in the definition of a clustering model, as they provide a way to calculate dissimilarity between two time series. For any time series clustering algorithm the calculation of distances and cross-distance matrix is essential for the formation of groups that maximize similarity between group members and minimize similarity between groups (Sarda-Espinosa 2017).

This section will present the definitions of distance measurements that will be used in Chapter 4.

3.6.1 Euclidean

One of the most common distance measurements used in clustering operations is Euclidean distance. This measurement can be computed efficiently but is sensitive to noise, scale, time shifts and can only be used in situations where time series are of equal size (Sarda-Espinosa 2017).

Equation (3.5) gives the definition of Euclidean distance.

$$d(p, q) = d(q, p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \tag{3.5}$$

Where $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ are two objects in Euclidean n-space.

3.6.2 Dynamic time warping (DTW)

Dynamic time warping (DTW) is a dynamic programming algorithm that calculates an optimum warping path between two time series. The first step in calculating the DTW distance is to create a local cost matrix (LCM), which has a dimension of $n \times m$, for each comparison between two time series (Sarda-Espinosa 2017).

The equation (3.6) presents the local cost matrix calculation formula based on a Euclidean space. It is denoted by the v in the equation that this method allows time series with multivariables.

$$lcm(i, j) = \sqrt{\sum_{v} \mid x_i^v - y_i^v \mid^2} \tag{3.6}$$

The equation (3.6) considers x and y as input series, where for each element (i, j) of the lcm matrix the norm between x_i and y_j is calculated.

In the second step, the DTW algorithm looks for an alignment between x and y that minimizes the aggregate cost through an iterative process of walking over the lcm matrix, starting at

$lcm(1,1)$ and ending at $lcm(n,m)$. In every step the algorithm goes in the direction in which the aggregate cost increases less given certain constraints (Giorgino 2009; Sarda-Espinosa 2017).

Equation (3.7) defines the calculation of the distance DTW between two time series.

$$DTW(x,y) = \sqrt{\sum \frac{m_\phi lcm(k)^2}{M_\phi}}, \forall k \in \phi \tag{3.7}$$

In which $\phi = \{(1,1),...,(n,m)\}$ and m_ϕ are the set of points that belongs to the optimal path and the per-step weighting, respectively. The variable M_ϕ is a normalization constant.

Figure 3.3 illustrates the optimum warping path and alignment between two time series.

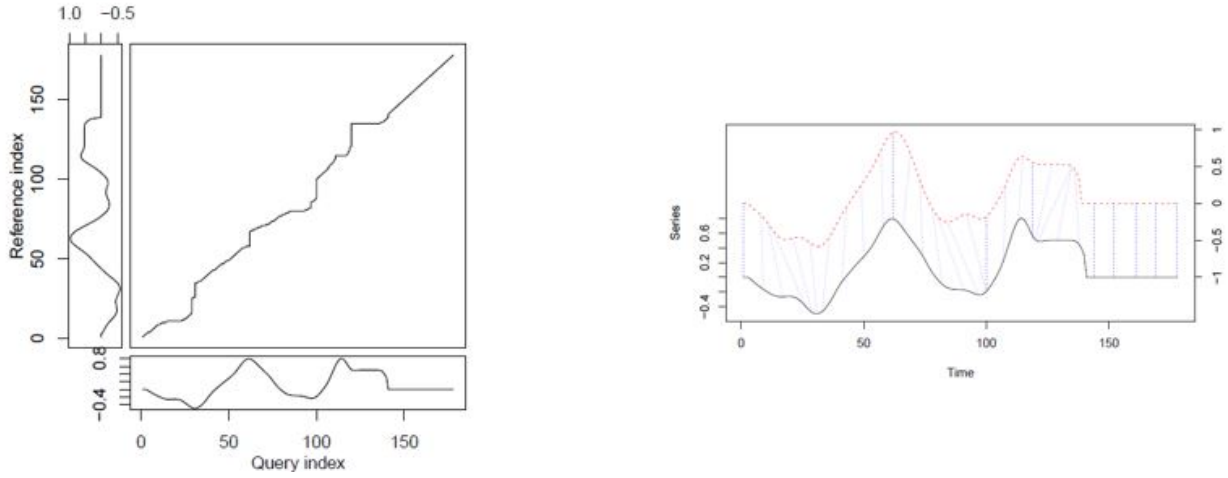

Figure 3.3: Otimum path found (on the left) and alignment (on the right) between two time series, based on Sarda-Espinosa (2017).

DTW is computationally expensive. If x has length n and y has length m, the DTW distance between them can be computed in O(nm) time, which is quadratic if m and n are similar. Additionally, the DTW distance can potentially deal with series of different length directly. This is not necessarily an advantage, as it has been shown before that performing linear reinterpolation to obtain equal length may be appropriate if m and n do not vary significantly (Ratanamahatana and Keogh 2004).

One of the possible modifications of DTW to deal with complexity is to use window constraints. These limit the area of the LCM that can be reached by the algorithm. One of the most common ones is the Sakoe-Chiba window (Sakoe and Chiba 1978), with which an allowed region is created along the diagonal of the LCM. These constraints can marginally speed up the DTW calculation, but they are mainly used to avoid pathological warping. It is common to use a window whose size is 10% of the series' length, although sometimes smaller windows produce even better results (Ratanamahatana and Keogh 2004; Sarda-Espinosa 2017).

Figure 3.4 shows Sakoe-Chiba window constraints in lcm matrix.

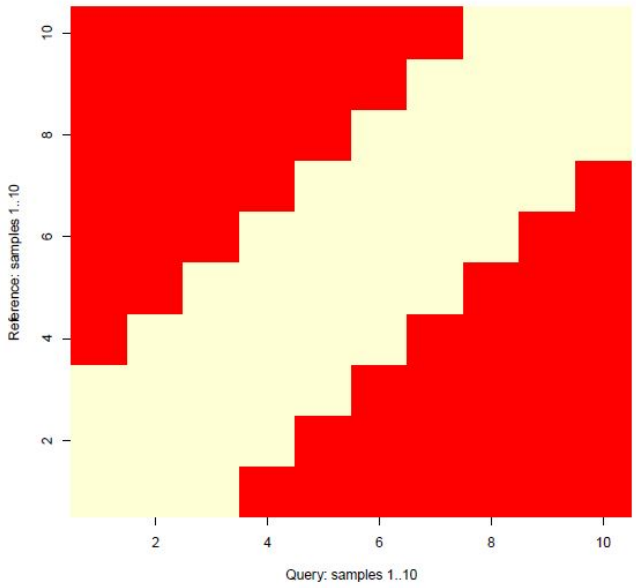

Figure 3.4: Sakoe-Chiba constraint for DTW. The red elements will not be considered by the algorithm when traversing the LCM (Sarda-Espinosa (2017)).

3.6.3 Global alignment kernel (GAK)

Global Alignment kernels (GAK) allow you to consider the cost of all alignments between two time series by calculating a soft minimum, enabling a more concise quantification of similarities than DTW (Sarda-Espinosa 2017). Equation (3.8) explains the calculation of similarity between two time series according to a global alignment kernel.

$$k_{GA}(x, y) = \sum_{\pi \in A(n,m)} \prod_{i=1}^{|\pi|} \kappa(x_{\pi_1(i)}, y_{\pi_2(i)}) \tag{3.8}$$

Where π and $\mid \pi \mid$ are the alignment between two series x and y and the length of π, respectively. The variables κ and $A(n, m)$ are the local similarity function and the set of all possible alignments constrained by the lengths of x and y, respectively.

The consideration of the various alignments makes this methodology have some limitations such as diagonal dominance and complexity O(nm) (Sarda-Espinosa 2017). However the diagonal dominance is not relevant as long as one of the series is not twice as long as the other (Cuturi 2011).

Regarding the complexity problem O(nm), it is possible to reduce complexity by using a tringular local kernel (TLK) which limits the number of alignments to be considered for GAK calculation of similarity. Equation (3.9) presents the formulation of the TLK.

$$w(i, j) = \left(1 - \frac{\mid i - j \mid}{T}\right) \tag{3.9}$$

Where T represents the order of the kernel.

Combining TLK with GAK gives the triangular global alignment kernel (TGAK) where the number of alignments to consider for calculating similarity is constrained by TLK. This kernel can be calculated with a complexity of O(Tmin(n,m)) (Sarda-Espinosa 2017).

Equation (3.10) formulates the GAK (based on Gaussian kernel).

$$k(x,y) = e^{-\left(\frac{1}{2\sigma^2}\|x-y\|^2 + log\left(2-e^{-\frac{\|x-y\|^2}{2\sigma^2}}\right)\right)} \tag{3.10}$$

In which:

$$\sigma = c \cdot med(\| x - y \|) \cdot \sqrt{med(| x |)} \tag{3.11}$$

Where $med(.)$ and c are the empirical median and a constant (the value of 1 will be used since it is the value adopted in R package dtwclust (Sarda-Espinosa 2019)), respectively. The variables x and y are subsampled vectors from the dataset.

The TGAK is then defined by equation (3.12).

$$TGAK(x,y,\sigma,T) = \frac{w(i,j)k(x,y)}{2 - w(i,j)k(x,y)} \tag{3.12}$$

when:

- $T = 0$ **or** $T \to \infty$ - all alignments are considered and TGAK converges to the original GAK;
- $T = 1$ - only compares series of equal length;
- $T > 1$ - only alignments that meet the constraint $-T < \pi_1(i) - \pi_2(i) < T$ are considered.

The similarity obtained by TGAK can be normalized between 0 and 1. The distance measurement is obtained by subtracting 1 by the value of the normalized TGAK similarity. Equation (3.13) formulates the distance measurement obtained through the normalized TGAK similarity:

$$D_{x,y} = 1 - e^{log(TGAK(x,y,\sigma,T)) - \frac{log(TGAK(x,x,\sigma,T)) + log(TGAK(y,y,\sigma,T))}{2}} \tag{3.13}$$

3.6.4 Shape-based distance (SBD)

Associated with the Shape-based clustering algorithm is the shape-based distance measurement (SBD). This measurement is based on cross-correlation with coefficients normalization (NCC_c) which makes it scale sensitive, so time series z-normalization must be done before applying this method (Paparrizos and Gravano 2015). After obtaining the NCC_c sequences, the SBD distance is calculated according to equation (3.14).

$$SBD(x,y) = 1 - \frac{max(NCC_c(x,y))}{\| x \|_2 \| y \|_2} \tag{3.14}$$

Where $\| \cdot \|$ is the Euclidean norm. This distance measure varies between 0 and 2, where 0 indicates perfect similarity between the two series. SBD presents itself as a faster alternative than DTW, efficiently utilizing Fast Fourier Transform (FFT) to obtain the NCC_c sequences (Paparrizos and Gravano 2015).

3.7 Prototype methods

Prototyping methods are an important component in defining a time series clustering model. This component has special relevance in partition clustering algorithms since prototypes are used as cluster centroids in the iterative group formation process. Apart from the clustering algorithm, the choice of prototype method is intrinsically related to the distance measurement to be used and similar to distance measurements, it is typically not known a priori which prototype method is the best (Sarda-Espinosa 2017).

In this section we will present the various prototype methods that will be evaluated in Chapter 4.

3.7.1 Mean

The use of arithmetic mean as a prototype of clusters is quite common when associated with the Euclidean distance average. However, due to the structure of the time series, the arithmetic mean prototype is considered a poor choice and may affect convergence of the clustering algorithm (Sarda-Espinosa 2017).

The cluster prototype calculation according to this method considers the average of each time-point i for all variables taking into account all time series belonging to the group. Equation (3.15) formulates the prototype calculation according to this method for a cluster C of size N.

$$\mu_i^v = \frac{1}{N} \sum_c x_{c,i}^v, \forall c \in C \tag{3.15}$$

In which $x_{c,i}^v$ is the i-th element of the v-th variable from the c-th series that belongs to cluster C.

3.7.2 Partition around medoids (PAM)

The prototype according to partition around medoids (PAM) is defined as an object of a cluster where the average distance to the other elements of the cluster is minimal. This approach has some advantages over mean approaches since medoid is always an element of the original data series (Kaufmann and Rousseeuw 1987; Sarda-Espinosa 2017).

The equation (3.16) defines the calculation of a medoid prototype

$$x_{medoid} = argmin_{y \in (x_1, x_2, ..., x_n)} \sum_{i=1}^{n} d(y, x_i) \tag{3.16}$$

In which $x_1, x_2, ..., x_n$ is a set of n points in a space with a distance function d.

3.7.3 DTW barycenter averaging (DBA)

DTW barycenter averaging (DBA) is a method that iteratively redefines an average sequence with the objective of minimize the sum of squared DTW distances from the average sequence to the set of sequences. This sum is formed by single distances between each coordinate of the average sequence and coordinates of sequences associated to it. Thus, the contribution of one coordinate of the average sequence to the total sum of squared distance is actually a sum of euclidean distances between this coordinate and coordinates of sequences associated to it during the computation of DTW. Note that a coordinate of one of the sequences may contribute to the new position of several coordinates of the average. Conversely, any coordinate of the average is updated with contributions from one or more coordinates of each sequence. In addition, minimizing this partial sum for each coordinate of the average sequence is achieved by taking the barycenter of this set of coordinates (Petitjean et al. 2011).

The process of refinement of the average sequence is composed by the following steps:

1. Computing DTW between each individual sequence and the temporary average sequence to be refined, in order to find associations between coordinates of the average sequence and coordinates of the set of sequences;
2. Updating each coordinate of the average sequence as the barycenter of coordinates associated with it during the first step.

The process of updating the average sequence can be defined by:

$$C_t' = barycenter(assoc(C_t)) \tag{3.17}$$

In which:

$$barycenter\left\{X_1, ..., X_\alpha\right\} = \frac{X_1 + ... + X_\alpha}{\alpha} \tag{3.18}$$

Where $assoc$ and α are the function that links each coordinate of the average sequence to one or more coordinates of sequences to be averaged and the number of sequences associated to C, respectively. The variables $C =< C_1, ..., C_T >$ and $C' =< C_1', ..., C_T' >$ represent the average sequence at iteration i and the update of C at iteration i+1, respectively. The variables $X_1 + ... + X_\alpha$ are coordinates of the set of sequences associated to C during the first step.

3.7.4 Shape extraction

This method to calculate time-series prototypes is part of the k-Shape algorithm. For this method centroids are selected by an optimization problem where the objective is to find the minimizer of the sum of squared distances to all other timeseries sequences. However, as cross-correlation intuitively captures the similarity rather than the dissimilarity of time series, we can express the computed sequence as the maximizer of the squared similarities to all other time-series sequence (Paparrizos and Gravano 2015; Sarda-Espinosa 2017).

As this approach is used in the context of iterative clustering, the previously computed centroid is used as reference and align all sequences towards this reference sequence. Since the previous centroid will be very close to the new centroid. For this alignment SBD

distance is used, which identifies an optimal shift for every sequence. Subsequently, as sequences are already aligned towards a reference sequence, it is performed a so-called maximization of Rayleigh Quotient to obtain the final prototype (Paparrizos and Gravano 2015; Sarda-Espinosa 2017).

3.7.5 Fuzzy-based prototype

The centroid function used by fuzzy c-means calculates the mean for each point across all members in the data, weighted by their degree of belongingness (Matteucci 2019a; Sarda-Espinosa 2017).

Equation (3.19) describes the calculation of centroids by this method.

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{3.19}$$

Where c_j and x_i are the d-dimension center of the cluster and the ith measured data of d-dimensional, respectively. The variables u_{ij} and m represent the degree of membership of x_i and the fuzziness exponent with a common value of 2, respectively.

3.8 Internal index methods

One of the important steps of this general methodology is to identify how the clusters formed by each algorithm fit the data. This question is complex to answer because different clustering algorithms produce different clusters and none of them are proven to be the best for all situations. Cluster validation is the process responsible for estimating how well formed clusters fit the underlying structure of the data (Arbelaitz et al. 2013).

In addition to the comparison between clustering algorithms, it should also be noted that cluster validation is also used in algorithms that a priori cannot determine the number of clusters that naturally exist in the data and need to initially provide the number of clusters to be formed. In these cases it is usual to run the algorithm several times and with different number of clusters to form and evaluate each of the iterations by a cluster validation process in order to obtain the number of ideal clusters to be formed for the dataset under analysis.

Cluster validation can be separated into external validation or internal validation as defined in chapter 2. Since the underlying structure of the data is not known for the dataset to be analyzed in chapter 4, cluster validation methods will be based on internal validation. The following sections will present internal index validation methods that directly estimate the quality of the formed clusters based on the measurement of the cohesion and separation of the formed clusters.

The internal validation indexes to be used in Fuzzy clustering algorithms are different from those used in other clustering approaches since for Fuzzy algorithms the degree of cluster membership has to be taken into account when calculating an internel index. Thus internal index algorithms will be classified as follows:

- **Internal indexes for hard partitions:** evaluation measures applicable to hard patition clustering algorithms (k-Means, k-Medoids and Hirarchical clustering);
- **Internal indexes for fuzzy partitiions** evaluation measures applicable to soft partition clustering algorithms (Fuzzy custering).

3.8.1 Internal indexes for hard partitions

Silhouette index

For silhouette index cohesion is measured based on the distance between all points in the same cluster and separation is based on the distance from the nearest neighbor(equation (3.20)). For this index a larger value indicates a better partition (Arbelaitz et al. 2013; Rousseeuw 1987).

$$Sil(C) = 1/N \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b\left(x_i, c_k\right) - a\left(x_i, c_k\right)}{max\left\{a\left(x_i, c_k\right), b\left(x_i, c_k\right)\right\}} \tag{3.20}$$

In which:

$$a\left(x_i, c_k\right) = 1/|c_k| \sum_{x_j \in c_k} d_e(x_i, x_j) \tag{3.21}$$

$$b\left(x_i, c_k\right) = \min_{c_l \in C \setminus c_k} \left\{ 1/|c_l| \sum_{x_j \in c_l} d_e(x_i, x_j) \right\} \tag{3.22}$$

Dunn index

This index is composed of the ratio between the estimated cohesion using nearest neighbor distance and the separation by the maximum cluster diameter value (equation (3.23)). For this index a larger value indicates a better partition (Arbelaitz et al. 2013; C. Dunn 1973).

$$D(C) = \frac{min_{c_k \in C}\left\{min_{c_{\in} C \setminus c_k}\left\{\delta(c_k, c_l)\right\}\right\}}{max_{c_k \in C}\left\{\triangle(c_k)\right\}} \tag{3.23}$$

In which:

$$\delta(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_l} \left\{d_e(x_i, x_j)\right\} \tag{3.24}$$

$$\triangle(c_k) = \max_{x_i x_j \in c_k} \left\{d_e(x_i, x_j)\right\} \tag{3.25}$$

COP index

This index is presented as the ratio between the cohesion estimated by distance from the points in a cluster to its centroid and the separation which is the furthest neighboring distance (equation (3.26)). For this index a smaller value indicates a better partition (Arbelaitz et al. 2013; Gurrutxaga et al. 2010).

$$COP(C) = \frac{1}{N} \sum_{c_k \in C} |c_k| \frac{1/|c_k| \sum_{x_i \in c_k} d_e(x_i, \overline{c_k})}{min_{x_i \notin c_k} max_{x_j \in c_k} d_e(x_i, x_j)} \tag{3.26}$$

Davies-Bouldin index

Through this index cohesion is estimated based on the distance of the points in a cluster to its centroid and the separation based on the distance between the centroids (equation (3.27)). For this index a smaller value indicates a better partition (Arbelaitz et al. 2013; L. Davies and Bouldin 1979).

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{d_e(\overline{c_k}, \overline{c_l})} \right\} \tag{3.27}$$

In which:

$$S(c_k) = 1 \diagup |c_k| \sum_{x_i \in c_k} d_e(x_i, \overline{c_k}) \tag{3.28}$$

Modified Davies-Bouldin index

This index is a variation of the previous method according to equation (3.29). For this index a smaller value indicates a better partition (Arbelaitz et al. 2013; Kim and Ramakrishna 2005).

$$DB^*(C) = \frac{1}{K} \sum_{c_k \in C} \frac{max_{c_l \in C \setminus c_k} \{S(c_k) + S(c_l)\}}{min_{c_l \in C \setminus c_k} \{d_e(\overline{c_k}, \overline{c_l})\}} \tag{3.29}$$

Calinski-Harabasz index

This index is based on the ratio of the estimated cohesion based on the distance of points in a cluster to its centroid and the separation based on the distance from centroids to a global centroid (equation (3.30)). For this index a larger value indicates a better partition (Arbelaitz et al. 2013; Caliński and JA 1974).

$$CH(C) = \frac{N - K \sum_{c_k \in C} |c_k| d_e(\overline{c_k}, \overline{X})}{K - 1 \sum_{c_k \in C} \sum_{x_i \in c_k} d_e(x_i, \overline{c_k})} \tag{3.30}$$

Score Function

In this index the separation is measured based on the distance between the cluster centroids and the global centroid and the cohesion is based on the distance from the points of a cluster to its centroid (equation (3.31)). For this index a larger value indicates a better partition (Arbelaitz et al. 2013; Saitta et al. 2007).

$$SF(C) = 1 - \frac{1}{e^{e^{bcd(C) + wcd(C)}}} \tag{3.31}$$

In which:

$$bcd(C) = \frac{\sum_{c_k \in C} |c_k| d_e(\overline{c_k}, \overline{X})}{N \times K} \tag{3.32}$$

$$wcd(C) = \sum_{c_k \in C} 1/|c_k| \sum_{x_i \in c_k} d_e(x_i, \overline{c_k}) \tag{3.33}$$

3.8.2 Internal indexes for fuzzy partitions

Modified Partition Coefficient index

This index only takes into account membership values by minimizing the overall content of pairwise fuzzy intersection in U (partition matrix). The index $PC(C)$ corresponds to the average relative amount of membership sharing done between pairs of fuzzy subsets. The $MPC(C)$ index is a modification to $PC(C)$ that reduces the monotonic evolution tendency imposed by number of clusters increases (c) (equation (3.34)). For this index a larger value indicates a better partition (Dave 1996; Wang and Zhang 2007).

$$MPC(C) = 1 - \frac{c}{c-1}(1 - PC(C)) \tag{3.34}$$

In which:

$$PC(C) = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} \tag{3.35}$$

Kwon index

This index takes into account membership values as well as the dataset itself (equation (3.36)). To deal with monotonic evolution tendency imposed by number of clusters increase, this index introduces a punishing function $\frac{1}{c}\sum_{i=1}^{c} \parallel v_i - \overline{v} \parallel^2$. For this index a smaller value indicates a better partition (Kwon 1998; Wang and Zhang 2007).

$$K(C) = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^2 \parallel x_j - v_i \parallel^2 + \frac{1}{c}\sum_{i=1}^{c} \parallel v_i - \overline{v} \parallel^2}{min_{i\neq k} \parallel v_i - v_k \parallel^2} \tag{3.36}$$

In which:

$$\overline{v} = \sum_{j=1}^{n} \frac{x_j}{n} \tag{3.37}$$

Improved Validation index

This index is a variation of the previous method according to equation (3.38) whith a different punishing function $\frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{\substack{k=1 \\ k\neq i}}^{c} \parallel v_i - \overline{v} \parallel^2$. For this index a smaller value indicates a better partition (Tang et al. 2005; Wang and Zhang 2007).

$$T(C) = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \parallel x_j - v_i \parallel^2 + \frac{1}{c(c-1)}\sum_{i=1}^{c}\sum_{\substack{k=1 \\ k\neq i}}^{c} \parallel v_i - \overline{v} \parallel^2}{min_{i\neq k} \parallel v_i - v_k \parallel^2 + \frac{1}{c}} \tag{3.38}$$

Validity Function

This index also takes into account membership values as well as the dataset itself, but introduces the concepts of fuzzy conpactness and fuzzy separation in the same way as the traditional internal index of validation presented in section 3.8.1 (equation (3.39)). The SC_1 component considers the degree of compaction and separation across membership and geometric properties of the data structure. The SC_2 component introduces the concept of fuzzy union and fuzzy intersection to achieve the degree of fuzzy conpactness/separation. For this index a larger value indicates a better partition (Wang and Zhang 2007; Zahid et al. 1999).

$$SC(C) = SC_1(c) - SC_2(c) \tag{3.39}$$

In which:

$$SC_1(c) = \frac{\sum_{i=1}^{c} \parallel v_i - \overline{v} \parallel^2 \, / c}{\sum_{i=1}^{c} \left(\sum_{j=1}^{n} (u_{ij}^m) \parallel x_j - v_i \parallel^2 \, / \sum_{j=1}^{n} u_{ij} \right)} \tag{3.40}$$

$$SC_2(c) = \frac{\sum_{i=1}^{c} \sum_{l=i+1}^{n} \left(\sum_{j=1}^{n} (min(u_{ij}, u_{lj}))^2 / \sum_{j=1}^{n} min(u_{ij}, u_{lj}) \right)}{\sum_{j=1}^{n} (max_{1 \leq i \leq c} u_{ij})^2 / \sum_{j=1}^{n} (max_{1 \leq i \leq c} u_{ij})} \tag{3.41}$$

PBMF index

Similar to the Validity Function presented earlier, this index also makes use of the traditional concepts of compactness and separation that the indexes in section 3.8.1 present (equation (3.42)). The $\frac{1}{c}$ component indicates the divisibility of a c cluster system. The $\frac{E_1}{J_m}$ component is a measure of c cluster system compactness. The D_c component is the maximum intercluster separation in c cluster system. For this index a larger value indicates a better partition (Pakhira et al. 2004; Wang and Zhang 2007).

$$PBMF(C) = \left(\frac{1}{c} \times \frac{E_1}{J_m} \times D_c \right)^2 \tag{3.42}$$

In which:

$$E_1 = \sum_{j=1}^{n} u_{ij} \parallel x_j - v \parallel \tag{3.43}$$

$$D_c = \max_{i,j=1}^{c} \parallel v_i - v_j \parallel \tag{3.44}$$

$$J_m = \sum_{j=1}^{n} \sum_{i=1}^{c} (u_{ij})^m \parallel x_j - v_i \parallel \tag{3.45}$$

3.8.3 Range applied to the internal indexes

Using the cvi function included in the R dtwclust package (Sarda-Espinosa 2019), it is possible to calculate the internal index measures specified in Section 3.8. The value of each internal index measure returned by the function is displayed on logarithmic scale of base 10.

However, in validating the cluster number to be formed (k) by a given clustering model, the values obtained by each of the internal indices will be scaled so that the best value according to a distance measure will be 1 and the worst value of the same measure will be assigned 0. A Total Score is also calculated based on the sum of the scaled values of the different internal index measures obtained for the same model with k clusters. This total Score measure will also be scaled between 1 and 0 to validate the optimal number of clusters for a given model.

The same methodology applies when validating which cluster initialization is best for a given clustering model (this only applies to k-Means and k-Medoids clustering).

3.9 Principal component analysis (PCA)

Principal component analysis (PCA) is considered to be one of the most popular methods of feature extraction and dimensional reduction of a dataset. This method has been used in various areas such as image processing, machine learning and general exploratory data analysis. For a dataset composed of p dimensional variables in $\mathbb{R}^p$, PCA allows to calculate the orthognal projection in a dimensional subspace composed of the same number or less dimensions than the original space. The principal components that make up this subspace will be the ones that capture the largest variance in the dataset (Seghouane et al. 2019). In this method of linear orthognal transformation, the coordinate system is transformed so that the largest variance occurs over the first principal component direction, the second largest data variance occurs over the second principal component, and so on (Pandey et al. 2019).

The PCA is performed according to the following steps:

1. The dataset matrix X has the dimension $n \times p$, where n corresponds to the number of observations and p to the number of variables.
2. First, the empirical mean of the corresponding column is subtracted from all matrix values in order to obtain a mean-subtracted data matrix B:

$$B = [x_{i,j} - \mu_j] \tag{3.46}$$

Where $x_{i,j}$ and μ_j are an element of matrix X and the empirical mean of column j, respectively.

3. Then compute the covariance matrix of the matrix B:

$$C = \frac{1}{n-1} B^* B \tag{3.47}$$

In which $n-1$ and * are the Bessel's correction and the conjugate transpose operator which in case of $\mathbb{R}$ corresponds to regular transpose(^{T}), respectively.

4. The eigenvalues associated with matrix C are calculated:

$$det(C - \lambda I) = 0 \tag{3.48}$$

Where λ and I are the eigenvalues and the identity matrix, respectively.

5. Subsequently the eigenvectors are calculated:

$$C\overrightarrow{e_i} = \lambda_i \overrightarrow{e_i} \tag{3.49}$$

In which λ_i and $\overrightarrow{e_i}$ represent the eigenvalue of index i and the eigenvector corresponding to the eigenvalue of index i, respectively.

6. Divide each eigenvector obtained by its own norm:

$$\overrightarrow{e_i} = \frac{\overrightarrow{e_i}}{\| \overrightarrow{e_i} \|} \tag{3.50}$$

Where $\| \overrightarrow{e_i} \|$ is the eigenvector norm.

7. Project data according to a principal component:

$$\overrightarrow{a_i} = (\overrightarrow{a} - \overrightarrow{\mu}) \cdot \overrightarrow{e_i} \tag{3.51}$$

In which $\overrightarrow{a} - \overrightarrow{\mu}$ and $\overrightarrow{a_i}$ represent the centered observation and the projected observation according to the principal component i, respectively.

In Chapter 4, this method will be used to find the first 3 principal components that explain the greatest variability present in the dataset. These three principal components will compose a Euclidean space in which clusters formed will be projected in order to be able to characterize the clusters formed in each clustering algorithm tested.

3.10 Definition of clustering models and methodology application

In this section the clustering models, that will be analyzed in chapter 4, are defined taking into account the analysis methodology presented in section 3.2.

Figure 3.5 shows the various stages of model implementation and their integration into the methodology.

The first step will be the characterization and pre-processing of the initial dataset with outlier identification through Boxplots (see section 3.3). In this step a PCA will also be performed with characterization of the principal components obtained.

The next step details the 12 clustering models to be studied according to the components that define each model (Figure 3.5):

- Clustering approach (see section 3.5);
- Distance measure (see section 3.6);
- Prototype (see section 3.7);
- Window of comparison (see section 3.6).

In the Internal index evaluation step of Figure 3.5, the internal index evaluation methods that will be applied to each clustering model typology are described (see section 3.8). In this step a set of iterations will be performed for each model to evaluate the optimal number of clusters ($k_{optimal}$), by varying the number of clusters at each iteration (k = 1 to 10) and evaluating the result of each one through the measurements of internal index.

After obtaining the model with optimal number of clusters ($k_{optimal}$), it is necessary to perform a new set of iterations to validate the best centroid initialization ($i_{optimal}$) for it. In this case the centroids are started randomly at each iteration (i = 1 to 20) and the results obtained through the internal index measurements are evaluated. The iteration that obtained the best result is chosen.

It should be noted that for the Hierarchical and Fuzzy Clustering models the second iterative process of evaluating centroid initialization is not applicable. In the case of Hierarchical Models, prototypes are not a necessary component for cluster formation (see section 3.5.1). In the case of Fuzzy Clustering, this approach does not initialize centroids but rather uses a membership matrix (U^0) representing the initial degree of belonging of each object to each cluster (see section 3.5.4).

In the PCA and centroids visualization step the best results for each clustering model are described through the observation of the centroids of the formed clusters that characterize the water demand profilling. By visualizing clusters according to the three principal components obtained by the PCA method (see section 3.9) the degree of separation of clusters formed by each clustering model is also described.

After characterizing the various clustering models, a selection of the ones that allow a better description of the dataset is performed. Taking these models into account, a combined model is created to assimilate the characteristics of them by combining the clusters.

In the last step the centroids of the combined model are characterized taking into account factors such as:

- Typology of the day (working days vs. weekends / holidays);
- Geographic data (region of the country);
- Dry months vs. wet months.

These analyzes allows to characterize the water demand profiles and the existence of anomalous behaviors that may lead to inefficient water use.

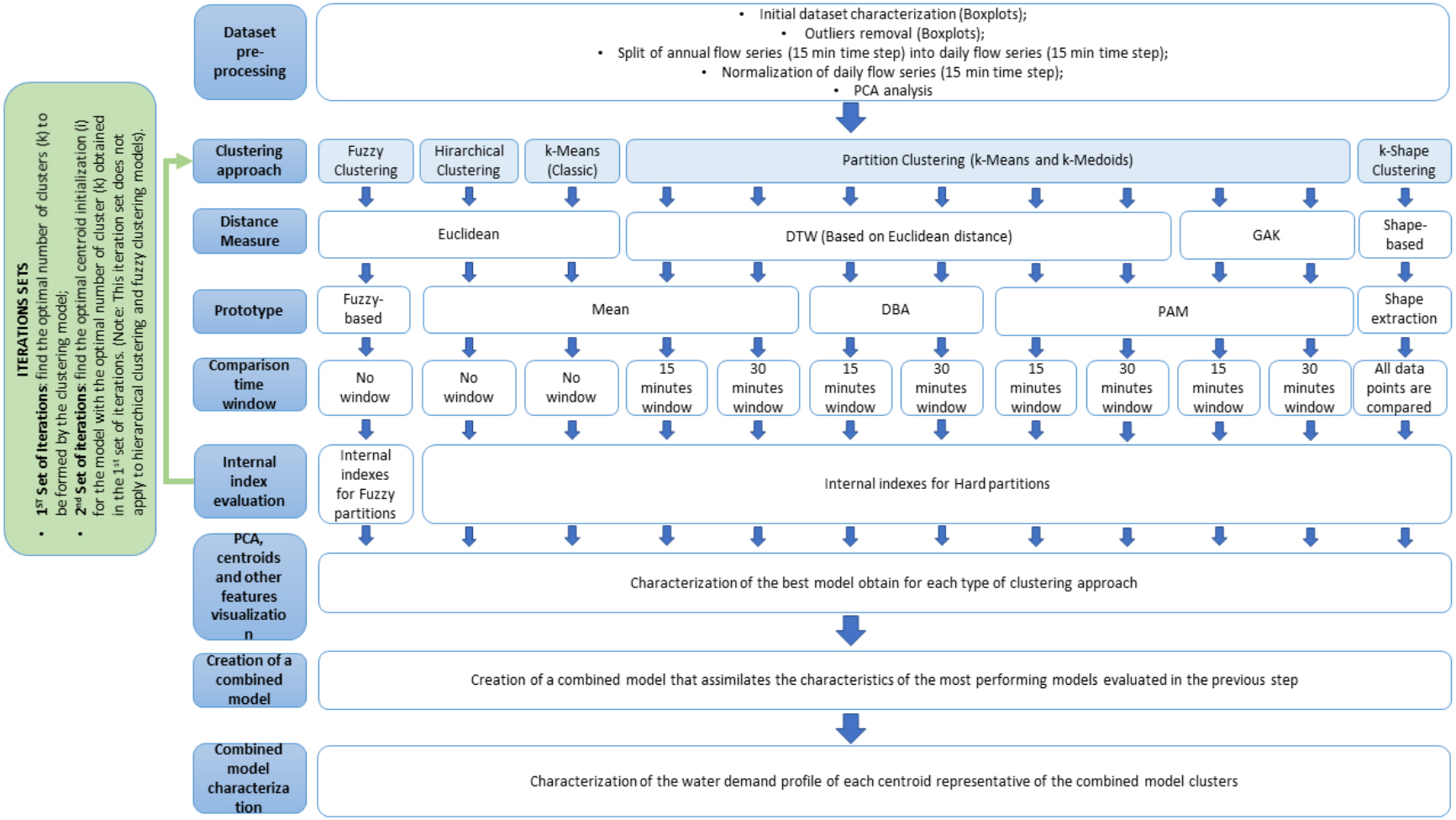

Figure 3.5: Clustering models definition and methodology application.

Chapter 4

Results and Discussion

4.1 Overview

Chapter 3 presented the theoretical foundation of clustering techniques, distance measurements, prototypes, internal index evaluation measures, PCA analysis, normalization and outlier removal methods. In the present chapter will be applied the techniques to the dataset in order to characterize it and extract knowledge about the various identifiable behaviors in daily mean flow patterns.

The present chapter is organized as follows:

- **4.2 Data characterization, preprocessing and PCA analysis**: initial data characterization, preprocessing, outlier removal and PCA analysis;
- **4.3 Application of clustering models with inelastic distance measures**: this section presents the method for analyzing inelastic clustering models. The application, evaluation and characterization of clustering models with inelastic distance measures is presented in Appendix A;
- **4.4 Application of clustering models with elastic distance measures**: this section presents the method for analyzing elastic clustering models. The application, evaluation and characterization of the most performing clustering models with elastic distance measures is presented in sections 4.4.1, 4.4.2 and 4.4.3. The remaining models are presented in Appendix B;
- **4.5 Summary of clustering models analysis**: comparison of clustering models and selection of the most performing models;
- **4.7 Combined model analysis**: evaluation and characterization of a combined model consisting of a combination of clusters of the most performing models.

4.2 Data characterization, preprocessing and PCA analysis

In this section the characterization of the initial data and data preprocessing operations will be performed in order to prepare the data for principal component analysis and clustering operations. It will also be described the PCA analysis and the representativeness of the principal components.

4.2.1 Raw dataset characterization

The dataset used in the analyzes of this chapter is a 15-minute average flow rate (m^3/h) series collected over a year. Each series corresponds to a year-round data collection from a District Metered Area (DMA).

Figure 4.1 indicates the geographical location of each of the series:

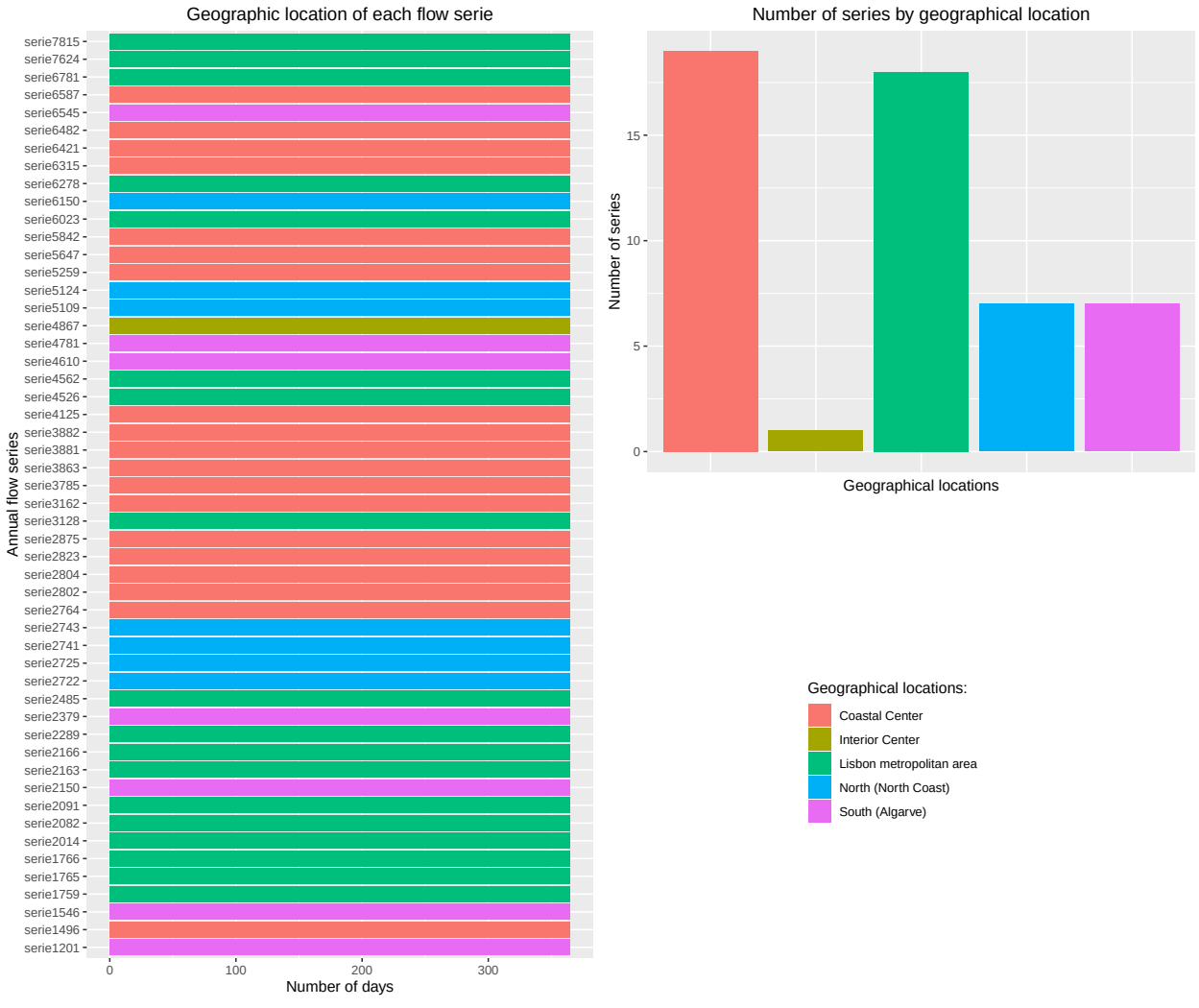

Figure 4.1: Geographical location.

The dataset presents a total of 52 series from which it can be seen that most of the series belong to the Coastal Center region, with 19 series, and the Lisbon Metropolitan Area, with 18 series. The North and South regions have 7 series each. The Interior Center region is the least represented with only 1 series.

Figure 4.2 shows the mean flow values (m^3/h) with a 15-minute time step for the months of January, February, June July, November, and December of the 1759 series.

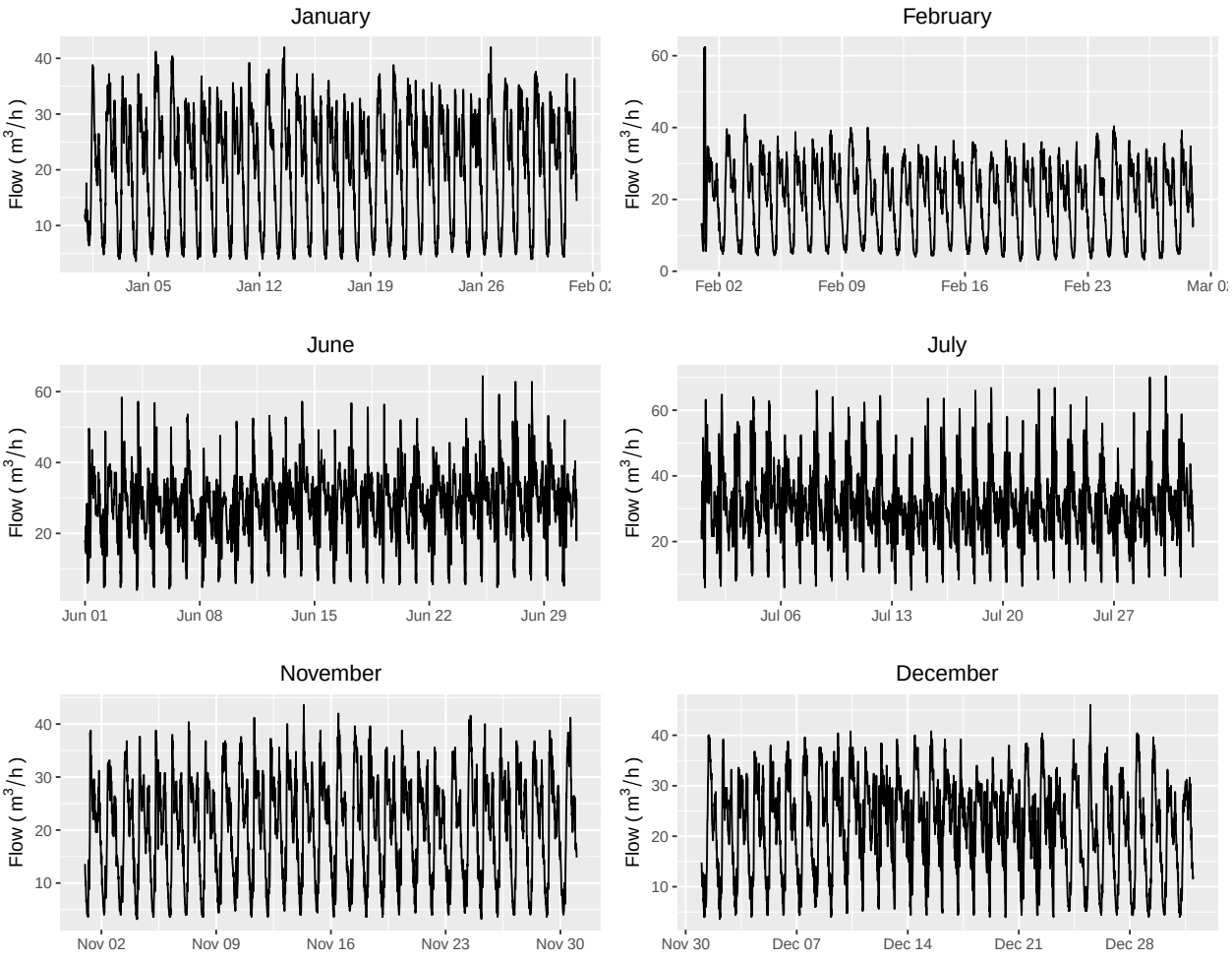

Figure 4.2: Series 1759 flow data.

4.2.2 Statistical characterization of the dataset

Graphical view of the median of the annual flow series

Figure 4.3 shows the median flow rate (m^3/h) of the annual series. This analysis allows to verify the order of magnitude of the flow values.

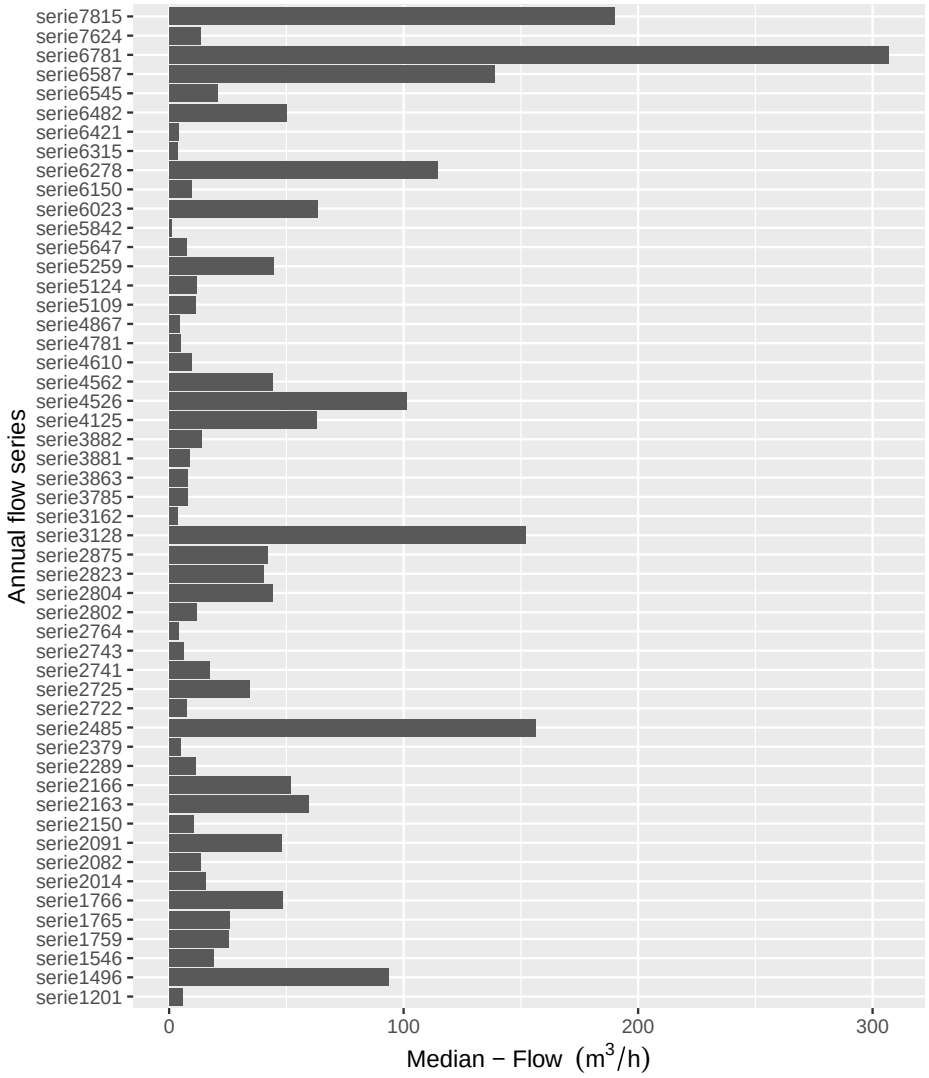

Figure 4.3: Meadin flow of each annual series.

Through the analysis of the median of the flow values (Figure 4.3), it is verified that most series have a median flow value of less than 50 m^3/h, with only the existence of 13 series with values of median flow greater than 50 m^3/h.

Boxplot analysis

This analysis allows us to visualize the main statistical characteristics associated with each series, as defined in section 3.3.

Figure 4.4 shows the Boxplots for the series with a median flow rate of less than 50 m^3/h.

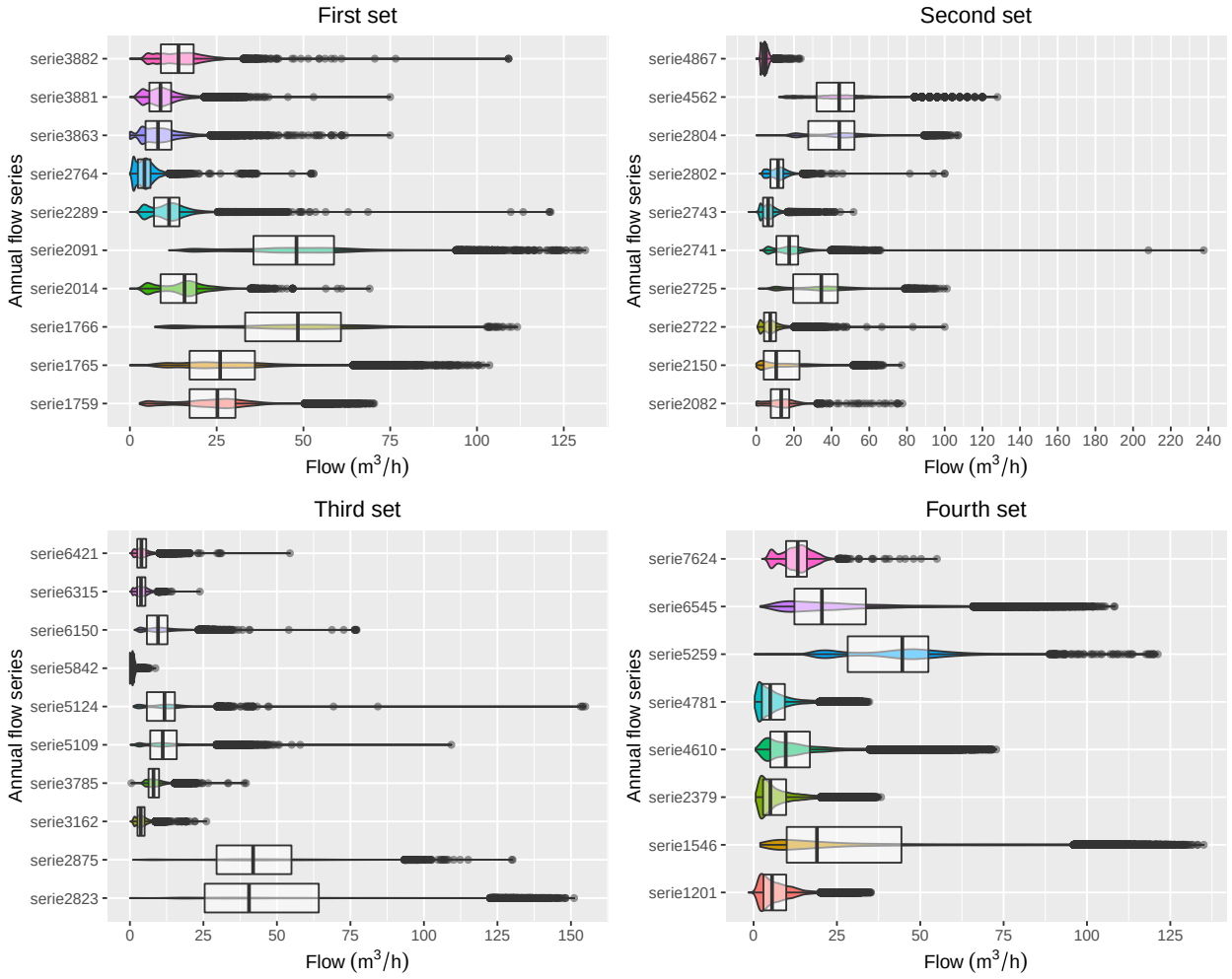

Figure 4.4: Boxplot of Series with median flow rate of less than 50 m^3/h.

Through the analysis of Figure 4.4, the following set of observations was verified:

- **1^{st} set**: this set of series there are outliers present in all series. The series with a median of more than 25 m^3/h have distances between quartiles higher than the others, indicating a greater variability of flow values throughout the year. Flow series less than 25 m^3/h are more compact and denser in the area between quartiles;
- **2^{nd} set**: the same characteristics are observed in this set. The 2741 series stands out by presenting outliers much higher than the distance between quartiles. In the series 2743 the presence of negative flow values is verified;
- **3^{rd} set**: this set also verifies the same patterns evidenced in the previous sets. The 5124 series has a medium flow rate of 12.5 m^3/h and a maximum flow rate greater than 150 m^3/h. The 6545, 5109, 2875 and 2823 series also have outliers greater than 100 m^3/h and longer distances between quartiles;
- **4^{th} set**: This set verifies the same patterns evidenced in the previous sets. Most series have a median of 12.5 m^3/h or less, with the exception of series 5259 and 1546 which

present higher median values, greater distances between quartiles and values of outliers greater than 100 m^3/h. In the series 1201 the presence of negative flow values is verified.

Figure 4.5 shows the Boxplots for the series with a median flow rate of over 50 m^3/h.

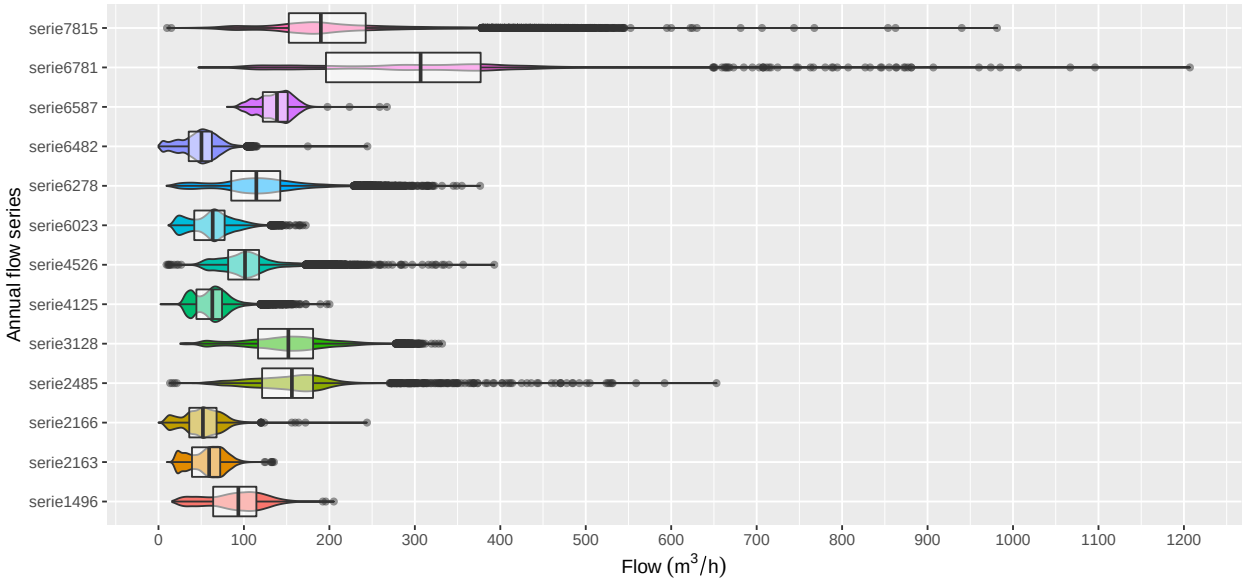

Figure 4.5: Boxplot of Series with median flow rate of over 50 m^3/h.

This set of series is characterized by having a median of more than 50 m^3/h. It is observed that tendentially the series with higher median value also present larger distances between quartiles. The 7815 and 6781 series are characterized by their maximum flow values higher than 950 m^3/h and also the series 2485 stands out for having a maximum flow rate of more than 600 m^3/h.

4.2.3 Dataset preprocessing for clustering operations

In order to prepare the dataset for clustering operations, the following set of steps must be performed:

- **Split operation**: clustering and PCA analysis will focus on the formation of groups based on the forms of daily flow series. It is necessary to split the annual flow series into daily flow series;
- **Outliers removal**: based on the analysis performed in section 4.2.2, daily series with Outliers present will be removed;
- **Normalization**: normalization of the flow values will be performed taking into account the Z-normalization method described in section 3.4.

Figure 4.6 presents the preprocessing steps applied to the 1759 series:

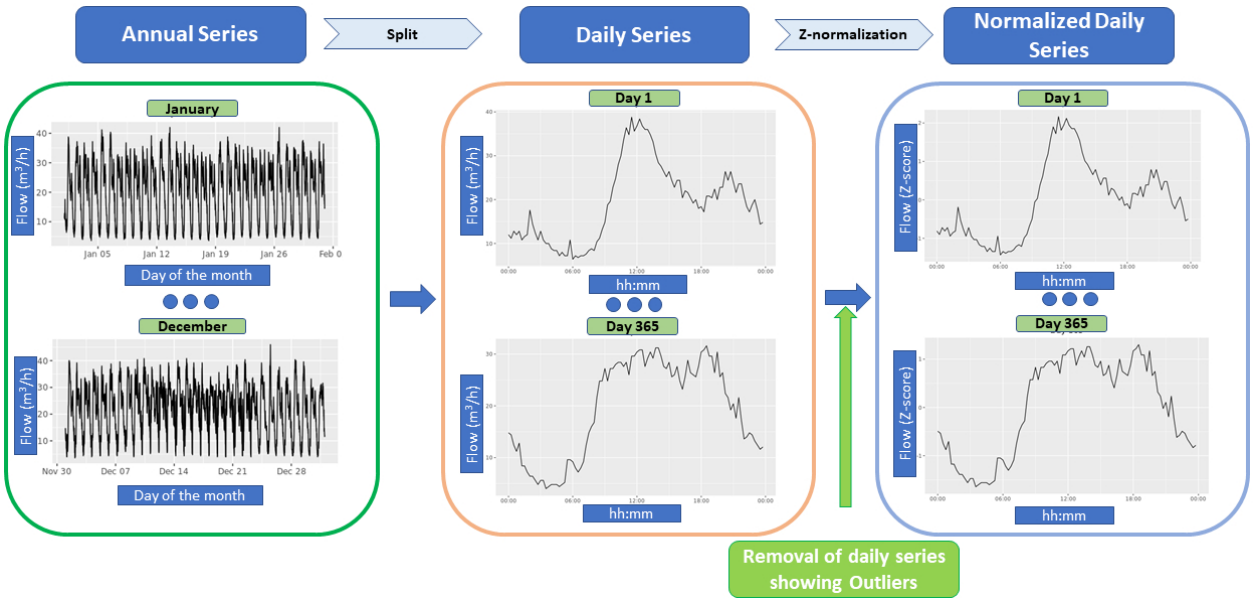

Figure 4.6: Series 1759 pre-processing.

The methodology of Figure 4.6 was applied to all annual series.

Regarding the Outliers removal process, after analyzing the results obtained in Section 4.2.2 with the Boxplot method, it was verified that for almost all series there are flow values above 1.5× IQR and these cases correspond to atypical values. It was decided to maintain these flow values in order to assess how cluster analysis can be used to differentiate atypical behaviors. In the case of negative values (correspond to inversions in the flow direction), since the focus of the analysis is on the study of the flow rate provided to network sectors, these values should be removed.

Table 4.1 identifies the daily flow series with negative flow values:

Table 4.1: Daily flow series with negative flow values.

Series ID	Month	Day	Minimum Flow (m^3/h)	Number of negative flow values	Proportion of negative flow values (%)
serie1201	10	27	-1.33	20	20.83
serie1201	10	28	-1.55	12	12.50
serie1201	11	3	-1.23	12	12.50
serie1201	11	4	-0.74	12	12.50
serie2743	8	31	-4.18	68	70.83
serie2743	9	1	-0.82	1	1.04

From Table 4.1 it is verified that 6 daily series have negative flow values. These series were removed from the datasets and not taken into account in the analysis performed in the next sections.

After the removal of daily series with negative flow values, the normalization process was performed according to the Z-normalization method (see section 3.4). At the end of this operation the Dataset is pre-processed. However the dataset has to be organized in a tidy way in order to enable operations of Principal component analysis and Clustering in later chapters.

The Dataset was organized as follows:

- **Variables**: each variable corresponds to a time instant of 15 minutes. Making a total of 96 variables that make up a day;
- **Observations**: each observation corresponds to a day of the year in a DMA.

Figure 4.7 shows how the Dataset was organized for PCA and clustering operations:

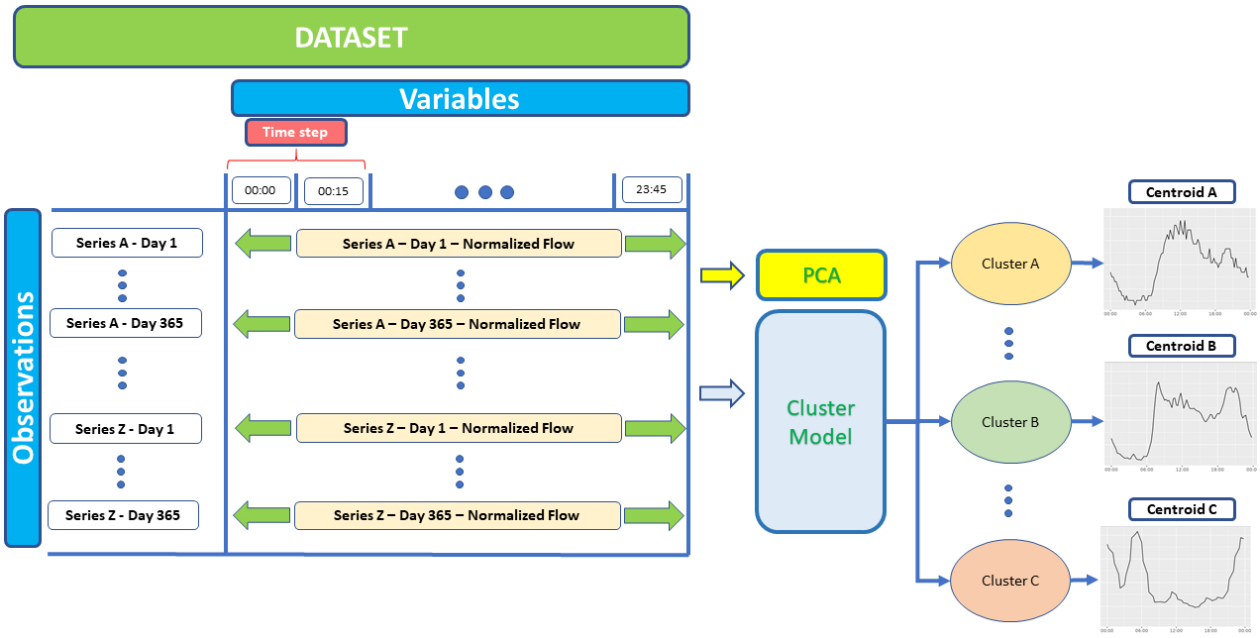

Figure 4.7: Dataset organization for PCA and Clustering Operations.

4.2.4 Principal component analysis

As described in section 3.9, the principal component analysis method allows to perform an orthogonal transformation to find the set of directions in space that allow you to describe the greatest variability of the data, thus allowing a dimensional reduction. This technique is particularly important for the visualization of the formed clusters and to evaluate the degree of separation of them according to dimensional space composed by 3 dimensions that explain the variability present in the dataset.

After applying the method and obtaining the principal components, it is necessary to evaluate the variance that is explainable by each of the components.

Figure 4.8 shows the variance explained by the principal components:

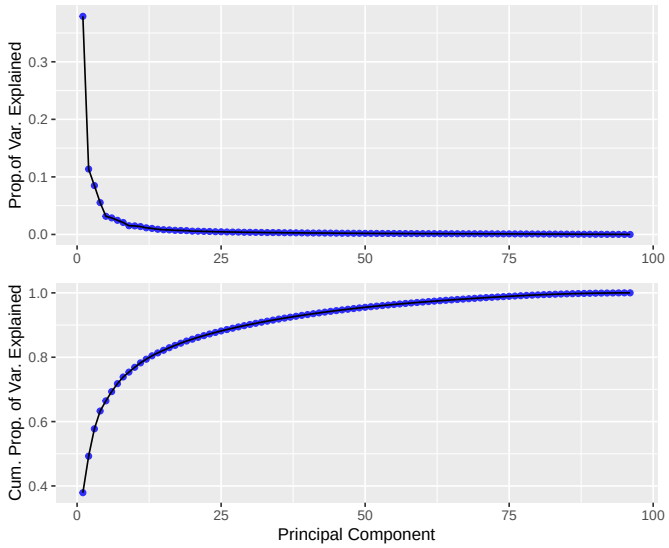

Figure 4.8: PCA - Variance explained by the principal components.

From Figure 4.8 it can be seen that the first 2 principal components explain about **50%** of the variability present in the data. The set of the first 3 principal components can explain about **60%**, and with the first 25 principal components it is possible explain about **90%** of the variability of the data. Finally, the set of the first 50 main components explains about **98%** of the variability of the model data.

In the next sections clustering models are applied. The clusters formed will be graphically represented according to the first 3 principal components. It should be noted that these graphs only allow to verify the separation of the clusters formed in a dimensional space that represents only **60%** of the variability of the data.

Figure 4.9 shows the weights assigned by each of the first 3 principal components at each time point:

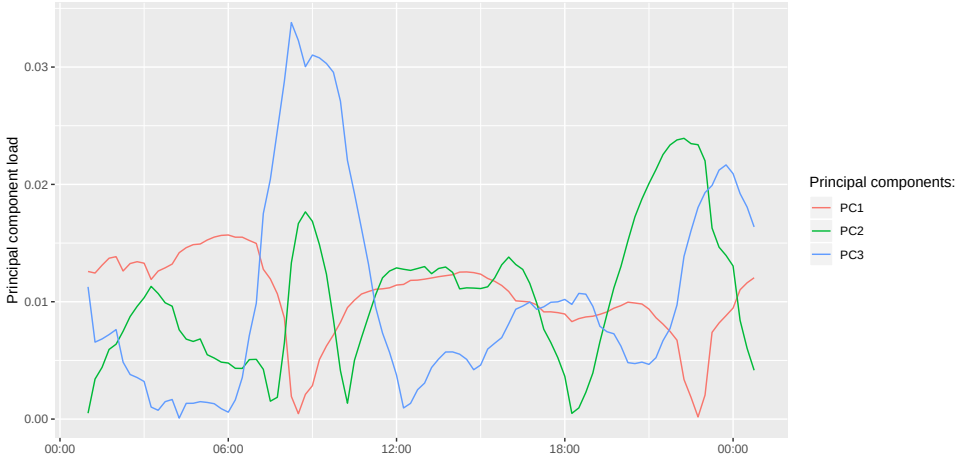

Figure 4.9: PCA - Principal Component Loads.

Through Figure 4.9 it is possible to conclude that:

- **Principal component 1 (PC1)**: uniformly represents all time periods, except for time periods near 07:30 and time periods near 21:30;
- **Principal component 2 (PC2)**: has a non-uniform distribution over all time points. This principal component is more representative of the time periods near 7:30 and the time periods near 21:30 (with maximum weight around 21:00);
- **Principal component 3 (PC3)**: similar to principal component 2, has a non-uniform distribution over all time points. However, it has a higher representation of time periods close to 7:30 than principal component 2, and a lower representation of time periods near 21:30 (with maximum weight close to 22:30).

Through this analysis it can be seen that PC1 is complementary to PC2 and PC3, since in the time periods where PC1 has little representation, it corresponds to zones where PC2 and PC3 components are more represented. Regarding the comparison between PC2 and PC3, it is noteworthy that the maximums in the zones with greater representativeness during the night period do not coincide. There is also a distinct behavior around 18:00, where PC2 has a minimum weight, while PC3 has a local maximum.

4.3 Application of clustering models with inelastic distance measures

In this section we will perform clustering operations according to the algorithms:

- K-means (hard partitioning);
- Hierarchical (hard partitioning);
- Fuzzy (soft partitioning).

The measure of distance used will be Euclidean in an inelastic way, that is, for the calculation of distances only the values of flow belong to the same time instant in the various daily patterns that make up the dataset will be compared.

Figure 4.10 shows the evaluation and characterization procedures of the clustering models with inelastic distance measures.

In **Appendix A** the following models are evaluated and characterized:

- **A.1**: K-means Clustering;
- **A.2**: Hierarchical Clustering;
- **A.3**: Fuzzy Clustering.

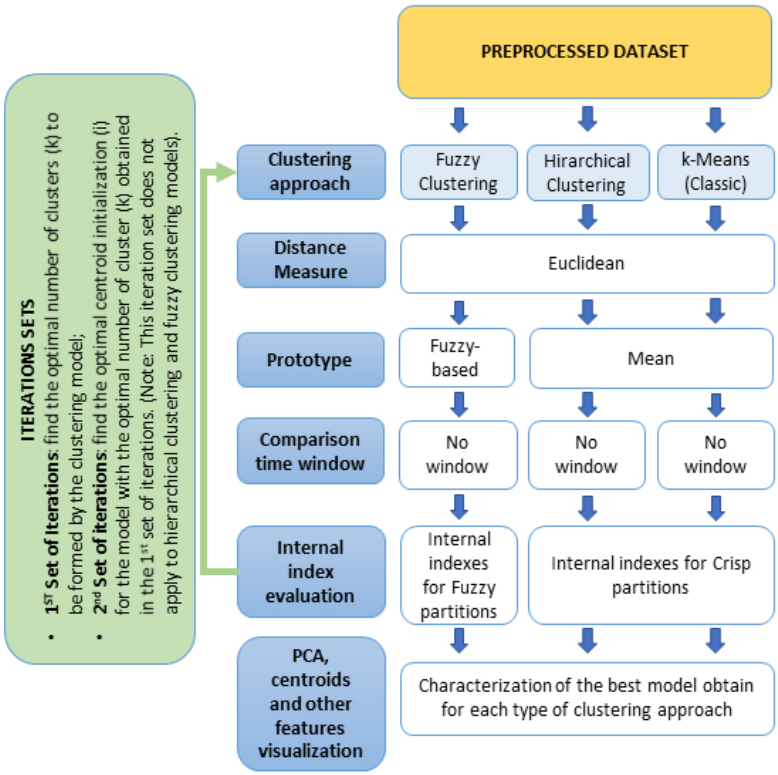

Figure 4.10: Characterization and evaluation workflow of cluster models with inelastic distance measurements.

4.4 Application of clustering models with elastic distance measures

In this section we will perform clustering operations according to the algorithms with elastic distance measurements that allow the comparison of flow values belonging to different time periods.

The following clustering approaches will be used:

- Partitional clustering (k-Means and k-Meadois);
- k-Shape clustering.

In order to compare the flow values at different time intervals between time series, the following elastic distance measurements were used:

- Dynamic time warping distance with window constrains;
- Global alignment kernel distance with window constrains;
- Shape-based distance.

In order to visualize the characteristics of each formed cluster, the following alternatives of representation of prototypes were used:

- Mean;

- Partition around medoids;
- DTW barycenter averaging;
- Shape Extraction (only when Shape-based distance is used).

Figure 4.11 shows the evaluation and characterization procedures of the clustering models evaluated in this section:

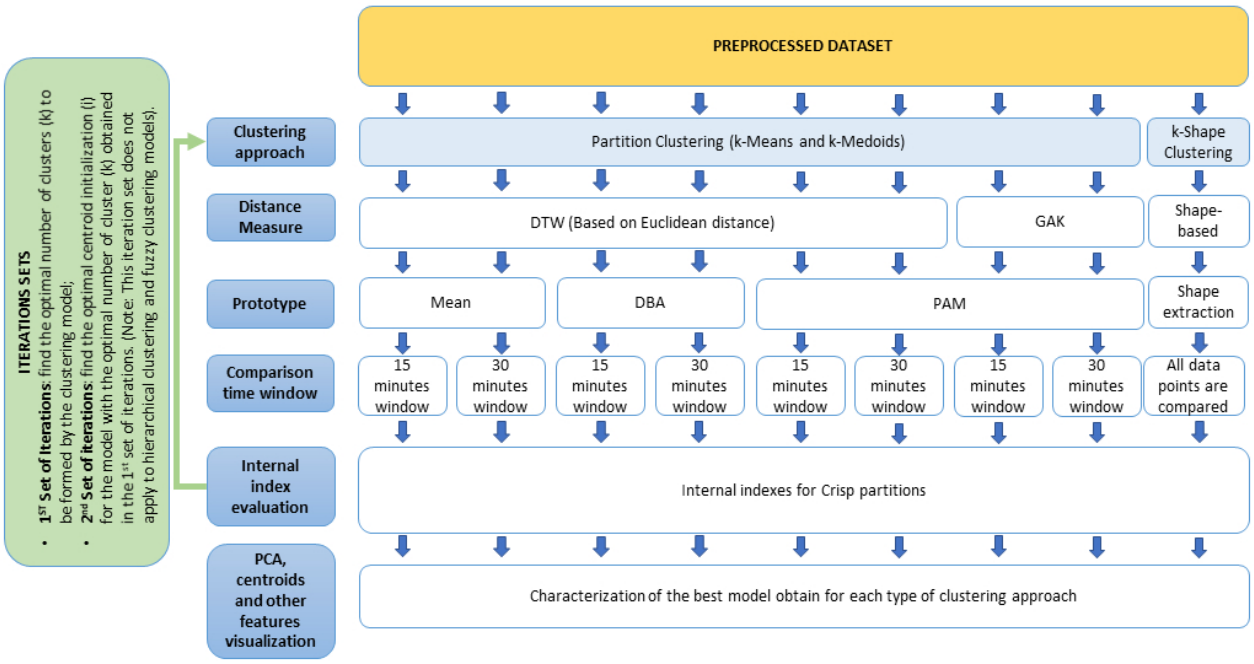

Figure 4.11: Characterization and evaluation workflow of cluster models with elastic distance measurements.

In **Appendix B** the following models are evaluated and characterized:

- **B.1**: Partitional Clustering with DTW, Mean prototype and 15 minutes time window;
- **B.2**: Partitional Clustering with DTW, Mean prototype and 30 minutes time window;
- **B.3**: Partitional Clustering with DTW, PAM prototype and 30 minutes time window;
- **B.4**: Partitional Clustering with DTW, DBA prototype and 15 minutes time window;
- **B.5**: Partitional Clustering with DTW, DBA prototype and 30 minutes time window;
- **B.6**: Partitional Clustering with GAK, PAM prototype and 30 minutes time window.

In the **next sections** will be presented the models that from the point of view of knowledge extraction allowed to obtain more information about the behaviors existing in the daily patterns present in the dataset. The **sections** are organized as follows:

- **4.4.1**: Partitional Clustering with DTW, PAM prototype and 15 minutes time window;
- **4.4.2**: Partitional Clustering with GAK, PAM prototype and 15 minutes time window;
- **4.4.3**: K-shape Clustering.

4.4.1 Partitional Clustering with DTW, PAM prototype and 15 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 15 minutes (see section 3.6.2).

Clustering model internal index evaluation

Figure 4.12 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

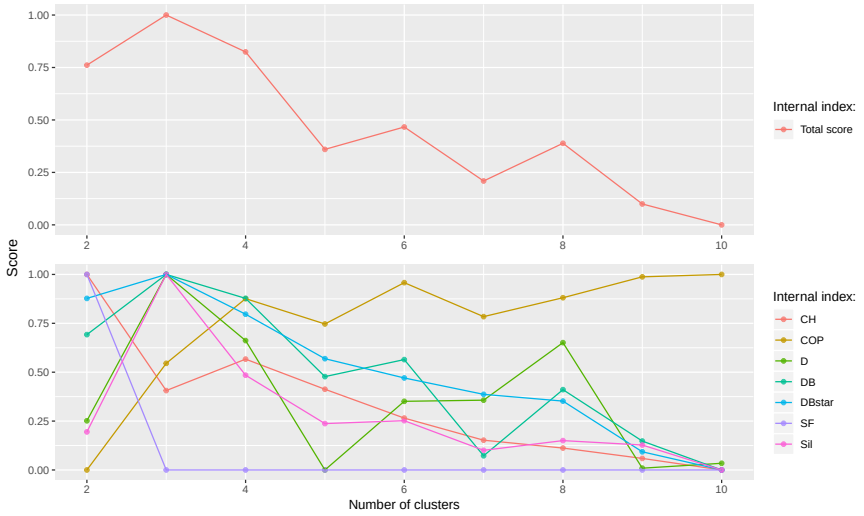

Figure 4.12: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, PAM Prototype and 15 minutes time window.

Figure 4.12 shows that the best result (Total score) was with the formation of 3 clusters. This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure 4.13 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

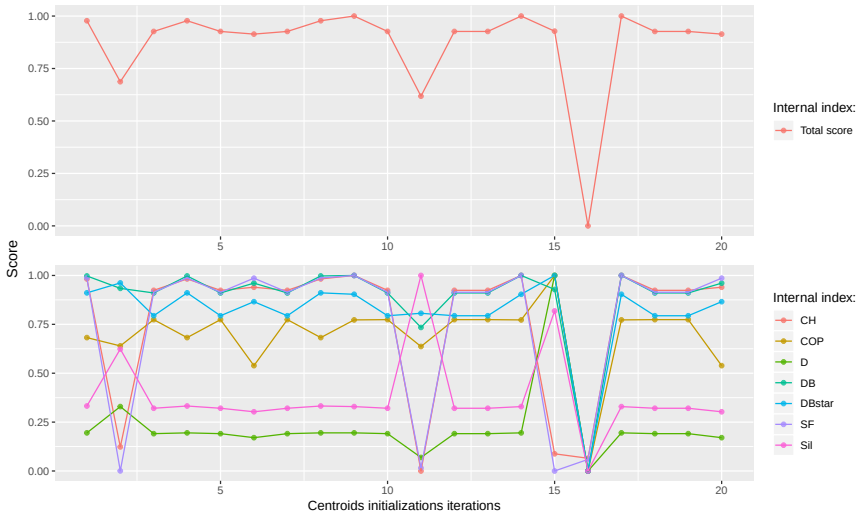

Figure 4.13: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, PAM Prototype and 15 minutes time window.

Figure 4.13 shows that the 1^{st}, 4^{th}, 9^{th}, 14^{th} and 17^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 1^{st} iteration clustering model with the formation of 3 clusters will be analyzed.

Clustering model characterization

Figure 4.14 shows the visualization of the clusters formed by the model according to the first 3 principal components. As can be seen from Figure 4.14, there is a distinction between cluster 3 and the group formed by clusters 1 and 2, except in zones close to the value of -5 in the first principal component.

For clusters 1 and 2, the projection under principal components 1 and 2 allows to distinguish between the two groups except in areas close to the value of 0 in the principal component 2. Observing clusters 1 and 2 according to the projection on the principal components 1 and 3 it is not possible to clearly distinguish between the two clusters.

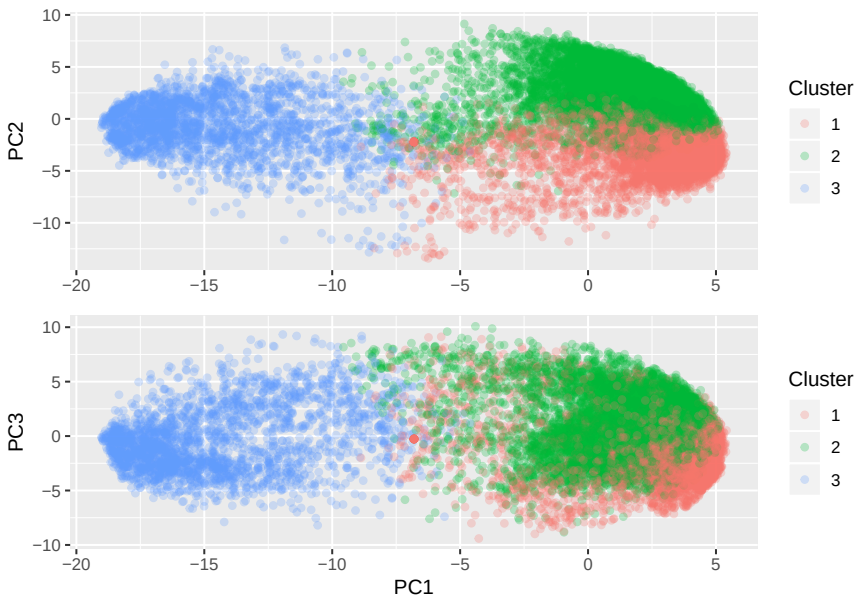

Figure 4.14: Clusters formed through the Partition Clustering model with DTW distance, PAM prototype and 15m window visualized through the 3 principal components of PCA.

Figure 4.15 shows the respective centroids of the clusters formed:

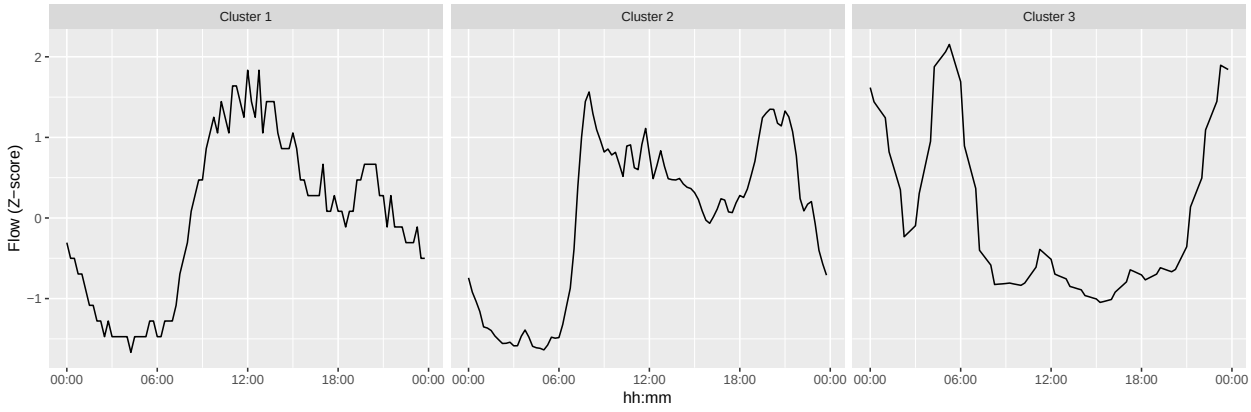

Figure 4.15: Partition Clustering model with DTW distance, PAM prototype and 15m window centroids.

Clusters 1 and 2 present higher consumption peaks during the day period, while Cluster 3 shows higher consumption during the night time period.

Cluster 1 shows the maximum consumption value at 12:00, a local minimum near 18:00 and a local maximum around 20:00. From this moment the consumption falls to the minimum value registered at 04:00. The described behavior represents a typical weekend period pattern, since the first peak of day consumption is only recorded near 12:00.

Cluster 2 has a maximum consumption peak near 08:00, another local maximum at 12:00 and reaches a local minimum around 16:00. From this period consumption increases again until around 20:00 which is a local maximum. Another local maximum is recorded around 21:00. After this period the consumption drops back down to 05:00 which corresponds to the minimum value of consumption. This behavior represents a typical pattern of a working day.

Cluster 3 shows peak consumption in the midnight period and in the period near 05:00 am. The predominance of this cluster by nocturnal consumption may be due to the use of water is predominantly associated with irrigation of gardens.

Figure 4.16 shows the size of each of the clusters formed:

Figure 4.16: Partition Clustering model with DTW distance, PAM prototype and 15m window clusters sizes.

The graphic shows that most of the patterns belong to Cluster 2 with 10354 dailly flow patterns, followed by Cluster 1 presents with 6175 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Cluster 3 has 2445 associated daily patterns that represent predominantly nocturnal consumption.

Figure 4.17 evaluates the degree of membership of each of the annual series to the formed clusters. It is observed that in all the annual series the daily patterns belong mostly to Clusters 1 and 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous clustering methods, since most clusters belong to a pattern with predominantly diurnal consumption.

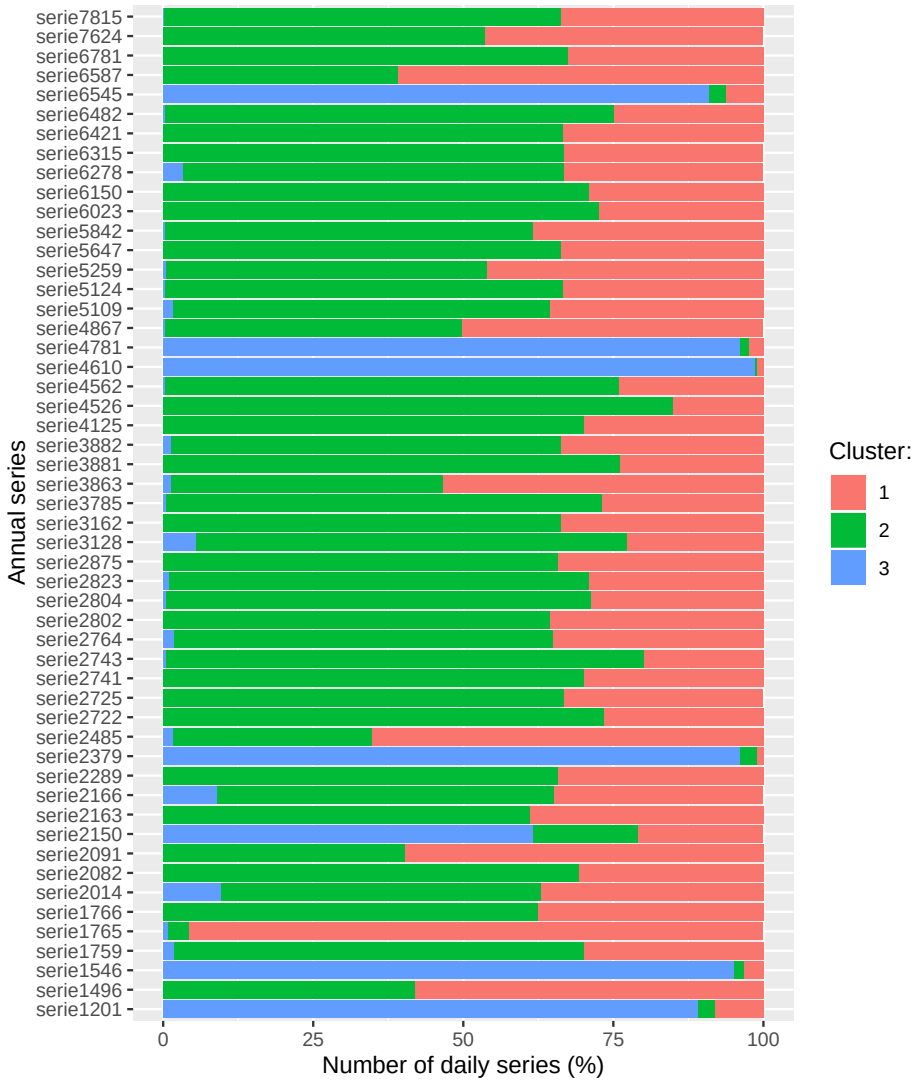

Figure 4.17: Partition Clustering model with DTW distance, PAM prototype and 15m window annual series membership.

Table 4.2 shows a set of statistical characteristics of the clusters formed:

Table 4.2: Partition Clustering model with DTW distance, PAM prototype and 15m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**	**Cluster 3 (m^3/h)**
Min.	0.00	0.00	0.00
1st Qu.	7.41	7.20	4.74
Median	20.28	18.47	10.43
Mean	46.40	45.95	19.24
3rd Qu.	58.90	57.69	22.21
Max.	1067.00	1207.00	530.50
IQR	51.49	50.49	17.47

Figure 4.18 identifies the influence of weekend or holiday days have on the formation of clusters. In Cluster 2 the percentage of weekend or national holiday patterns is around 12%, proving that this cluster is associated with typical working day behavior. In the case of cluster 1, the percentage of weekend or holiday patterns is around 70%, proving that this cluster is associated with typical weekend or holiday behavior. For Cluster 3 the percentage of weekends and holidays is around 30%. These values indicate that the formed cluster do not allow to identify a distinct behavior between a working day and a weekend or holiday.

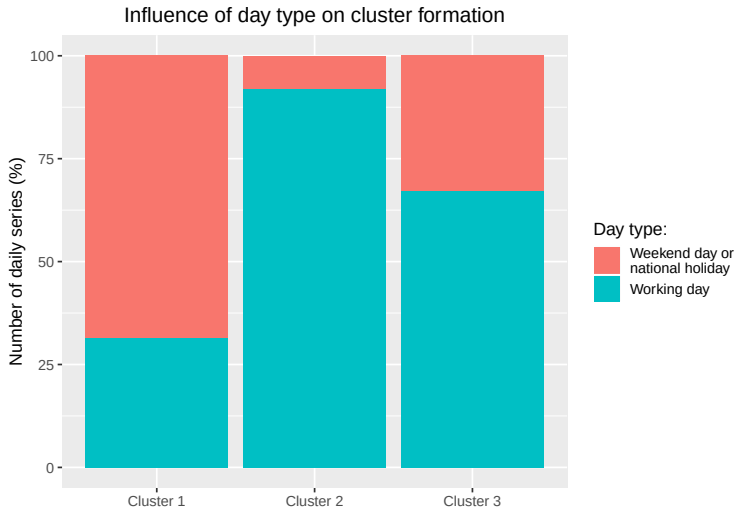

Figure 4.18: Partition Clustering model with DTW distance, PAM prototype and 15m window influence of day typology on the formation of clusters.

Figure 4.19 allows identifying the influence of day typology in each annual series by cluster type. As can be seen in Cluster 2, the annual series show mostly a higher percentage of daily patterns in working days. In the case of Cluster 1 for most of the annual series, there are a greater number of daily patterns of weekend day or national holiday type. Except for the annual series 6587, 5259, 3863, 1765 and 1496 which present a greater number of daily patterns associated to the working days. The annual series 6545, 4781, 4610, 2379, 2150, 1546, and 1201 also exhibit a greater number of daily patterns associated with working days, but are poorly represented in Cluster 1. Cluster 3, which represents daily patterns with higher nocturnal consumption, shows that in the annual series in which this cluster is the most representative, the proportions between working days and national holiday / weekend are indicative that typology of the day does not have significant influence on this cluster.

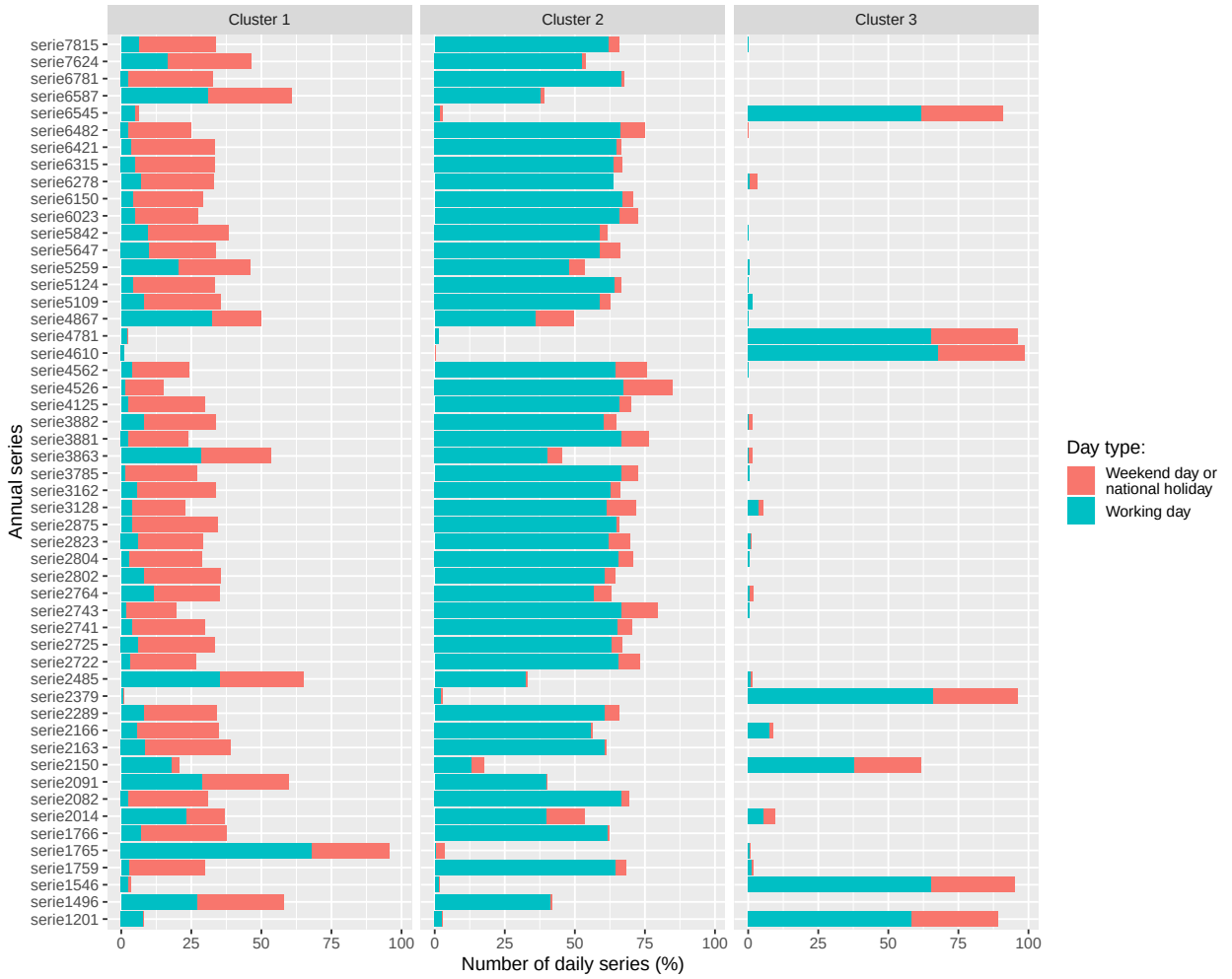

Figure 4.19: Partition Clustering model with DTW distance, PAM prototype and 15m window influence of day typology on each series by clusters.

4.4.2 Partitional Clustering with GAK, PAM prototype and 15 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: GAK (see section 3.6.3);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 15 minutes (see section 3.6.2).

Clustering model internal index evaluation

Figure 4.20 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

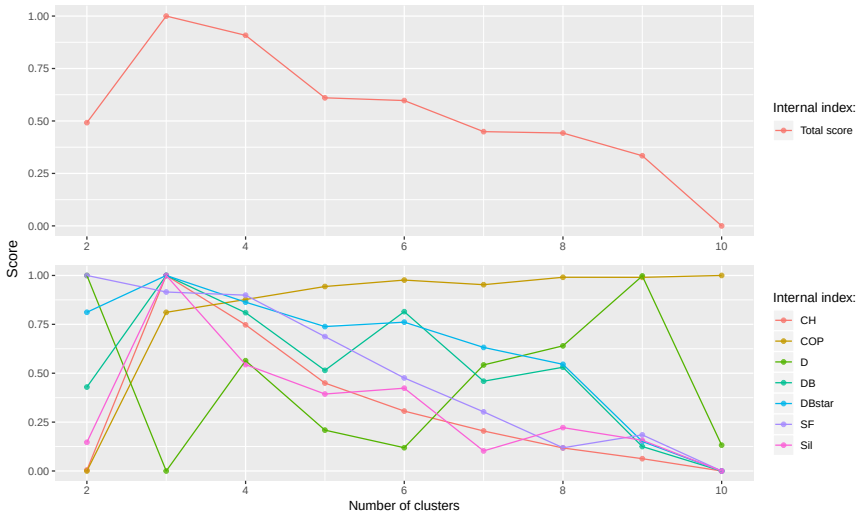

Figure 4.20: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with GAK, PAM Prototype and 15 minutes time window.

Figure 4.20 shows that the best result (Total score) was with the formation of 3 clusters. This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure 4.21 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

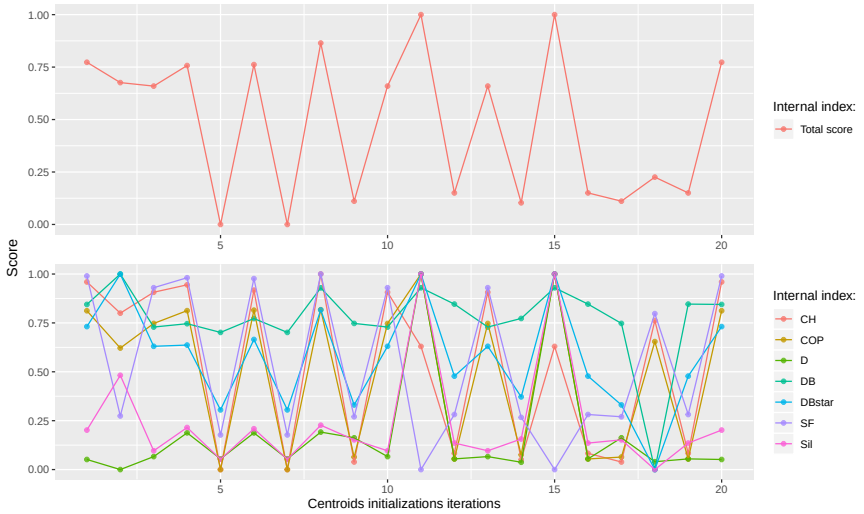

Figure 4.21: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with GAK, PAM Prototype and 15 minutes time window.

Figure 4.21 shows that the 11^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 15^{th} iteration clustering model with the formation of 3 clusters will be analyzed.

Clustering model characterization

Figure 4.22 shows the visualization of the clusters formed by the model according to the first 3 principal components:

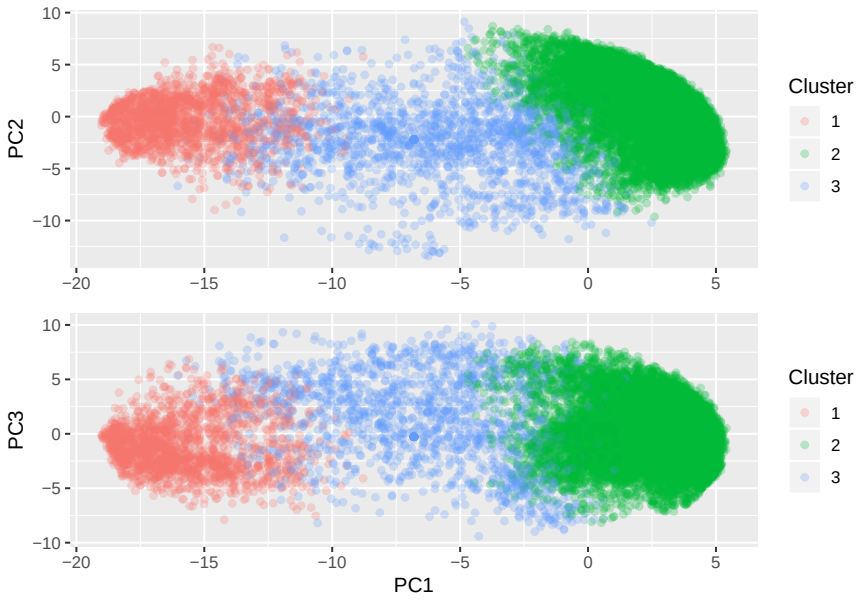

Figure 4.22: Clusters formed through the Partition Clustering model with GAK distance, PAM prototype and 15m window visualized through the 3 principal components of PCA.

In Figure 4.22 it is possible to see a separation of the clusters, being that Cluster 1 tends to be located tendentially in zones of value inferior to -12 in the principal component 1, Cluster 2 is tended in zones of value superior to -2.5 of the principal component 1. Cluster 3 is located in the intermediate zone between clusters 1 and 2.

The results obtained with the formation of 3 clusters are quite different in the location of the clusters compared to the results of previously presented in clustering models with DTW distance and PAM centroid that formed 3 clusters.

Figure 4.23 shows the respective centroids of the clusters formed:

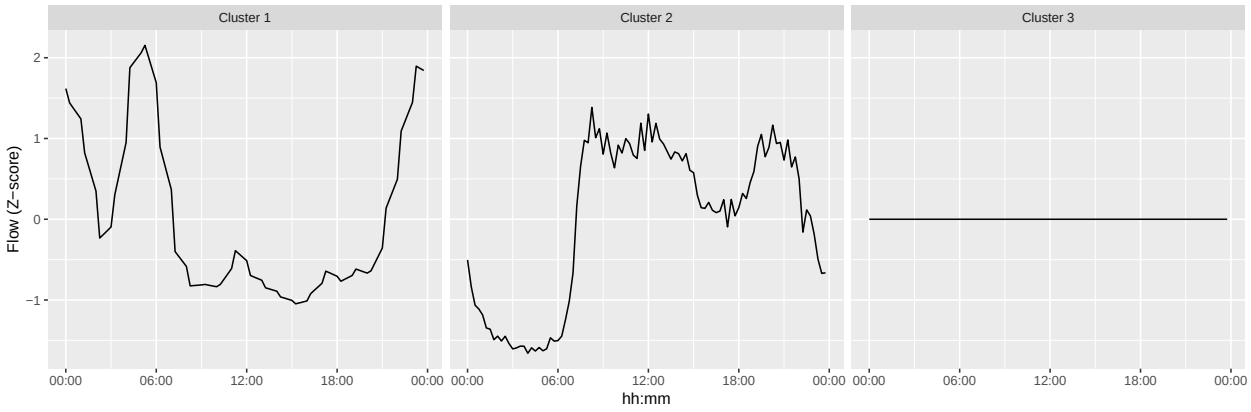

Figure 4.23: Partition Clustering model with GAK distance, PAM prototype and 15m window centroids.

Cluster 1 shows peak consumption in the 23:00 period and in the period near 05:00 am. The predominance of this cluster by nocturnal consumption may be due to the use of water is predominantly associated with irrigation of gardens.

Cluster 2 has a maximum consumption peak near 08:00, another local maximum at 12:00 and reaches a local minimum around 16:00. From this period consumption increases again until around 20:00 which is a local maximum. After this period the consumption drops back down to 05:00 which corresponds to the minimum value of consumption.

The centroid of cluster 3 has a constant flow rate throughout the day.

Clusters 2 present higher consumption peaks during the day period, while Cluster 1 shows higher consumption during the night time period. Cluster 3 identifies a group of daily patterns that exhibit a behavior of constant flow throughout the day.

Figure 4.24 shows the size of each of the clusters formed:

Figure 4.24: Partition Clustering model with DTW distance, PAM prototype and 15m window clusters sizes.

Figure 4.24 shows that most of the patterns belong to Cluster 2 with 15306 dailly flow patterns, followed by Cluster 1 presents with 1844 daily flow patterns. Cluster 3 presents 1824 daily flow patterns.

Figure 4.25 evaluates the degree of membership of each of the annual series to the formed clusters:

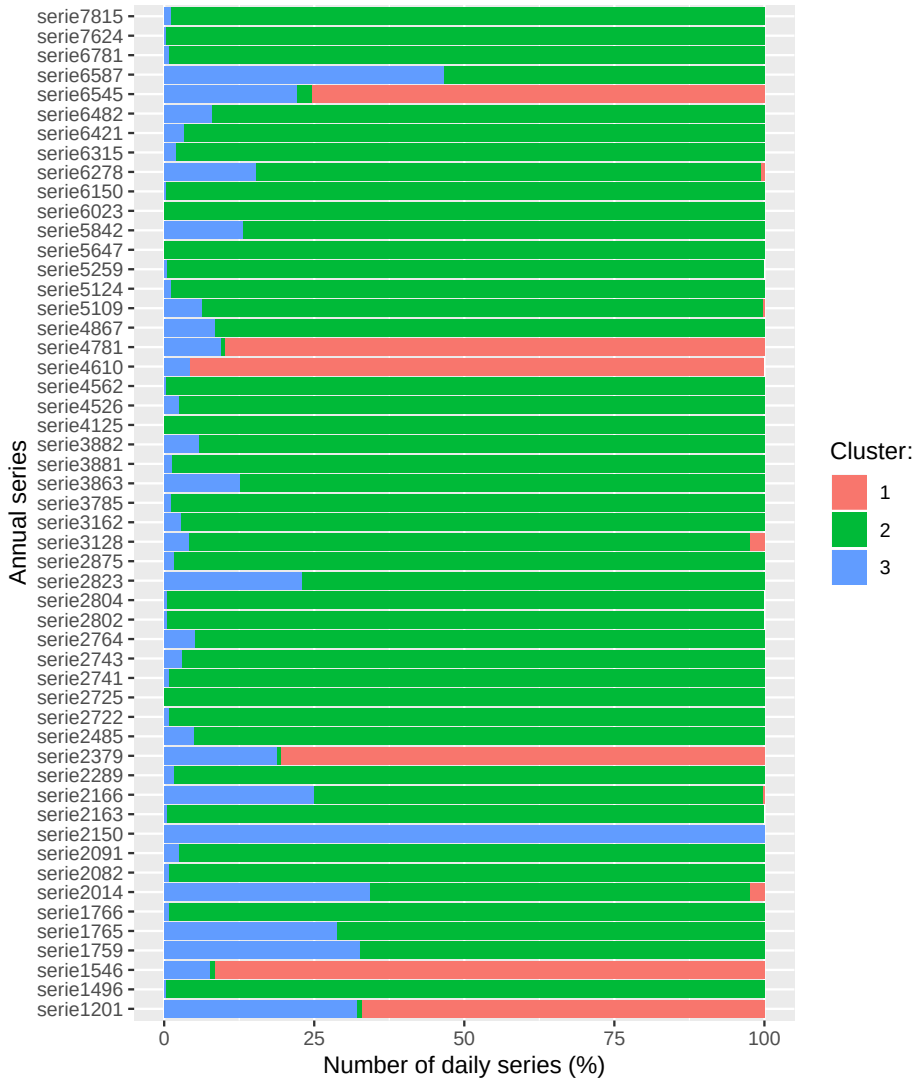

Figure 4.25: Partition Clustering model with DTW distance, PAM prototype and 15m window annual series membership.

It was observed that in all the annual series the daily patterns belong mostly to Clusters 2, indicated that most annual series present higher consumption during the daytime period. The exceptions are the series 6545, 4781, 4610, 2379, 1546 and 1201 that belong mostly to Cluster 1 and therefore show higher consumption during the night time.

The annual series 2150 belongs mainly to cluster 3. Other annual series such as 6587, 2166, 2014, 1765, 1759 and 1201 show a high percentage of daily patterns belonging to cluster 3.

Table 4.3 shows a set of statistical characteristics of the clusters formed:

Table 4.3: Partition Clustering model with GAK distance, PAM prototype and 15m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**	**Cluster 3 (m^3/h)**
Min.	0.00	0.00	0.00
1st Qu.	4.92	7.34	5.01
Median	10.37	18.82	18.01
Mean	18.18	46.15	38.09
3rd Qu.	21.60	57.76	47.20
Max.	312.25	1207.00	981.25
IQR	16.68	50.42	42.19

Figure 4.26 identifies the influence of weekend or holiday days have on the formation of clusters:

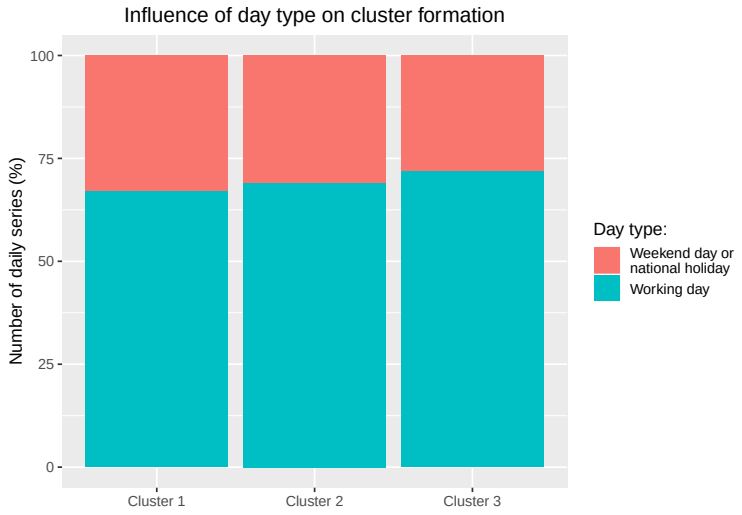

Figure 4.26: Partition Clustering model with GAK distance, PAM prototype and 15m window influence of day typology on the formation of clusters.

It is observed that the percentage of weekends and holidays for clusters is about 30%. This distribution indicates that these Clusters do not identify a distinct behavior between working day and weekend or holiday, since the assignment of the typology of days in a year is of the same order of magnitude.

Figure 4.27 allows identifying the influence of day typology in each annual series by cluster type:

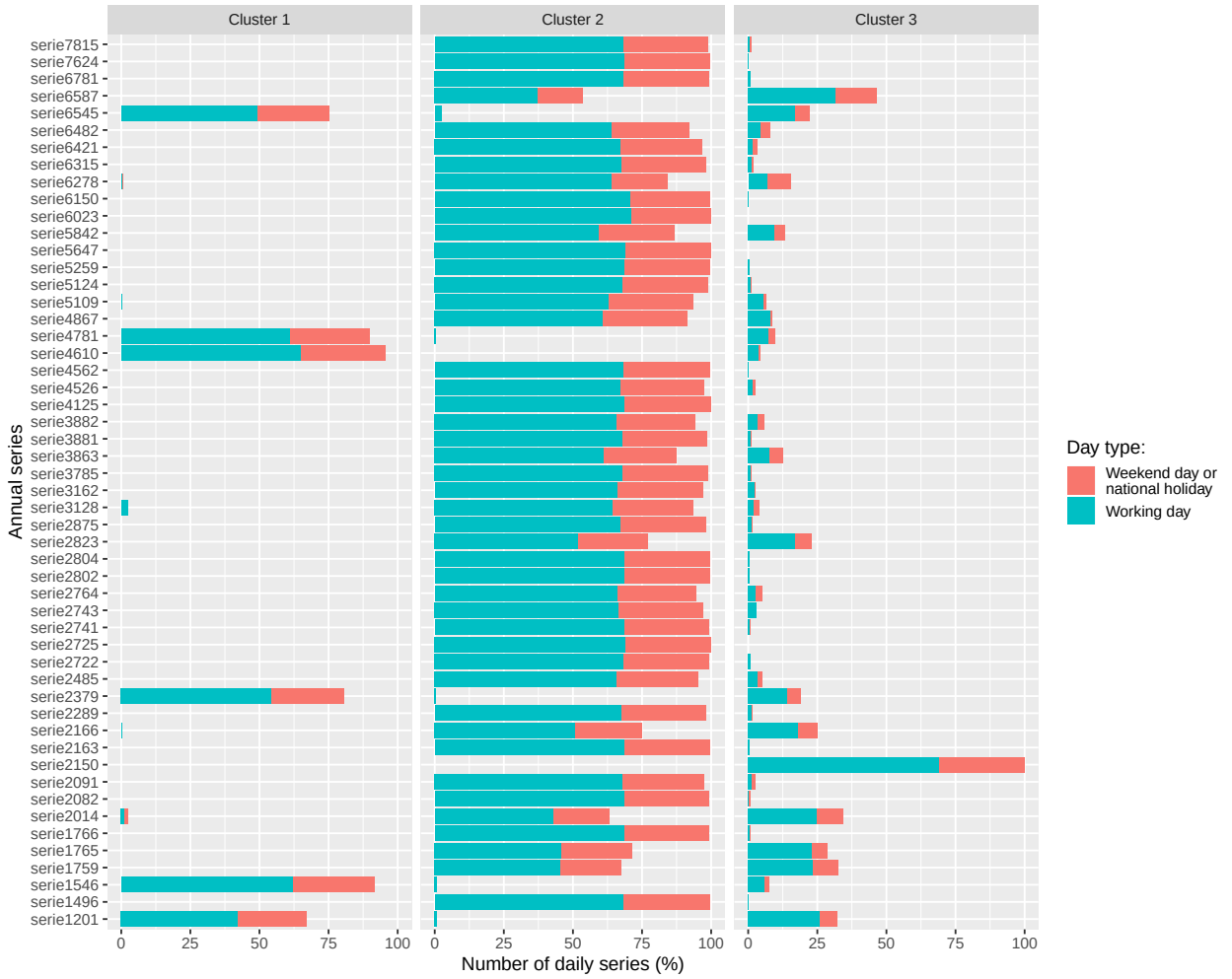

Figure 4.27: Partition Clustering model with GAK distance, PAM prototype and 15m window influence of day typology on each series by clusters.

As can be seen from Figure 4.27, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in the graph of the previous section, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day.

4.4.3 K-shape Clustering

In this section we will analyze a clustering model using the K-shape Clustering approach (see section 3.5.3) with the following components:

- Distance measure: Shape-based (see section 3.6.4);
- Prototype: Shape extraction (see section 3.7.4);
- Comparison time window: All data points are compared (see section 3.6.4).

Clustering model internal index evaluation

Figure 4.28 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

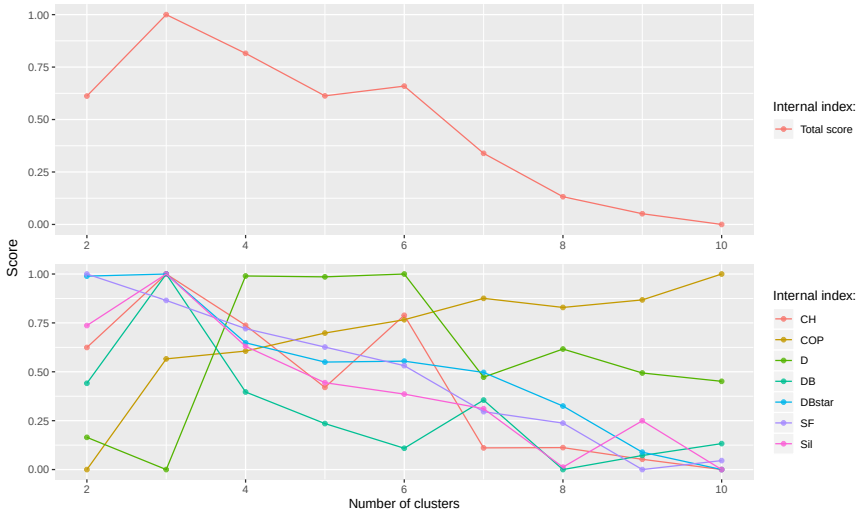

Figure 4.28: Internal index evaluation for 1^{st} iteration set of k-Shape.

Figure 4.28 shows that the best result (Total score) was with the formation of 3 clusters. This clustering approach needs to initially allocate centroids (see section 3.5.3), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures. Figure 4.29 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

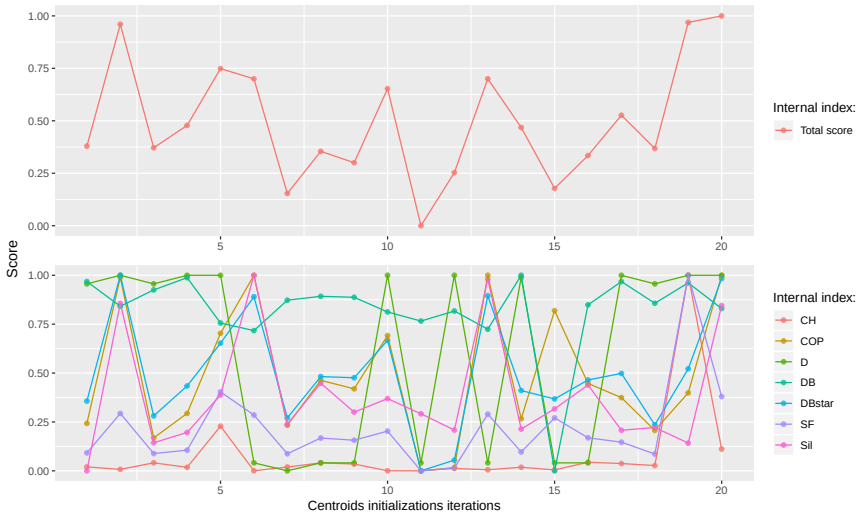

Figure 4.29: Internal index evaluation for 2^{nd} iteration set of k-Shape.

Figure 4.29 shows that the 20^{th} iteration provided the best performance in the internal indexes evaluation. In the next section the 20^{th} iteration clustering model with the formation of 3 clusters will be analyzed.

Clustering model characterization

Figure 4.30 shows the visualization of the clusters formed by the model according to the first 3 principal components:

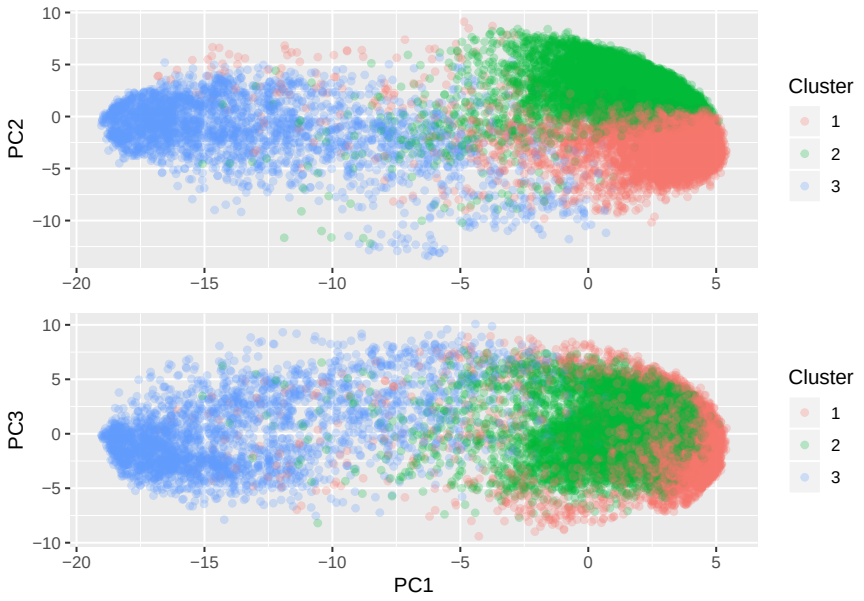

Figure 4.30: Clusters formed through the k-Shape model visualized through the 3 principal components of PCA.

As can be seen from Figure 4.30, there is a distinction between cluster 3 and the group formed by clusters 1 and 2, except in zones close to the value of -5 in the first principal component.

For clusters 1 and 2, the projection under principal components 1 and 2 allows to distinguish between the two groups except in areas close to the value of 0 in the principal component 2. Observing clusters 1 and 2 according to the projection on the principal components 1 and 3 it is not possible to clearly distinguish between the two clusters.

This results are very similar to the one obtained by the clustering partition model with DTW distance, PAM centroid and 15 minutes window.

Figure 4.31 shows the respective centroids of the clusters formed:

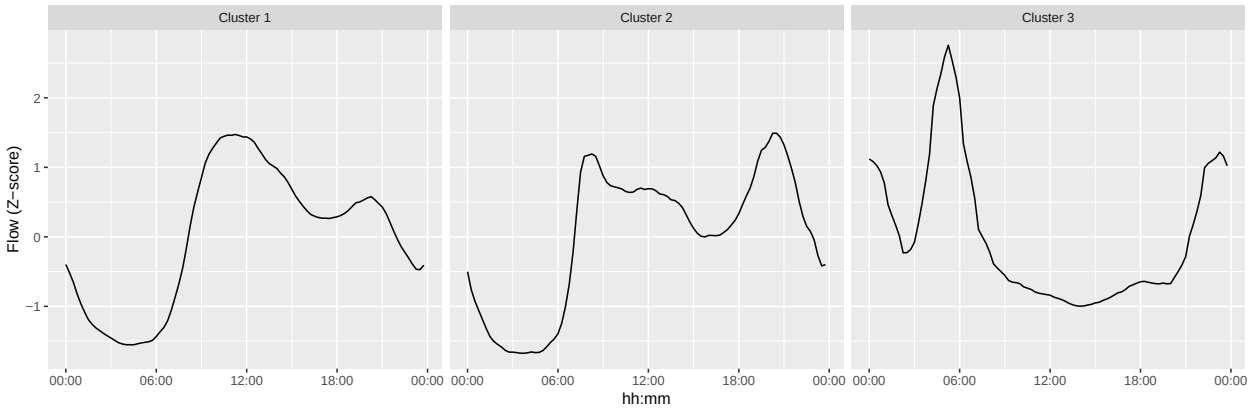

Figure 4.31: k-Shape model centroids.

Clusters 1 and 2 present higher consumption peaks during the day period, while Cluster 3 shows higher consumption during the night time period. Cluster 1 shows the maximum consumption value at 12:00, a local minimum near 17:00 and a local maximum around 20:00. From this moment the consumption falls to the minimum value registered at 04:00. The described behavior represents a typical weekend period pattern, since the first peak of day consumption is only recorded near 12:00. Cluster 2 has a maximum consumption peak near 07:30, another local maximum at 12:00 and reaches a local minimum around 16:30. From this period consumption increases again until around 20:00 which is a local maximum. After this period the consumption drops back down to 04:00 which corresponds to the minimum value of consumption. This behavior represents a typical pattern of a working day. Cluster 3 shows peak consumption at 23:00 and in the period near 05:00. The predominance of this cluster by nocturnal consumption may be due to the use of water is predominantly associated with irrigation of gardens.

This results are very similar to the one obtained by the clustering partition model with DTW distance, PAM centroid and 15 minutes window.

Figure 4.32 shows the size of each of the clusters formed:

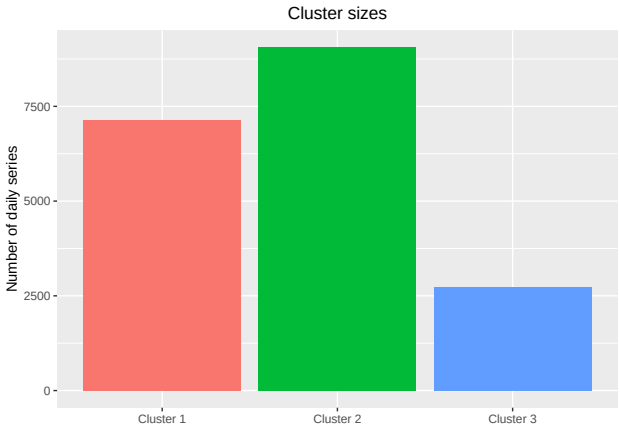

Figure 4.32: k-Shape model clusters sizes.

The graphic shows that most of the patterns belong to Cluster 2 with 9052 dailly flow patterns, followed by Cluster 1 presents with 7147 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Cluster 3 has 2740 associated daily patterns that represent predominantly nocturnal consumption.

Figure 4.33 evaluates the degree of membership of each of the annual series to the formed clusters:

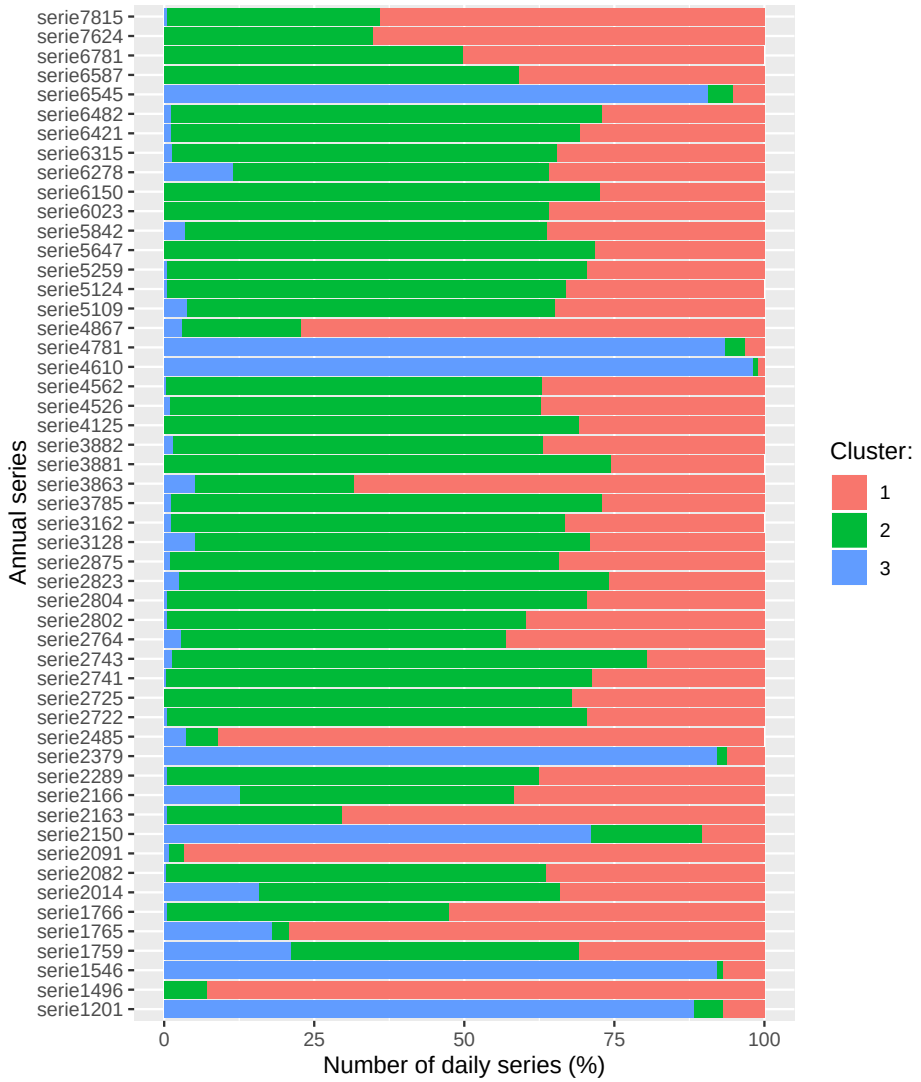

Figure 4.33: k-Shape model annual series membership.

In Figure 4.33 it is observed that in all the annual series the daily patterns belong mostly to Clusters 1 and 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous clustering methods, since most clusters belong to a pattern with predominantly diurnal consumption.

Table 4.4 shows a set of statistical characteristics of the clusters formed:

Table 4.4: k-Shape model clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**	**Cluster 3 (m^3/h)**
Min.	0.00	0.00	0.00
1st Qu.	8.00	6.80	5.00
Median	23.38	16.40	11.52
Mean	52.39	41.33	22.06
3rd Qu.	67.50	52.00	25.55
Max.	1067.00	1207.00	981.25
IQR	59.50	45.20	20.55

Figure 4.34 identifies the influence of weekend or holiday days have on the formation of clusters:

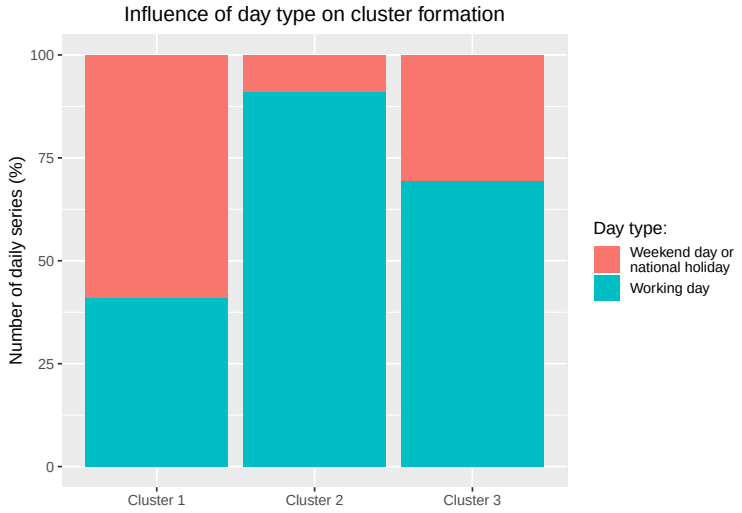

Figure 4.34: k-Shape model influence of day typology on the formation of clusters.

As it can be seen, for cluster 2 the percentage of weekend or national holiday patterns is around 12%, proving that this cluster is associated with typical working day behavior.

In the case of cluster 1, the percentage of weekend or holiday patterns is around 63%, proving that this cluster is associated with typical weekend or holiday behavior.

For Cluster 3 the percentage of weekends and holidays is around 30%. These values indicate that the formed cluster do not allow to identify a distinct behavior between a working day and a weekend or holiday.

These results are identical to those obtained by partition model with DTW distance, PAM prototype and 15m window. Although in k-Shape model Cluster 2 the percentage of daily patterns of weekend or holiday typology is lower.

Figure 4.35 allows identifying the influence of day typology in each annual series by cluster type:

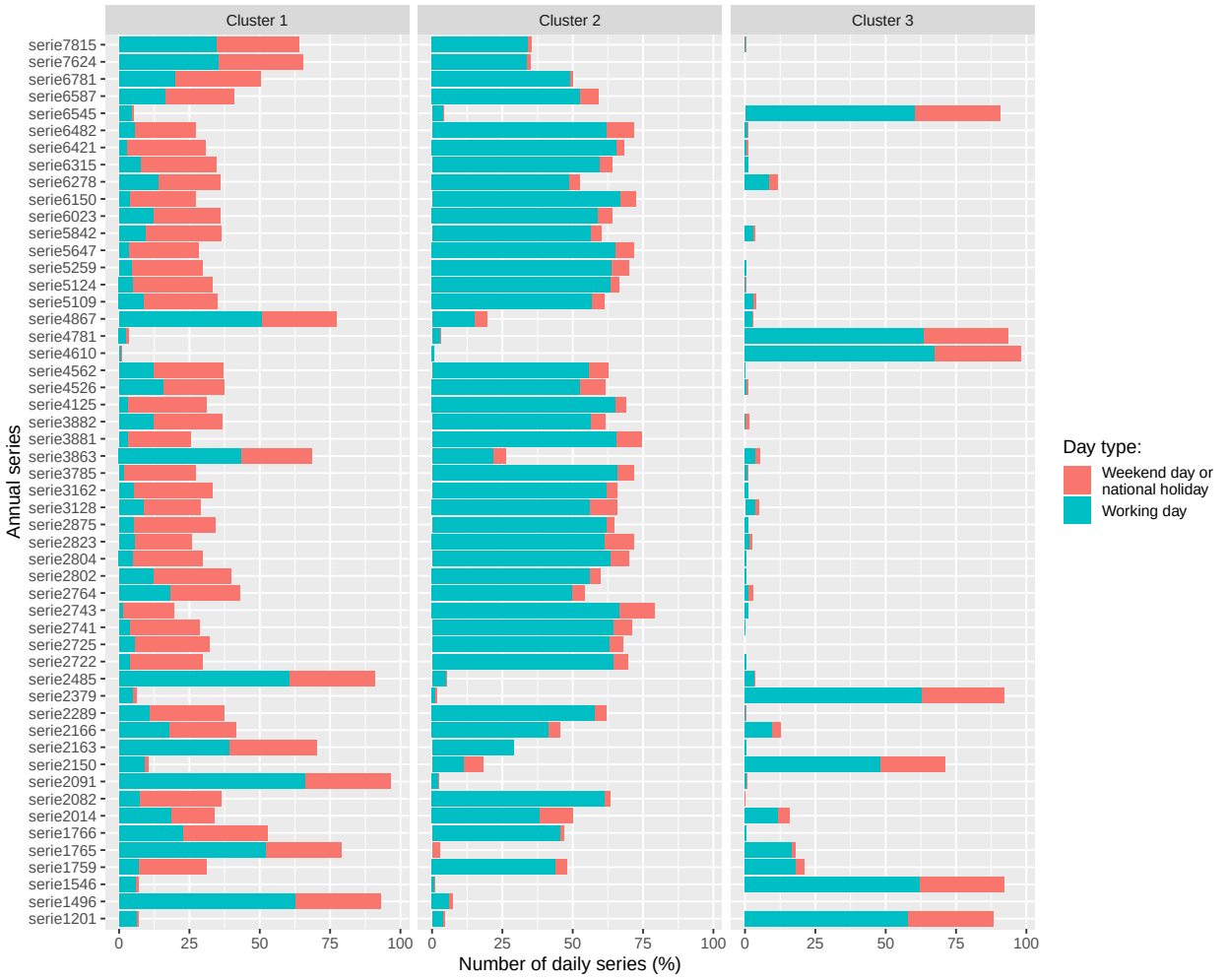

Figure 4.35: k-Shape model influence of day typology on each series by clusters.

As can be seen in Cluster 2, the annual series show mostly a higher percentage of daily patterns in working days.

In the case of Cluster 1 for most of the annual series, there are a greater number of daily patterns of weekend day or national holiday type.

Cluster 3, which represents daily patterns with higher nocturnal consumption, shows that in the annual series in which this cluster is the most representative, the proportions between working days and national holiday / weekend are indicative that typology of the day does not have significant influence on this cluster.

4.5 Summary of clustering models analysis

Table 4.5 summarizes the outputs and limitations of each cluster operation performed in the previous sub-chapters.

As shown in Table 4.5 the models where the best performance iteration only forms 2 clusters, differentiate the daily patterns by presenting higher consumption at night or higher consumption during the day. In the case of the hierarchical model (section A.2), it presents a large mix of clusters in the boundary zones between the two clusters (Figure A.10), which highlights the rigidity of this clustering algorithm that does not allow an element to change clusters in later iterations. In contrast, the Fuzzy Clustering model (section A.3) presented the best-defined boundaries (Figure A.17), highlighting the soft partitioning nature of the algorithm. All the remaining models that formed only 2 clusters present some mix in the cluster boundary zones although not as mixed as in the case of hierarchical clustering.

In the case of the models that in the best iteration formed 3 clusters, those are differentiated in different ways:

1. Models that can differentiate between daily patterns with predominantly daytime consumption from patterns with nighttime consumption. With regards to predominant daytime consumption patterns, they can still distinguish 2 subsets that differ according to the occurrence of daytime peak consumption. These 2 subsets can be distinguished by their peak consumption in the morning or peak consumption around 12:00 (weekend vs. workday pattern). This behavior presents the Partitional Clustering models with DTW, PAM prototype and 15m time window (section 4.4.1) and k-Shape Clustering (section 4.4.3). In the case of the Partitional Clustering model with DTW, PAM prototype and 30 min time window. (section B.3), the 2 subsets formed from the predominantly daytime daily consumption patterns are distinguished by patterns that have their peak consumption in the morning or patterns where their peak consumption near dinner time. Regarding the boundaries of the formed clusters, there is a greater mix at the boundaries between the subsets representing the predominantly daytime consuming patterns;
2. Models that can differentiate between predominantly nighttime consumption, predominantly daytime consumption and also a 3^{rd} set where daily patterns that do not differ greatly between daytime and nighttime consumption, are inserted. The Partitional Clustering models with GAK, PAM prototype and 15 or 30m time window (respectively sections 4.4.2 and B.6) exhibit this behavior. Regarding the boundaries between the clusters, it is observed that these models have a larger border mixture between the cluster that represents the daily patterns with night consumption and the cluster that represents the patterns that do not differ greatly between day and night consumption.

Given the above aspects, it can be concluded that the models that form 3 clusters can provide more information about the daily patterns present in the dataset under analysis.

Table 4.5 shows that k-Means models (section A.1), Partition Clustering with DTW, Mean prototype and 15m or 30m time window (respectively section B.1 and B.2), differ only in the

time window comparison (respectively 0, 15 and 30 minutes) and have the same result at the level of formed clusters, only 2 clusters, as well as present centroids of clusters with similar characteristics. When comparing models that differ only by using different prototypes, such as Partition Clustering with DTW, Mean centroid and 15m time window (section B.1) and Partition Clustering with DTW, PAM centroid and 15m time window (section 4.4.1), it is revealed that with the use of the PAM prototype 3 clusters are formed instead of the 2 formed by the Mean prototype model. This indicates that the time windows comparison is not as an important factor as choosing the type of prototype to use in partition clustering models.

In Table 4.5, looking at the prototypes used in the models and the number of clusters they produced, it was validated that the models that use as centroids the dataset objects that minimize the distances to the other cluster members to which they belong (PAM prototype and Shape- extraction prototype) tend to form 3 clusters as the best iteration of the model and thus better capture the characteristics present in the dataset under study. In the case of prototype typologies that use centroid as an arithmetic mean of the values of objects present in the cluster to which the centroid belongs (Mean prototype and DBA prototype), they tend to form 2 clusters as a better iteration of the model and do not capture the characteristics so effectively.

The following models were chosen as having the best performance in dataset feature extraction:

- Partitition Clustering with DTW, PAM prototype and 15 minutes time window (section 4.4.1);
- Partitition Clustering with GAK, PAM prototype and 15 minutes time window (section 4.4.2);
- k-Shape Clustering (section 4.4.3).

These models were chosen because they present: the formation of 3 clusters as the best iteration, relevant model outputs to characterize the dataset (Table 4.5), make use of different distance measurements, use prototypes that minimize distances to the remaining members of the cluster (PAM prototype and Shape extraction prototype), and thus better representing the cluster. Regarding the time window for Partition Clustering models, the minimum window studied for the model (15 minutes) was chosen, since it was revealed that the increment of the temporal window is not an important factor for feature extraction for this particular dataset.

Regarding the Outliers, in the Boxplot analysis of Figure 4.5 it is clear that the 7815, 6871 and 2485 series had maximum flow rates greater than 600 m^3/h. The decision was made not to remove these Outliers prior to clustering operations. Analyzing the centroids obtained by the clustering models (Figures 4.15, 4.23 and 4.31) and the degree of belonging of the series to the clusters (Figures 4.17, 4.25 and 4.33), it appears that the presence of these Outliers had little influence on the formation of clusters and selection of the clusters centroids.

Table 4.5: Summary of clustering models analysis.

Section	Clustering Method	Type of partition	Prototype	Distance Measure	Comparasion time window	Number of Clusters	Outputs	Limitations
4.4.1	Partitional Clustering (k-Medoids)	Hard partition	PAM	DTW (Euclidean)	15 minutes	3	• It forms 3 clusters: one represents patterns with nocturnal consumption superior to daytime consumption, another represents typical working day patterns with daytime consumption superior to nocturnal consumption and a third cluster with typical patterns of weekend or holiday with greater daily consumption than nocturnal; • The algorithm allows to compare flow values up to a time lag of 15 minutes; • The centroids formed correspond to actual patterns of the dataset.	• The projection according to the principal components 1 and 3 does not present a clear distinction between clusters 1 and 2.
4.4.2	Partitional Clustering (k-Medoids)	Hard partition	PAM	GAK	15 minutes	3	• It forms 3 clusters: one represents patterns with nocturnal consumption superior to diurnal consumption, another represents patterns with daytime consumption superior to nocturnal consumption and finally a cluster that represents daily series values without great variability of the value of flow throughout the day; • The algorithm allows to compare flow values up to a time lag of 15 minutes; • The centroids formed correspond to actual patterns of the dataset.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns.
4.4.3	k-Shape Clustering	Hard partition	Shape extraction	Shape based distance	All data points are compared	3	• It forms 3 clusters: one represents patterns with nocturnal consumption superior to daytime consumption, another represents typical working day patterns with daytime consumption superior to nocturnal consumption and a third cluster with typical patterns of weekend or holiday with greater daily consumption than nocturnal; • The algorithm allows to compare flow values up to a time lag of 15 minutes; • The centroids formed correspond to actual patterns of the dataset.	• The projection according to the principal components 1 and 3 does not present a clear distinction between clusters 1 and 2.
A.1	k-Means Clustering	Hard partition	Mean	Euclidean	No window	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption.	• The algorithm only compares flow rates at the same time instants; • Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.

Table 4.5: Summary of clustering models analysis. *(continued)*

Section	Clustering Method	Type of partition	Prototype	Distance Measure	Comparasion time window	Number of Clusters	Outputs	Limitations
A.2	Hierarchical Clustering	Hard partition	Mean	Euclidean	No window	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption.	• The algorithm only compares flow rates at the same time instants; • Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The projection according to the principal components 1 and 3 and also projection according to the principal components 1 and 2, do not present a clear distinction between the formed clusters, evidencing that this model is more rigid in the formation of the clusters; • The centroids formed do not correspond to actual patterns of the dataset.
A.3	Fuzzy Clustering	Soft partition	Mean	Euclidean	No window	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption; • The projections of the clusters according to the principal components allows to validate that there is a clear distinction between the clusters in the border zones. It is possible to conclude that this soft partitioning algorithm is able to define in a more assertive way the cluster to be assigned in the boundary zones.	• The algorithm only compares flow rates at the same time instants; • Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.
B.1	Partitional Clustering (k-Means)	Hard partition	Mean	DTW (Euclidean)	15 minutes	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption; • The algorithm allows to compare flow values up to a time lag of 15 minutes.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.
B.2	Partitional Clustering (k-Means)	Hard partition	Mean	DTW (Euclidean)	30 minutes	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption; • The algorithm allows to compare flow values up to a time lag of 30 minutes.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.

Table 4.5: Summary of clustering models analysis. *(continued)*

Section	Clustering Method	Type of partition	Prototype	Distance Measure	Comparasion time window	Number of Clusters	Outputs	Limitations
B.3	Partitional Clustering (k-Medoids)	Hard partition	PAM	DTW (Euclidean)	30 minutes	3	• It forms 3 clusters: one represents patterns with nocturnal consumption superior to daytime consumption, another represents daytime consumption superior to nocturnal consumption with peak consuption in the morning and a third cluster that represents daytime consumption superior to nocturnal consumption with peak consuption in near dinner time; • The algorithm allows to compare flow values up to a time lag of 30 minutes; • The centroids formed correspond to actual patterns of the dataset.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The projection according to the principal components 1 and 3 does not present a clear distinction between clusters 1 and 2.
B.4	Partitional Clustering (k-Medoids)	Hard partition	DBA	DTW (Euclidean)	15 minutes	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption; • The algorithm allows to compare flow values up to a time lag of 15 minutes.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.
B.5	Partitional Clustering (k-Medoids)	Hard partition	DBA	DTW (Euclidean)	30 minutes	2	• It forms 2 clusters: one represents patterns with nocturnal consumption superior to daytime consumption and the other represents patterns with daytime consumption superior to nocturnal consumption; • The algorithm allows to compare flow values up to a time lag of 30 minutes.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns; • The centroids formed do not correspond to actual patterns of the dataset.
B.6	Partitional Clustering (k-Medoids)	Hard partition	PAM	GAK	30 minutes	3	• It forms 3 clusters: one represents patterns with nocturnal consumption superior to diurnal consumption, another represents patterns with daytime consumption superior to nocturnal consumption and finally a cluster that represents daily series values without great variability of the value of flow throughout the day; • The algorithm allows to compare flow values up to a time lag of 30 minutes; • The centroids formed correspond to actual patterns of the dataset.	• Formed groups do not distinguish typical daily patterns of workdays from typical weekend or holiday patterns.

4.6 Further analysis on best clustering models

In this section the characteristics of geographic distribution and the prepronderance of dry months vs. wet months in the clusters formed by the best models will be analyzed.

The geographical distribution of the dataset comprises the following regions:

- Lisbon metropolitan area;
- North (North Coast);
- Coastal Center;
- Interior Center;
- South (Algarve).

The preponderance of dry months with humid months will have the following constitution:

- Dry months: June, July, August and September;
- Wet months: October, November, January, February, March, April and May.

The following models will be evaluated:

- Partitition Clustering with DTW, PAM prototype and 15 minutes time window (section 4.4.1);
- k-Shape Clustering (section 4.4.3);
- Partitition Clustering with GAK, PAM prototype and 15 minutes time window (section 4.4.2).

4.6.1 Evalution of Partition Clustering with DTW, PAM prototype with 15 minutes time window

Figure 4.36 shows the centroids formed in clustering analysis performed in section 4.4.1.

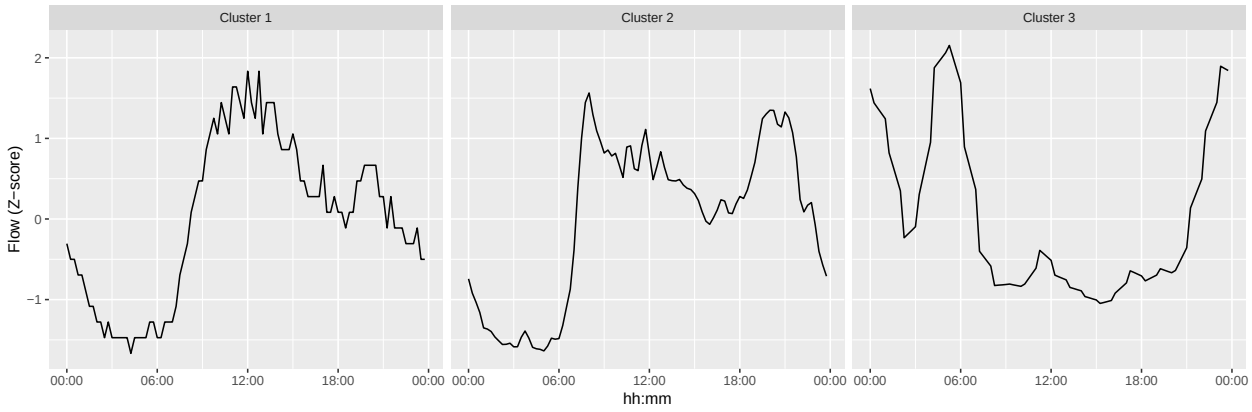

Figure 4.36: Partition Clustering model with DTW distance, PAM prototype and 15m window centroids for further cluster analysis.

Figure 4.37 shows that Cluster 1 representing typical weekend dailly patterns and Cluster 2 representing a typical working day daily pattern have a similar distribution with respect to

geographical location. With greater weight over the Coastal Center and Lisbon Metropolitan Area regions.

Cluster 3, which represents daily patterns with predominantly nocturnal consumption, belongs mainly to the South (Algarve) region. Indicating that in this location there is a strong component of irrigation in the water use.

Figure 4.38 shows that for Cluster 1 about 25% of the daily patterns that make up the cluster belong to dry months. In the case of Cluster 2 about 37.5% of the daily patterns that make up the cluster belong to dry months.

Regarding Cluster 3, it presents about 37.5% of the daily patterns that make up the cluster as belonging to dry months. This behavior was not expected as it is a cluster that represents daily patterns with predominantly nocturnal consumption due to water use irrigation and should occur more in dry months than in humid months. It may indicate that in the south (Algarve) region, to which this cluster mainly belongs, water sprinklers may not be programmed according to the humidity of the terrain and the registered rainfall.

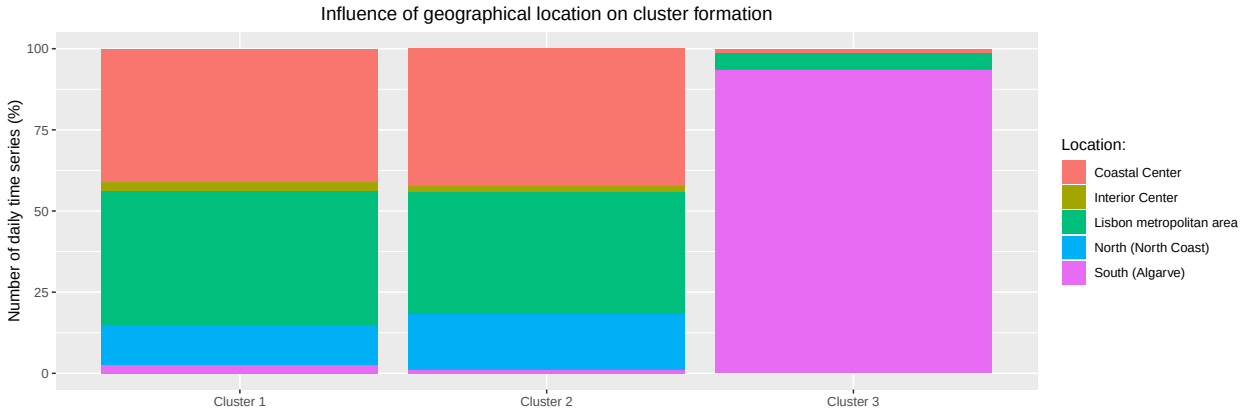

Figure 4.37: Geographic distribution of the clusters formed for Part. Clust. model with DTW, PAM prototype and 15m window.

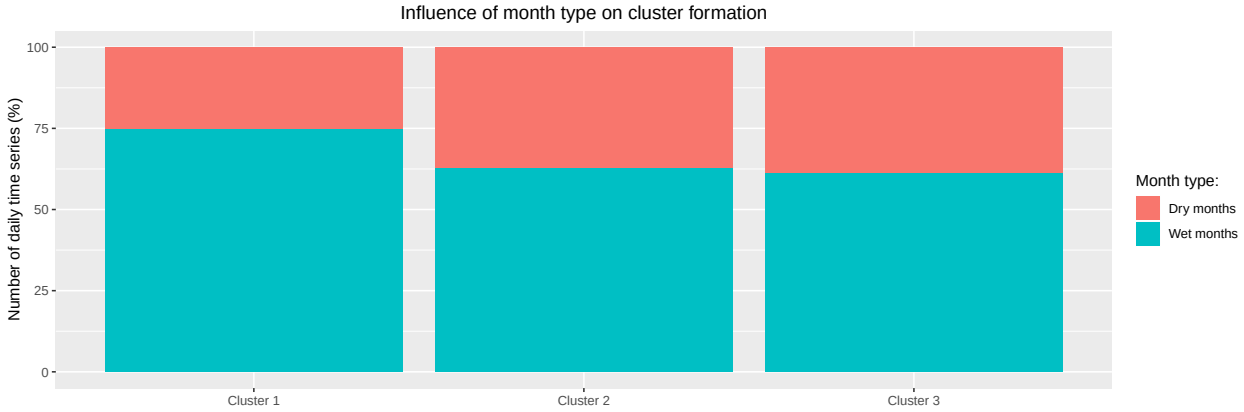

Figure 4.38: Distribution of wet months and dry months for Part. Clust. model with DTW, PAM prototype and 15m window.

4.6.2 Evalution of k-Shape Clustering

Figure 4.39 shows the centroids formed in clustering analysis performed in section 4.4.3.

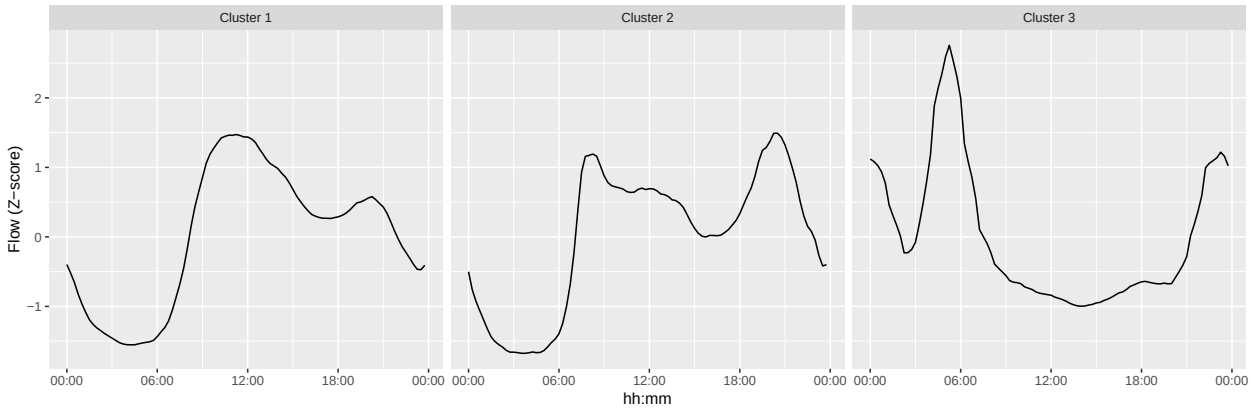

Figure 4.39: k-Shape model centroids for further cluster analysis.

Figure 4.40 shows that Cluster 1 representing typical weekend dailly patterns and Cluster 2 representing a typical working day daily pattern have a similar distribution with respect to geographical location to the presented for Partition Clustering with DTW, PAM prototype with 15 minutes time window. With greater weight over the Coastal Center and Lisbon Metropolitan Area regions.

Cluster 3 also presents similar results to the Partition Clustering with DTW, PAM prototype with 15 minutes time window. This cluster represents daily patterns with predominantly nocturnal consumption and belongs mainly to the south (Algarve) region. Indicating that in this location there is a strong component of irrigation in the water use.

Figure 4.41 shows that for Cluster 1 about 25% of the daily patterns that make up the cluster belong to dry months. In the case of Cluster 2 about 37.5% of the daily patterns that make up the cluster belong to dry months. This results are similar to the Partition Clustering with DTW, PAM prototype with 15 minutes time window.

Regarding Cluster 3, it presents about 44% of the daily patterns that make up the cluster as belonging to dry months. This behaviour is similar to the Partition Clustering with DTW, PAM prototype with 15 minutes time window, and was not expected as it is a cluster that represents daily patterns with predominantly nocturnal consumption due to water use irrigation and should occur more in dry months than in humid months. It may indicate that in the South (Algarve) region, to which this cluster mainly belongs, water sprinklers may not be programmed according to the humidity of the terrain and the registered rainfall.

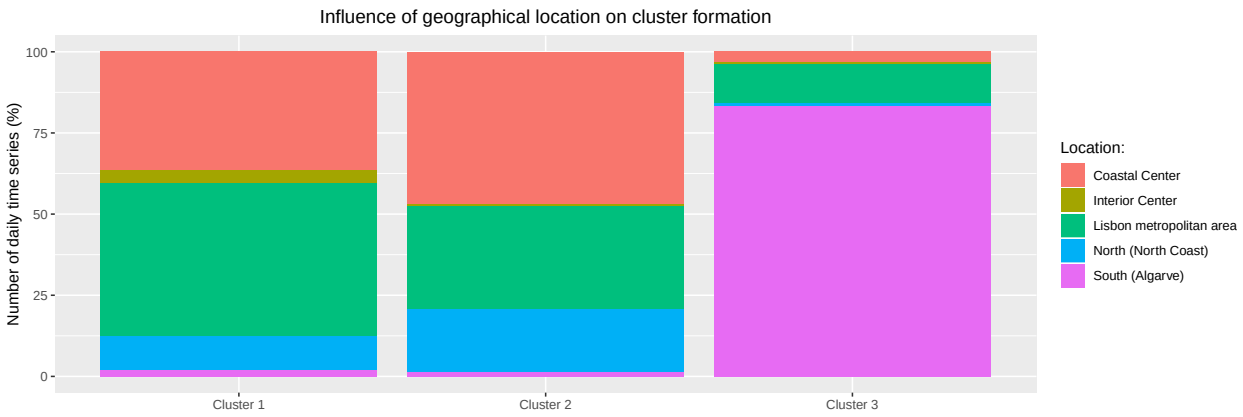

Figure 4.40: Geographic distribution of the clusters formed for k-Shape Clust. model.

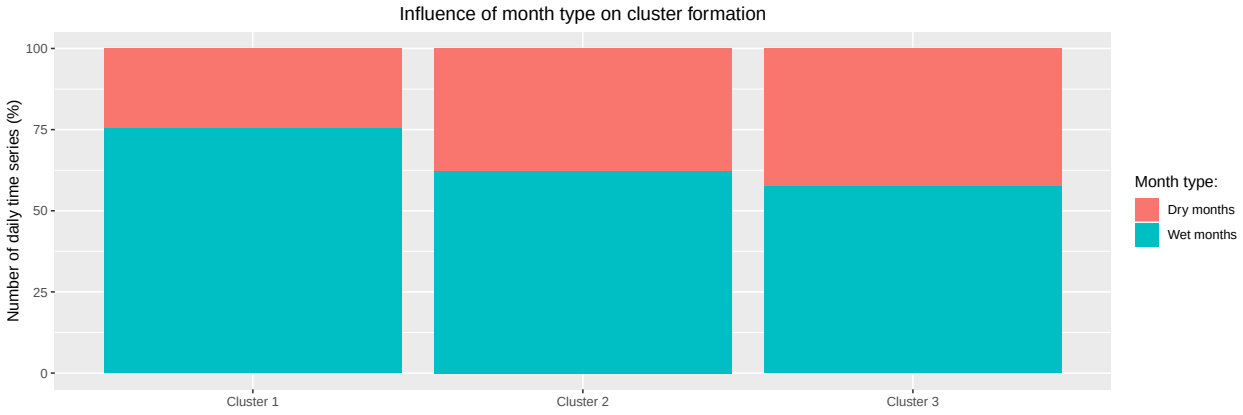

Figure 4.41: Distribution of wet months and dry months for k-Shape Clust. model.

4.6.3 Evalution of Partition Clustering with GAK, PAM prototype with 15 minutes time window

Figure 4.42 shows the centroids formed in clustering analysis performed in section 4.4.2.

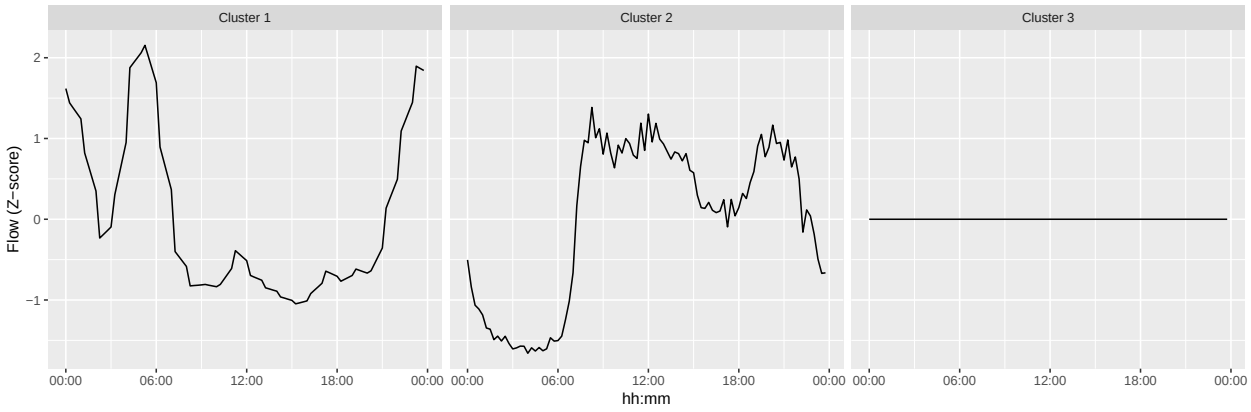

Figure 4.42: Partition Clustering model with GAK distance, PAM prototype and 15m window centroids for further cluster analysis.

Figure 4.43 shows that Cluster 1 represents daily patterns with predominantly nocturnal consumption similar to the Cluster 3 from the previous models. About 98% of this cluster belongs to the South (Algarve) region.

Cluster 2 is represented daily patterns with predominantly daytime consumption. The regions most represented by this cluster are Costal Center, Lisbon Metropolitan Area and North (North Coast) regions.

The centroid of cluster 3 has a constant flow rate over time. Since centroid PAM is a real dataset pattern with the minimum distance to the remaining cluster members, it may represent all the patterns that have the lowest variability over time and little variation between daytime and nighttime consumption. For this cluster it is verified that all regions are represented, with emphasis on the South (Algarve), Lisbon metropolitan area and Coastal Center regions.

Figure 4.44 shows that for Cluster 1 about 37.5% of the daily patterns that make up the cluster belong to dry months. This behaviour is similar to the Cluster 3 of Partition Clustering with DTW, PAM prototype with 15 minutes time window and the Cluster 3 of the k-Shape clustering model. This behaviour is not expected since it is a cluster that represents daily patterns with predominantly nocturnal consumption due to water use irrigation and should occur more in dry months than in humid months. It may indicate that in the south (Algarve) region, to which this cluster mainly belongs, water sprinklers may not be programmed according to the humidity of the terrain and the registered rainfall.

In the case of Cluster 2 about 31% of the daily patterns that make up the cluster belong to dry months. As for Cluster 3, about 44% of the daily patterns that make up the cluster belong to dry months.

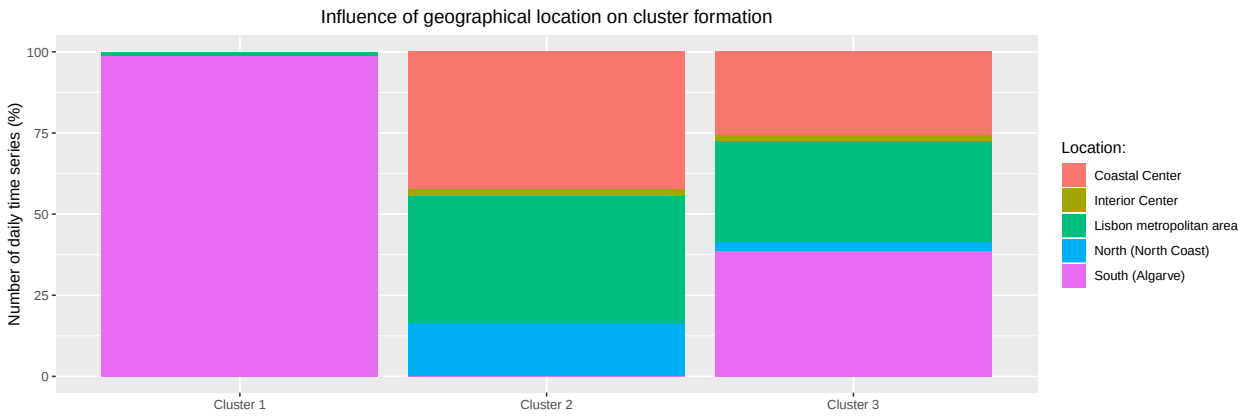

Figure 4.43: Geographic distribution of the clusters formed for Part. Clust. model with GAK, PAM prototype and 15m window.

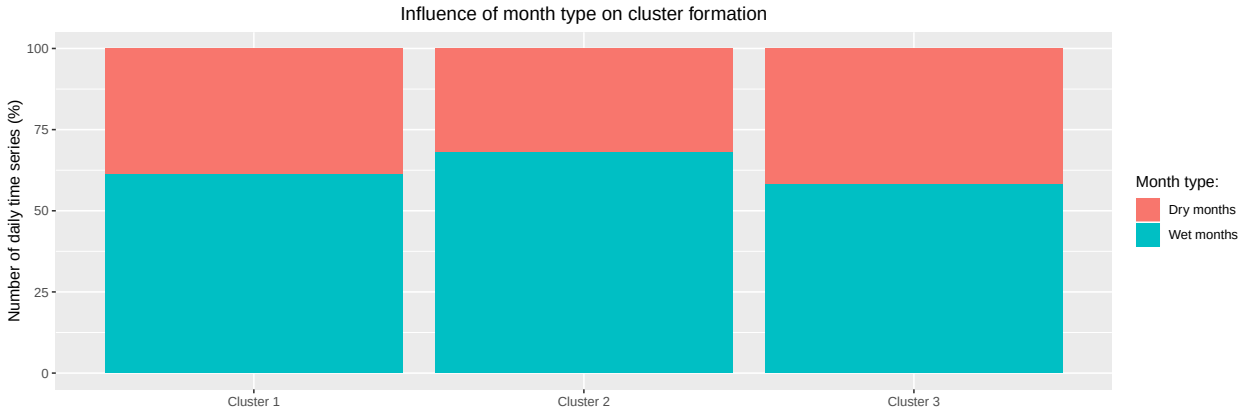

Figure 4.44: Distribution of wet months and dry months for Part. Clust. model with GAK, PAM prototype and 15m window.

4.6.4 Summary on further analysis on best clustering models

From the subsequent analysis it was found that the Partition Clustering with DTW, PAM prototype with 15 minutes time window and k-Shape Clustering models formed identical clusters and presented similar distributions at the level of geographic distribution and distribution of wet months vs. dry months.

It was also validated that the cluster that in each model represents daily patterns of predominantly nocturnal consumption and irrigation, belongs mostly to the South (Algarve) region and represents mostly wet months. This behavior was not expected and indicates that there may be incorrect management of water use for irrigation, as this cluster is more associated with wet months and it is recommended that irrigation systems be regulated to take into account the soil humidity level and the recorded rainfall in the South (Algarve) region.

From the point of view of the representation of clusters formed by each model, Partition Clustering with DTW, PAM prototype with 15m time window and k-Shape Clustering models present the same set of clusters:

- Cluster 1: represents typical weekend daily patterns with peak consumption near 12:00;
- Cluster 2: represents typical working day daily patterns with peak consumption near 7:30;
- Cluster 3: represents daily patterns with predominantly nighttime consumption associated with irrigation water use.

In the case of Partition Clustering with GAK, PAM prototype with 15m time window model, the formed clusters represent:

- Cluster 1: represents daily patterns with predominantly nighttime consumption associated with irrigation water use;
- Cluster 2: represents daily patterns with predominantly daytime consumption;
- Cluster 3: represents all the patterns that have the lowest variability over time and little variation between daytime and nighttime consumption.

In the next section we will propose a combined model that aggregates the characteristics of the Partition Clustering with DTW, PAM prototype with 15m time window and Partition Clustering with GAK, PAM prototype with 15m time window models described in the present analysis. From Partition Clustering with DTW, PAM prototype with 15m time window model, Clusters 1 and 2 will be incorporated in the combined model, in order to assimilate groups representing daily weekend patterns and daily working day patterns, respectively. From the Partition Clustering with GAK, PAM prototype with 15m time window model, Cluster 3 will be incorporated into the combined model, in order to assimilate a group that represents all the patterns that have the lowest variability over time and little variation between daytime and nighttime consumption.

Given that the Partition Clustering with DTW, PAM prototype with 15m time window and k-Shape Clustering models are identical in the model outputs, only the Partition Clustering with DTW, PAM prototype with 15m time window model was used in the combined model since it presents the same prototype typology (PAM) as the Partition Clustering with GAK, PAM prototype with 15m time window model.

4.7 Combined model analysis

In this section a combined model will be presented that will consist of a combination of the clusters of the best performing models analyzed in the previous chapter:

- Partition Clustering with DTW, PAM prototype with 15m time window;
- Partition Clustering with GAK, PAM prototype with 15m time window.

Regarding the Partition Clustering with DTW, PAM prototype with 15m time window model the following clusters will be incorporated:

- Cluster 1: represents typical weekend daily patterns with peak consumption near 12:00;

- Cluster 2: represents typical working day daily patterns with peak consumption near 7:30;
- Cluster 3: represents daily patterns with predominantly nighttime consumption associated with irrigation water use.

Cluster 2 of the Partition Clustering with GAK, PAM prototype with 15m time window model, which represents daily patterns with predominantly daytime consumption, will also be incorporated into the combined model. This cluster will overlap selected clusters from Partition Clustering model with DTW, PAM prototype with 15m time window, which implies a reduction of members in these clusters, being necessary to recalculate the PAM centroids in clusters 1,3 and 4 of the combined model.

Figure 4.45 shows the process of combining clusters to create the combined model:

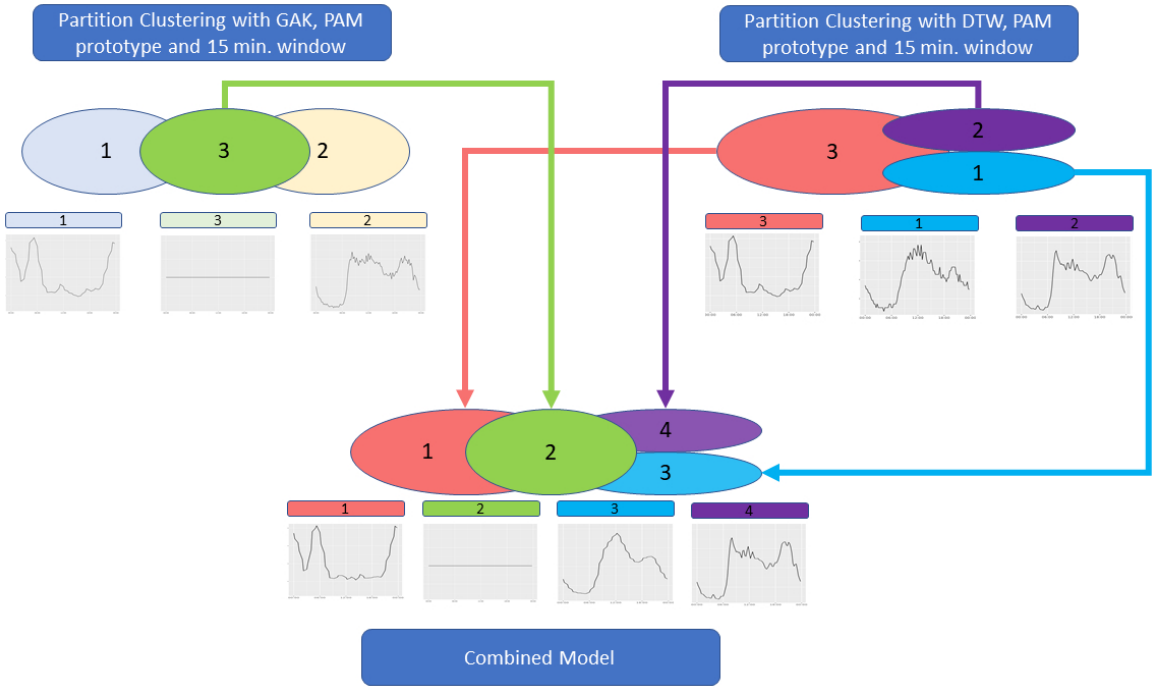

Figure 4.45: Characterization and evaluation workflow of cluster models with elastic distance measurements.

Figure 4.46 shows the visualization of the clusters formed by the model according to the first 3 principal components:

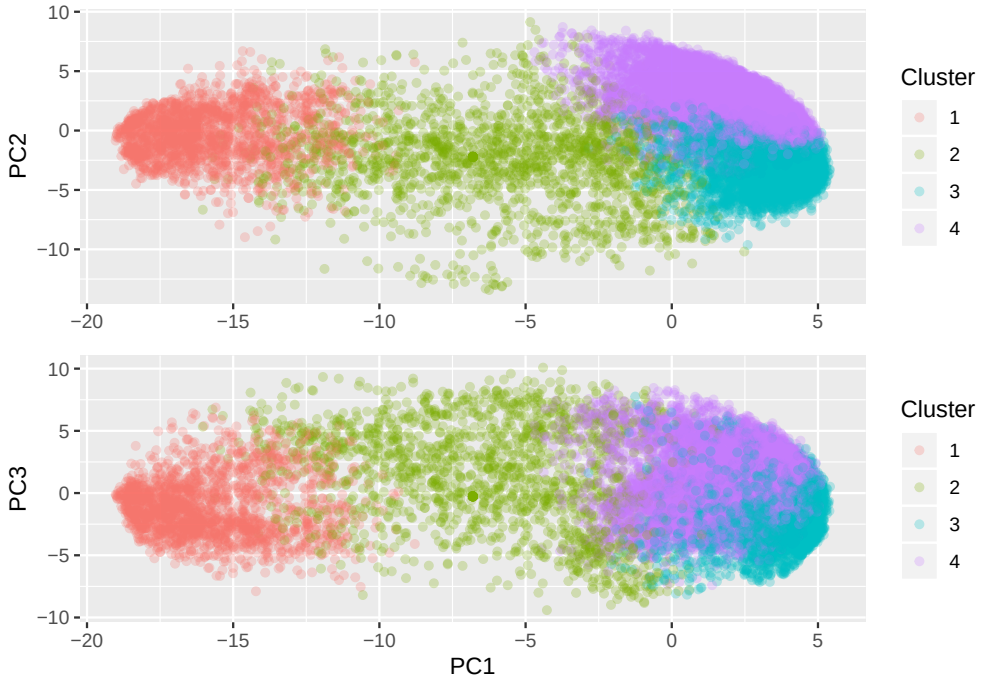

Figure 4.46: Clusters formed through the Combined Model visualized through the 3 principal components of PCA.

Figure 4.46 shows a greater dispersion of the members of Cluster 2. However, looking at the centroid of Cluster 2 and the fact that this cluster is in the zone between Cluster 1 and the zone composed by Clusters 3 and 4, indicates that this cluster is composed of daily patterns that do not have significant daytime and nighttime variability. In section 4.7.1 clustering will be done on Cluster 2 to characterize the subsets present in Cluster 2.

4.7.1 Cluster 2 - Application of clustering models with elastic distance measures

In this section clustering operations on Cluster 2 of the Combined Model will be performed according to the algorithms with elastic distance measurements that allow the comparison of flow values belonging to different time periods.

The following clustering approaches will be used:

- Partition Clustering with DTW, PAM prototype with 15m time window;
- Partition Clustering with GAK, PAM prototype with 15m time window.

Figure 4.47 shows the evaluation and characterization procedures of the clustering models evaluated in this section:

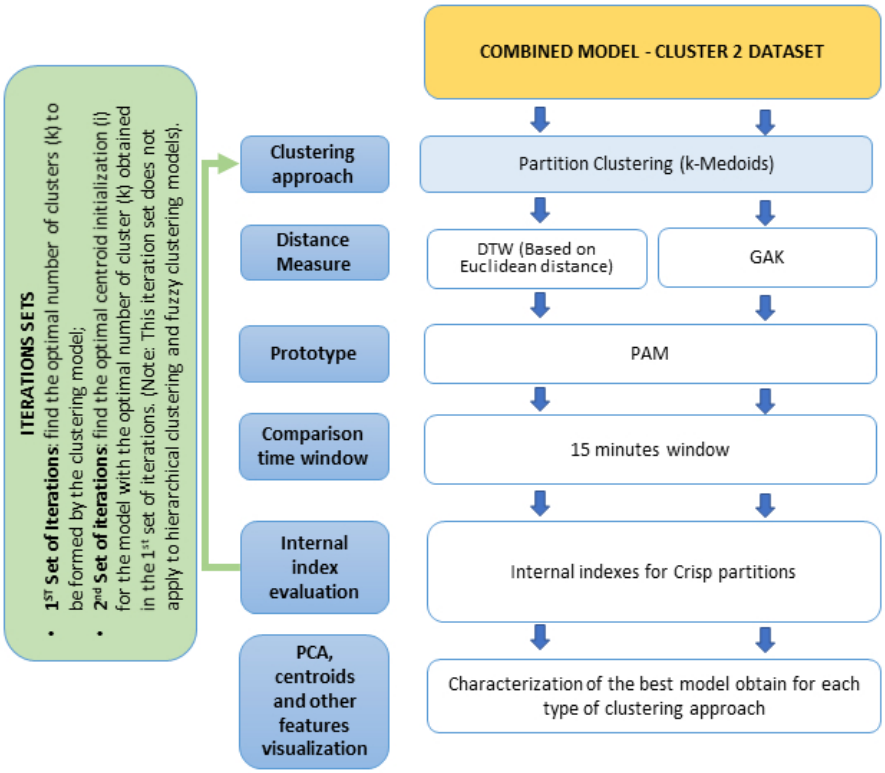

Figure 4.47: Characterization and evaluation workflow of cluster models with elastic distance measurements applied on Cluster 2.

In **Appendix C** the following models are evaluated and characterized:

- **C.1**: Cluster 2 - Partitional Clustering with DTW, PAM prototype and 15 minutes time window;
- **C.2**: Cluster 2 - Partitional Clustering with GAK, PAM prototype and 15 minutes time window.

Summary of analysis on cluster 2

From the analysis of Figures C.5 and Figure C.13 of clustering operations on Cluster 2, it was found that in both models there is a Cluster whose centroid is represented by a pattern that does not show flow variation over time and that aggregates most of the daily patterns. In Partitional Clustering with DTW model, the centroid is represented in Cluster 2.5 (see Figure C.5) and in Partitional Clustering with GAK model, the centroid is represented in Cluster 2.7 (see Figure C.13). These clusters group the cases of daily patterns that show less variability between nighttime and daytime consumption compared to the daily patterns belonging to the remaining clusters.

Figure 4.48 shows the size of clusters 2.5 and 2.7 compared to the remaining clusters formed in their clustering models.:

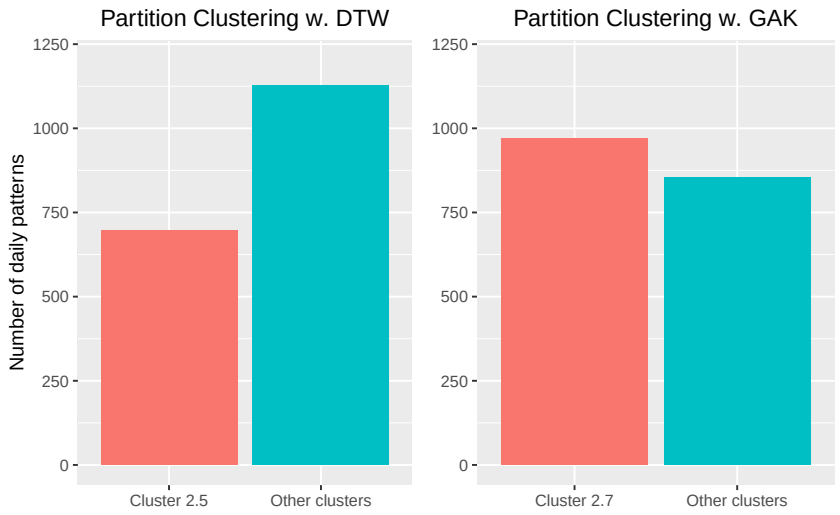

Figure 4.48: Cluster 2 - Clusters size comparison.

Figure 4.48 shows that in Partitional Clustering with DTW model, Cluster 2.5 has a smaller size which makes the remaining formed clusters larger and more representative of the behaviors present in the set of patterns that make up Cluster 2. In the case of Partitional Clustering with DTW model, Cluster 2.7 has a larger size which makes the remaining formed clusters to have fewer elements and be less representative.

In this analysis it was decided to use clusters formed through Partitional Clustering with DTW model to represent the subsets present in Cluster 2 of the Combined Model. In the next section the Combined Model will be presented taking into account the subsets of Cluster 2 formed through the Partitional Clustering with DTW model.

4.7.2 Combined model final representation

Clustering model characterization

Figure 4.49 shows the visualization of the clusters formed by the model according to the first 3 principal components:

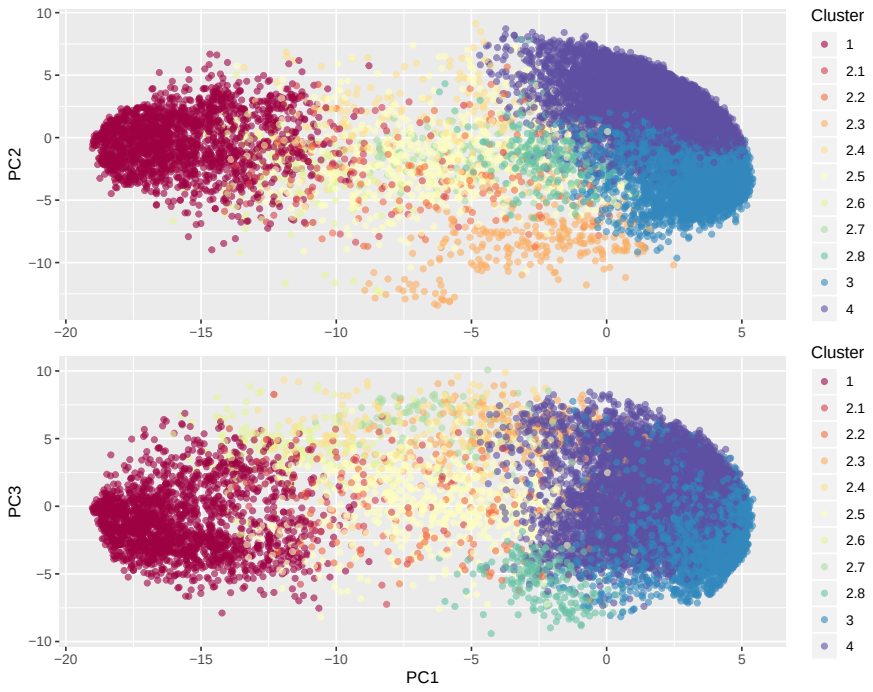

Figure 4.49: Combined Model final representation visualized through the 3 principal components of PCA.

From Figure 4.49 it can be seen that Cluster 1, associated with predominantly nocturnal consumptions, is well defined on the left according to Principal Component 1. Clusters 3 and 4, of predominantly daytime consumption, are on the right according to Principal Component 1. The boundary separating Clusters 1 and 2 is well defined according to Principal Component 2, with values greater than 0, in Principal component 2, being associated with Cluster 4 which represents typical working day patterns. In the case of values below 0, according to the principal component 2, they are associated with Cluster 3 which represents typical weekend patterns. The subsets of Cluster 2 are in the middle zone of Principal Component 1 meaning that the differences between the predominance of nighttime consumption versus daytime consumption in these clusters are not so evident.

Figure 4.50 shows the size of each of the clusters formed:

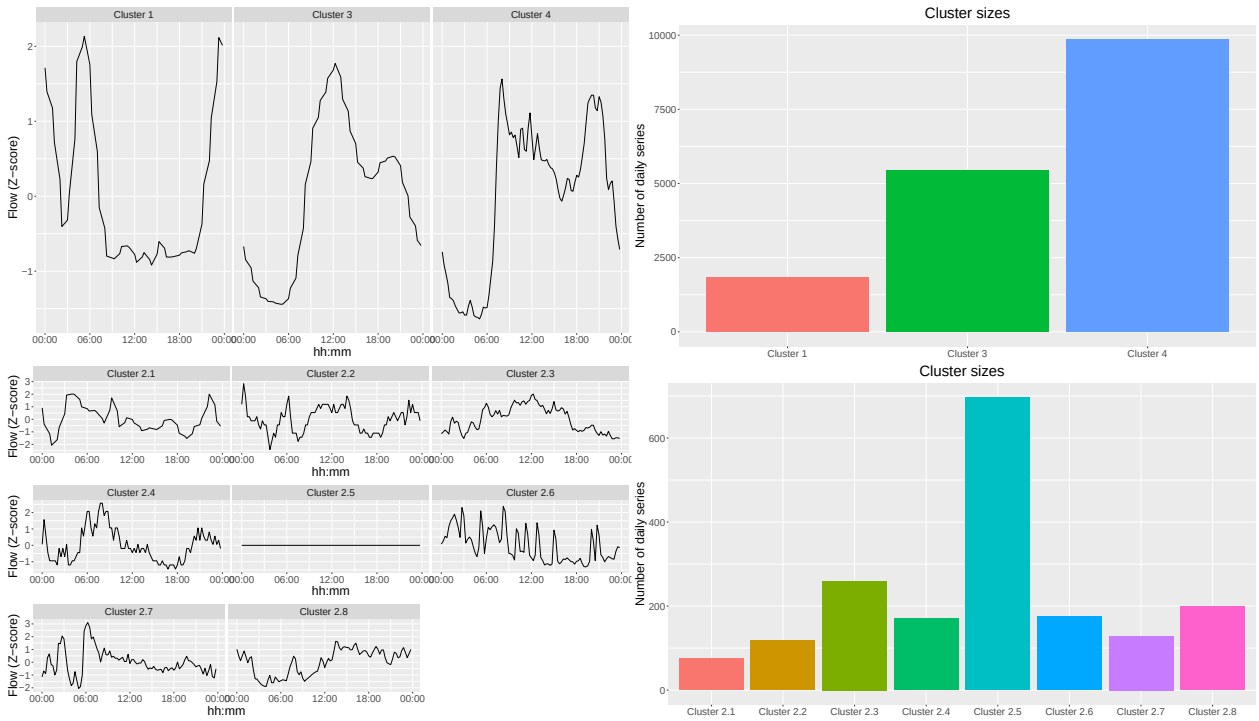

Figure 4.50: Combined Model final representation of clusters sizes.

Figure 4.50 shows that Clusters 1, 3 and 4 aggregate a greater number of daily patterns. Cluster 4, representing a typical daytime working day pattern is the largest cluster, followed by Cluster 3 representing a typical weekend daytime pattern behavior. Cluster 1 represents predominantly nighttime consumption patterns that are smaller in size than previously mentioned clusters.

In the case of the subsets of Cluster 2, they are much smaller in size than Clusters 1, 3 and 4. Cluster 2.5 stands out in size, this cluster groups patterns that show little consumption variability over time. The remaining subsets are in the same order of magnitude at the 75 to 296 member repsentatity level.

Table 4.6 shows a set of statistical characteristics of the clusters formed:

Table 4.6: Combined Model final representation of clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2.1 (m^3/h)**	**Cluster 2.2 (m^3/h)**	**Cluster 2.3 (m^3/h)**	**Cluster 2.4 (m^3/h)**	**Cluster 2.5 (m^3/h)**	**Cluster 2.6 (m^3/h)**	**Cluster 2.7 (m^3/h)**	**Cluster 2.8 (m^3/h)**	**Cluster 3 (m^3/h)**	**Cluster 4 (m^3/h)**
Min.	0.00	0.97	0.00	0.00	0.00	0.00	0.07	0.95	0.00	0.00	0.00
1st Qu.	4.92	6.86	18.40	8.75	10.13	2.35	11.12	22.00	103.61	7.52	7.27
Median	10.37	12.12	60.00	20.80	52.00	5.90	20.04	28.40	128.12	19.99	18.22
Mean	18.18	34.65	62.44	25.95	55.79	17.29	29.02	31.30	110.76	46.04	46.21
3rd Qu.	21.60	19.39	95.42	34.80	76.00	16.19	31.69	34.40	145.99	57.37	58.00
Max.	312.25	881.00	208.75	251.08	376.42	981.25	299.62	766.37	184.24	1067.00	1207.00
IQR	16.68	12.53	77.02	26.05	65.86	13.84	20.57	12.40	42.38	49.85	50.73

Figure 4.51 identifies the influence of weekend or holiday days have on the formation of

clusters:

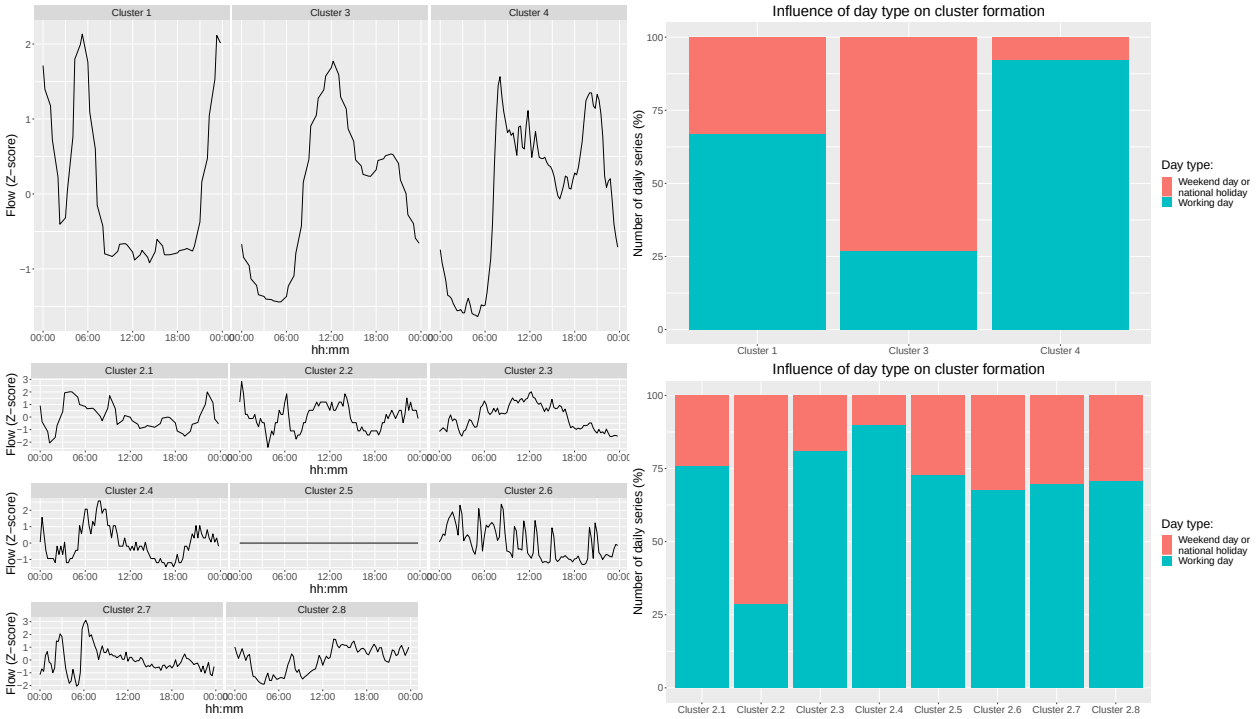

Figure 4.51: Combined Model final representation - influence of day typology on the formation of clusters.

Figure 4.51 shows that only Clusters 3 and 2.2 are more associated with weekend days or national holidays. The remaining clusters are mostly associated with working days.

Figure 4.52 shows the geographic distribution of the clusters formed by the model:

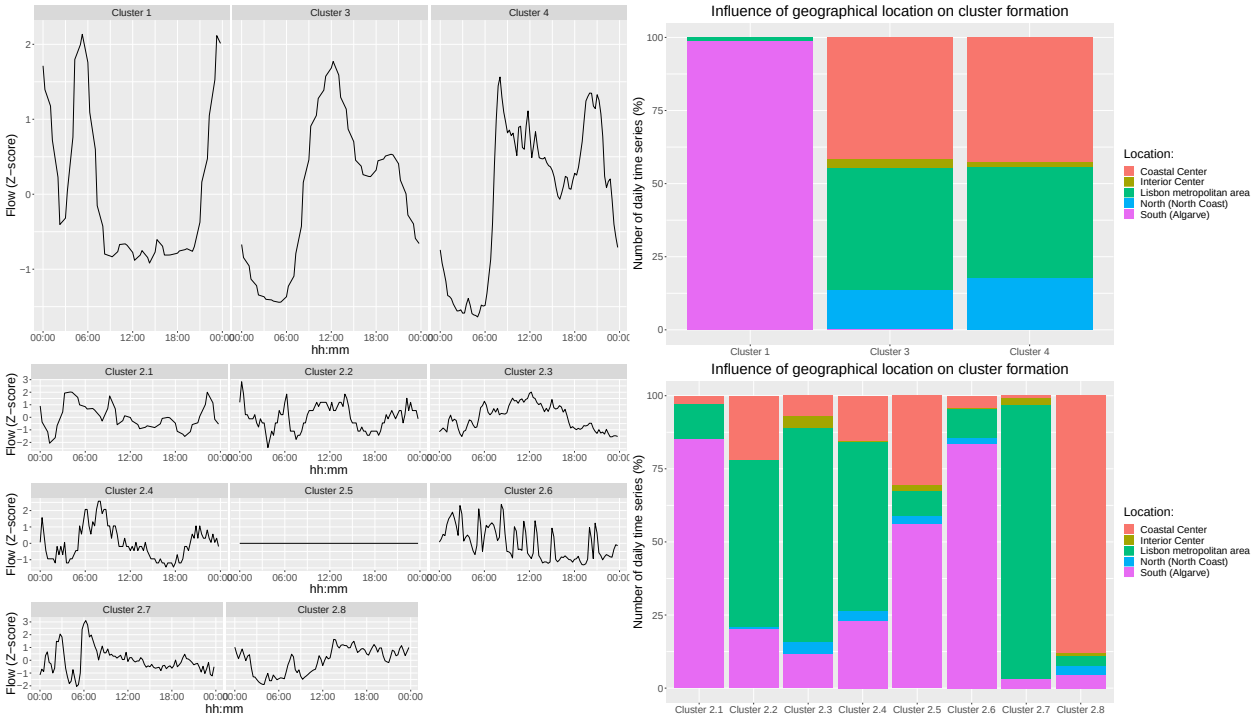

Figure 4.52: Combined Model final representation - geographic distribution of the clusters formed.

Figure 4.52 shows that Clusters 3 and 4 show similar distribution of locations, these clusters represent typical weekend and working day behaviors, respectively. All locations are well represented in these clusters except the South (Algarve) region. Cluster 3, which represents patterns with mostly nocturnal consumption, belongs mostly to the South (Algarve) region.

In the case of Cluster 2 subsets, Clusters 2.1, 2.5, 2.6 have a greater presence in the South (Algarve) region. Clusters 2.2, 2.3, 2.4 and 2.7 have more presence in the Lisbon Metropolitan Area. Cluster 2.8 mainly represents Costal Center region.

Figure 4.53 shows the distribution of wet months and dry months in the clusters formed by the model:

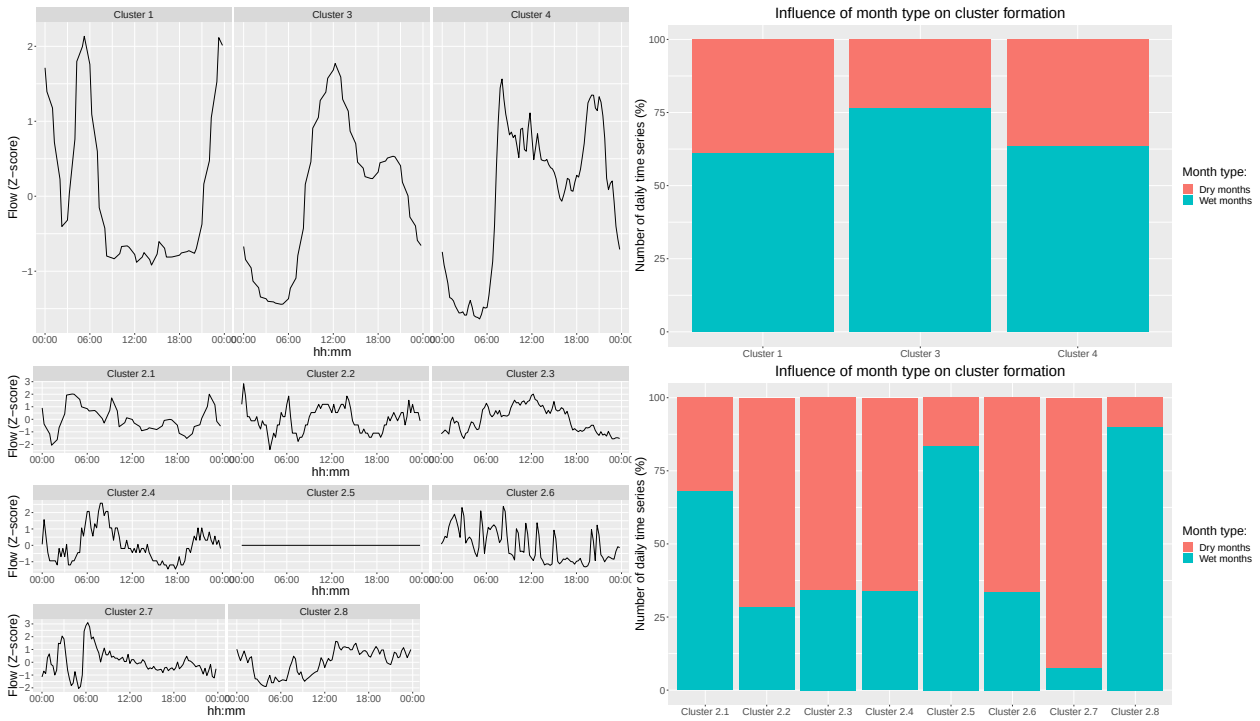

Figure 4.53: Combined Model final representation - distribution of wet months and dry months in the clusters formed.

Figure 4.53 shows that Clusters 1, 3 and 4 are more associated with wet months than with dry months. In the case of Cluster 1, which represents predominantly nighttime consumption associated with irrigation, it was not expected that this cluster would be more associated with wet months, which may indicate that irrigation controllers could be programmed at the same frequency regardless of the time of year, indicating a less sustainable use of water.

For subsets of cluster 2, it is found that most clusters belong more to dry months, except for Clusters 2.1, 2.5 and 2.8.

4.7.3 Summary of the combined model analysis

The Combined Model aggregates the clusters of the best performing models in order to highlight the main characteristics of the consumption profiles present in the dataset.

With this model it was possible to identify clusters with predominantly nocturnal consumption behaviour such as Cluster 1, and clusters with predominantly daytime consumption such as Clusters 3 and 4.

Relatively to typical weekend or holiday daily patterns and typical workday patterns, this model validates the existence of Clusters 3 and 2.3, which are predominantly associated with typical weekend behaviors. The remaining clusters are more associated with typical workday behaviors.

At the region level, it was found that the South (Algarve) region is closely associated with

the Cluster 1 of predominantly nightly consumption and irrigation water use. Cluster 2.1 is also strongly associated with this region, although in this case there is high nighttime consumption, there is also significant daytime consumption.

In the case of Cluster 2.7, this is associated with the Lisbon Metropolitan Area region, with a large component of irrigation at 06:00 and moderate daytime consumption.

Cluster 2.8 is associated with the Coastal Center region and is characterized by not having a significant consumption in the morning.

This model also identifies 2 clusters in which water use for irrigation may not be efficient:

- **Cluster 1**: It has a predominantly nocturnal consumption for irrigation and is more associated with humid months than dry months. This may show that irrigation systems are scheduled to be operated independently of irrigation needs over the time period of the year.
- **Cluster 2.6**: show instant consumption peaks throughout the day associated with irrigation systems. Operating these irrigation systems during periods of increased heat and sun exposure may not be effective and sustainable as some of the water evaporates before it is absorbed into the soil.

Both clusters are associated with the southern region (Algarve). In this region, the operation of irrigation systems should be rethought to allow more sustainable and efficient water use.

Chapter 5

Conclusions and future developments

This dissertation intends to contribute to the study of clustering methods applicable to medium flow time series, focusing on average daily flow patterns (m^3/h) with a time step of 15 minutes.

The following methods were validated as the best performing clustering algorithms for the studied dataset: Partitional Clustering with DTW distance, PAM centroid; k-Shape Clustering; Partitional Clustering with GAK distance, PAM centroid.

The first two methods presented, as the best approach, the formation of three clusters in which the respective centroids describe a mostly nocturnal consumption pattern, a typical weekend pattern and a typical work day pattern. The k-Shape method also obtained as the best approach the formation of three clusters, but the centroids of these clusters describe a pattern with mostly nighttime consumption, a pattern with mostly daytime consumption and an intermediate pattern that represents the cases where the variability between daytime consumption and nighttime consumption is not so pronounced.

In terms of the distance measurements used, this dissertation validated that elastic distance measurements, namely DTW, GAK and Shape distance, allow clustering approaches to produce better results than inelastic distance measurements such as Euclidean. This result validates that elastic measurements allow a better capture of the "shape" of the patterns present in the dataset. Regarding the choice of the temporal window in the case of DTW distance measurement, it was found that increasing the temporal window from 15m to 30m had no major influence on clusters formation compared to choosing the distance measurement or choosing the prototype.

At the prototype level, the analyzes performed validated that the prototype typologies that represent a real dataset pattern that minimizes the distances to the other cluster members (PAM prototype and Shape extraction prototype), when used in cluster approaches, tend to form 3 clusters, whereas the use of prototypes that represent a medium standard (DBA prototype and Mean prototype) in clustering approaches tend to form only 2 clusters. It was concluded that for the analyzed dataset, the PAM and Shape extraction prototypes produce better results in clustering operations.

Regarding the methods of internal index evaluation, in the present study the following measures were used in the hard partitioning approaches: Silhouette Index, Dunn Index, COP Index, Davies-Bouldin Index, Modified Davies-Bouldin Index, Calinski-Harabasz Index and Score Function. In the case of soft partitioning, the following measures were used: Modified Partition Coefficient Index, Kwon Index, Improved Validation Index, Validity Function and PBMF Index. In both cases, a majority voting approach was used to evaluate a clustering methodology, which proved to be a robust approach, as the different points of view proposed by the different internal index methods in assessing the degree of cohesion and separation of clusters formed were considered.

The present work characterized the behaviors of the daily average flow pattern series present in the dataset through prototype of clusters formed in the different approaches. In the final combined model (see section 4.7.2), two clusters indicating inefficient water uses can be identified (Cluster 1 and Cluster 2.6). Cluster 1 is associated with predominantly nocturnal consumption for irrigation. This cluster is more associated with wet months than dry months, indicating that when programming irrigation systems the time of year is not considered. In the case of Cluster 2.6, there are maximum consumptions throughout the day, which also indicates that irrigation systems may be operating at warmer times of day, causing further evaporation if watering occurs during these periods. Both clusters are mostly associated with the South (Algarve) region and corrective measures should be taken in the areas affected by these clusters in order to program irrigation systems to take into account the time of year and the time of day.

For future developments the following analyzes should be considered:

- Dataset clustering analysis using the best performing clustering methods considering only the nighttime period. Restricting nighttime period, may enable the identification of clusters associated with leakage events in the water supply systems;
- Comparison of results using various dataset normalization methodologies (e.g.: mean vs. median);
- Application of clustering methodologies with inelastic distance measurements and a previous dimensional reduction of the dataset using PCA. Comparison of the results obtained with the clustering methods with elastic distance measures presented in this dissertation;
- Application of the methodology proposed in this dissertation to a dataset composed of time series with flow and pressure data, in order to obtain clusters that best represent the behaviors present in water distribution systems.

Appendix A

Clustering models with inelastic distance measures

A.1 K-means Clustering

In this section we will analyze a clustering model using the K-means approach (see section 3.5.2) with the following components:

- **Distance measure**: Euclidean (see section 3.6.1);
- **Prototype**: Mean (see section 3.7.1).

This model is the application of a classic K-means model with Euclidean distance and without using time comparison windows.

A.1.1 Clustering model internal index evaluation

Figure A.1 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

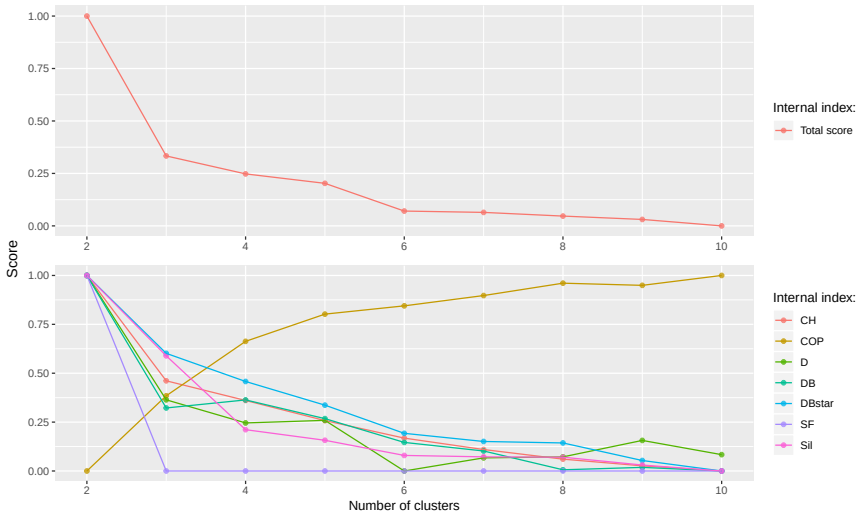

Figure A.1: Internal index evaluation for 1^{st} iteration set of K-Means Clustering.

Figure A.1 shows that the best result (Total score) was with the formation of 2 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure A.2 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

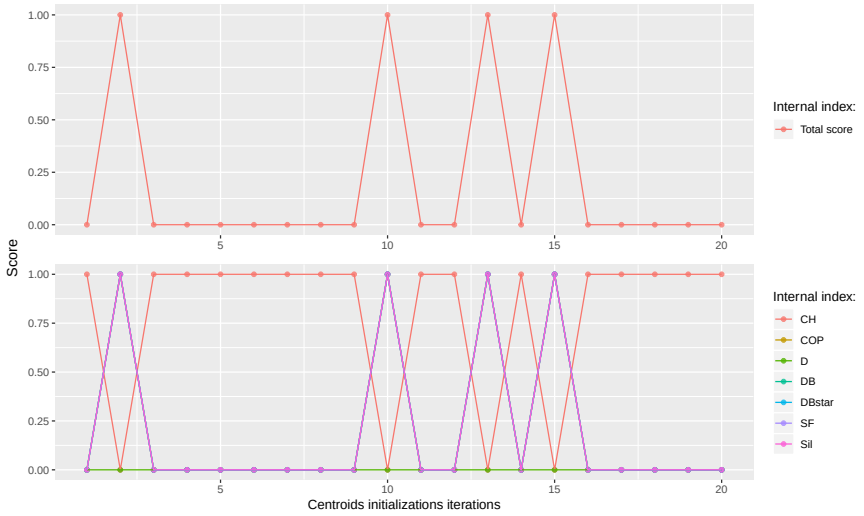

Figure A.2: Internal index evaluation for 2^{nd} iteration set of K-Means Clustering.

Figure A.2 shows that the 2^{nd}, 10^{th}, 13^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation.

In the next section the 2^{nd} iteration clustering model with the formation of 2 clusters will be analyzed.

A.1.2 Clustering model characterization

Figure A.3 shows the visualization of the clusters formed by the model according to the first 3 principal components:

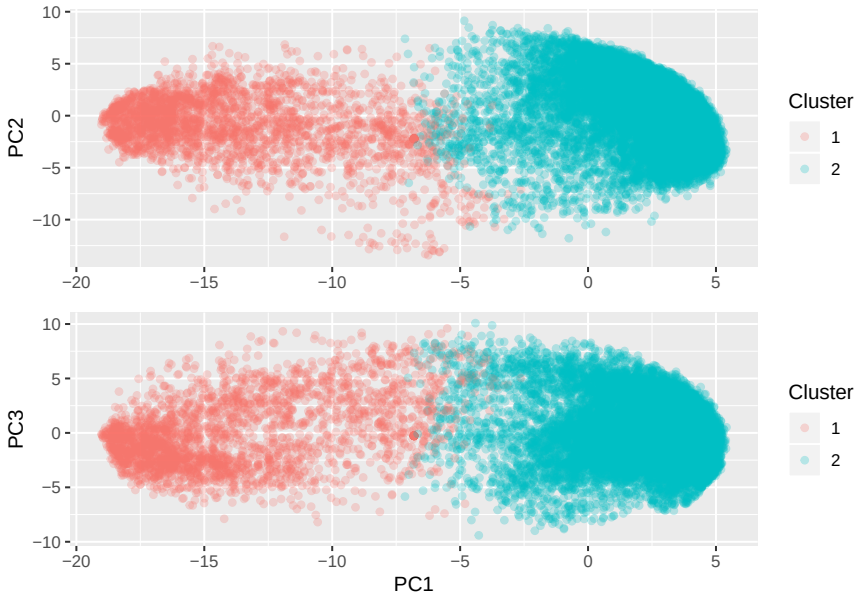

Figure A.3: Clusters formed through the k-Means model visualized through the 3 principal components of PCA.

In Figure A.3 it is possible to verify that most of the daily patterns belong to Cluster 2. It is also observed that in the zone in the neighborhood of the value corresponding to -5 of the Principal Component 1, there is no clear distinction between the two clusters formed. Another relevant aspect is that the distinction between the two clusters is made according only to the principal component 1.

Figure A.4 shows the respective centroids of the clusters formed:

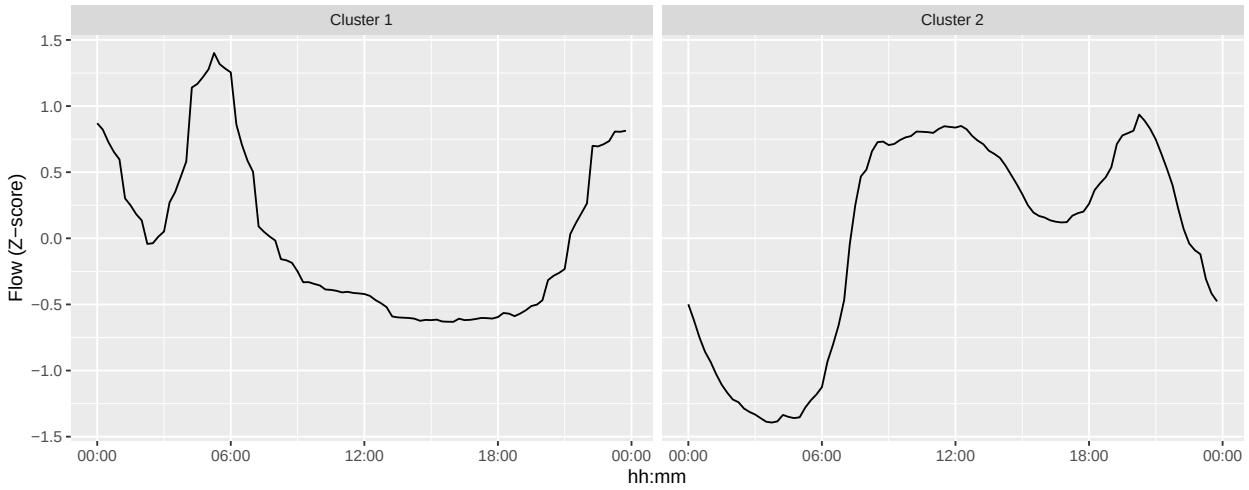

Figure A.4: k-Means model centroids.

Cluster 2 centroid has periods of higher consumption in the morning, lunch period and dinner period. In contrast, the centroid of cluster 1 has the periods of highest consumption occurring in the night period, the maximum values of flow are reached at midnight and also at 5:00.

From the previous graph we can see that the k-means algorithm with the formation of 2 clusters, has learned to group the patterns taking into account the fact that the patterns present higher consumption in the night period or in the daytime period.

Figure A.5 shows the size of each of the clusters formed:

Figure A.5: k-Means model clusters sizes.

Figure A.5 shows that most of the daily series belong to Cluster 2, and Cluster 1 presents only 2718 daily flow series. Indicating that most daily series have predominantly peak flows during the daytime period.

Figure A.6 evaluates the degree of membership of each of the annual series to the formed clusters:

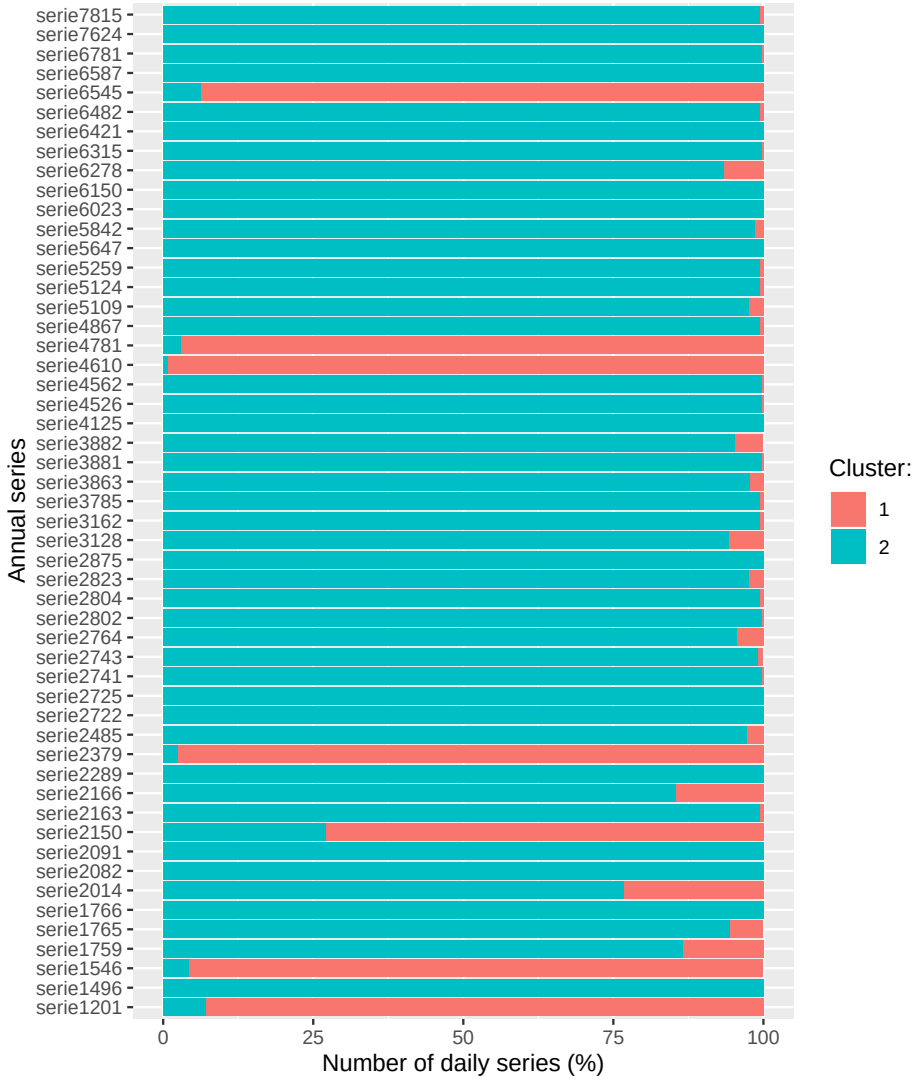

Figure A.6: k-Means model annual series membership.

In Figure A.6 it is observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201.

Table A.1 shows a set of statistical characteristics of the clusters formed:

Table A.1: k-Means model clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.82	7.28
Median	10.80	19.25
Mean	20.39	46.38
3rd Qu.	23.38	58.61
Max.	981.25	1207.00
IQR	18.56	51.33

Figure A.7 identifies the influence of weekend or holiday days have on the formation of clusters:

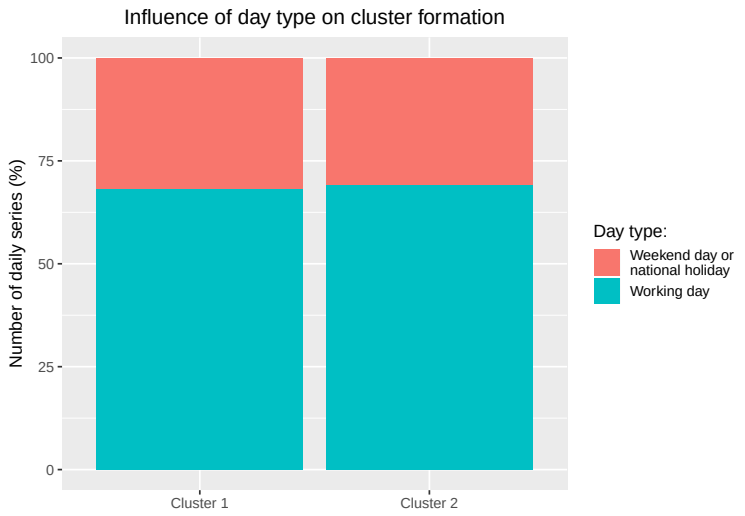

Figure A.7: k-Means model influence of day typology on the formation of clusters.

As can be seen from Figure A.7, the percentage of weekends and holidays is around 31% in both clusters. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure A.8 allows identifying the influence of day typology in each annual series by cluster type. As can be seen from Figure A.8, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure A.7, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day.

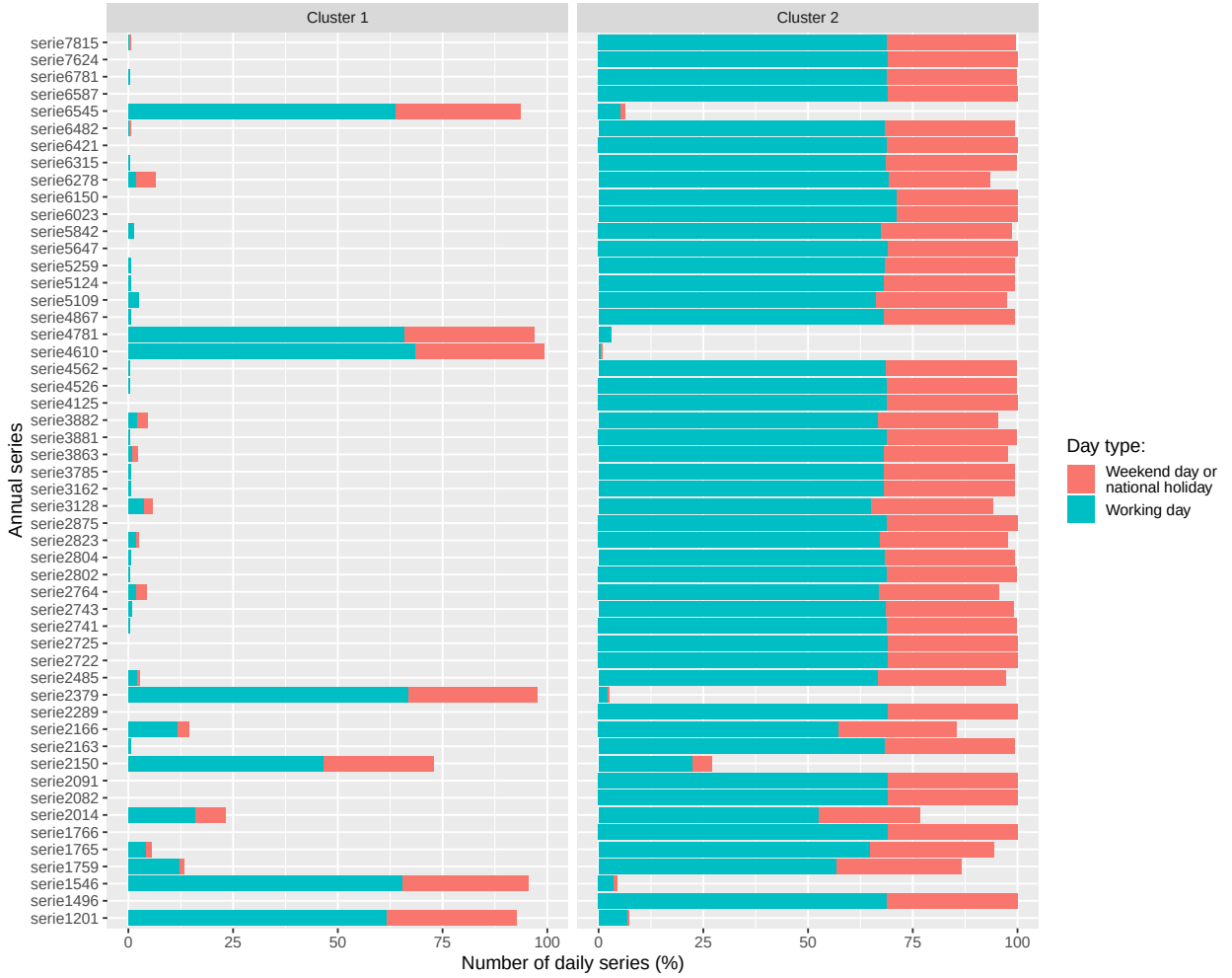

Figure A.8: k-Means model influence of day typology on each series by clusters.

A.2 Hierarchical Clustering

In this section we will analyze a clustering model using the Hierarchical approach with the buttom-up strategy and complete-linkage method (see section 3.5.1). The model will also incorporate the following components:

- **Distance measure**: Euclidean (see section 3.6.1);
- **Prototype**: Mean (see section 3.7.1).

A.2.1 Clustering model internal index evaluation

Figure A.9 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

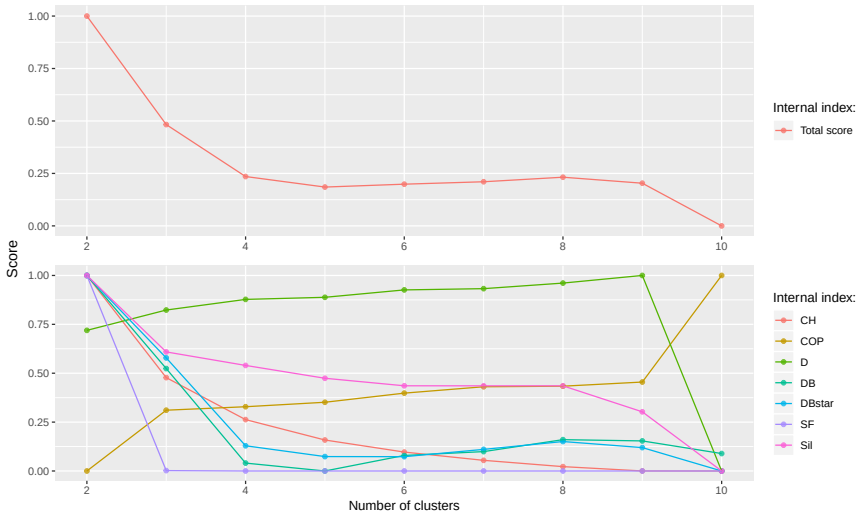

Figure A.9: Internal index evaluation for 1^{st} iteration set of Hierarchical Clustering.

Figure A.9 shows that the best result (Total score) was with the formation of 2 clusters. Since the hierarchical clustering model does not require centroid initialization (see section 3.5.1), there is no need to perform an analysis of the best centroid initialization (2^{nd} iteration set) as it did for K-means clustering (see section A.1). In the next section the clustering model with the formation of 2 clusters will be analyzed.

A.2.2 Clustering model characterization

Figure A.10 shows the visualization of the clusters formed by the model according to the first 3 principal components.

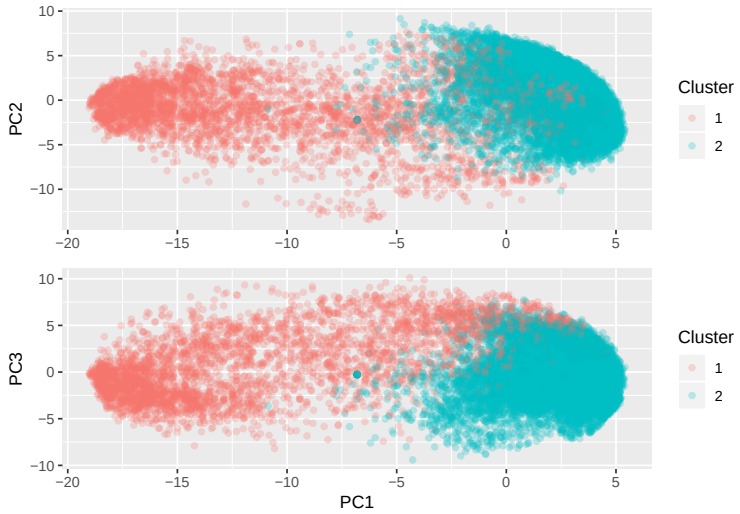

Figure A.10: Clusters formed through the Hierarchical model visualized through the 3 principal components of PCA.

Through Figure A.10 it is verified that Cluster 1 tends to negative zones according to the principal component 1 and Cluster 2 tends to the positive zones according to this component. The projection according to the principal components 1 and 3 gives a better distinction of the clusters formed than the projection according to the principal components 1 and 2.

Comparing with the projection according to the principal components obtained with the K-means model (see section A.1), the hierarchical model presents worse results, because the distinction between the two clusters is not so evident in the lower density zones.

Figure A.11 shows the respective centroids of the clusters formed:

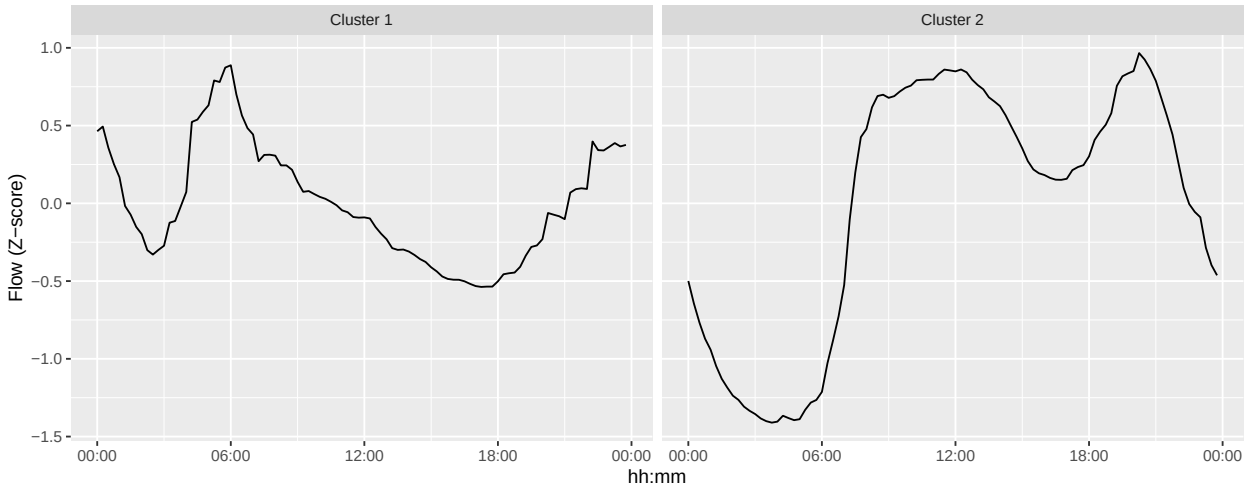

Figure A.11: Hierarchical model centroids.

Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around midnight and the second peak of consumption occurs around 06:00. The periods of minimum consumption occur at 2:30 and around 18:00.

In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 16:30. The absolute minimum consumption for Cluster 2 occurs around 4:00.

From the previous graph we can see that the hierarchical clustering algorithm with the formation of 2 clusters, has learned to group the patterns taking into account the fact that the patterns present higher consumption in the night period or in the daytime period.

Figure A.12 shows the size of each of the clusters formed. This Figure shows that most of the daily series belong to Cluster 2, and Cluster 1 presents only 3869 daily flow series. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure A.12: Hierarchical model clusters sizes.

Figure A.13 evaluates the degree of membership of each of the annual series to the formed clusters:

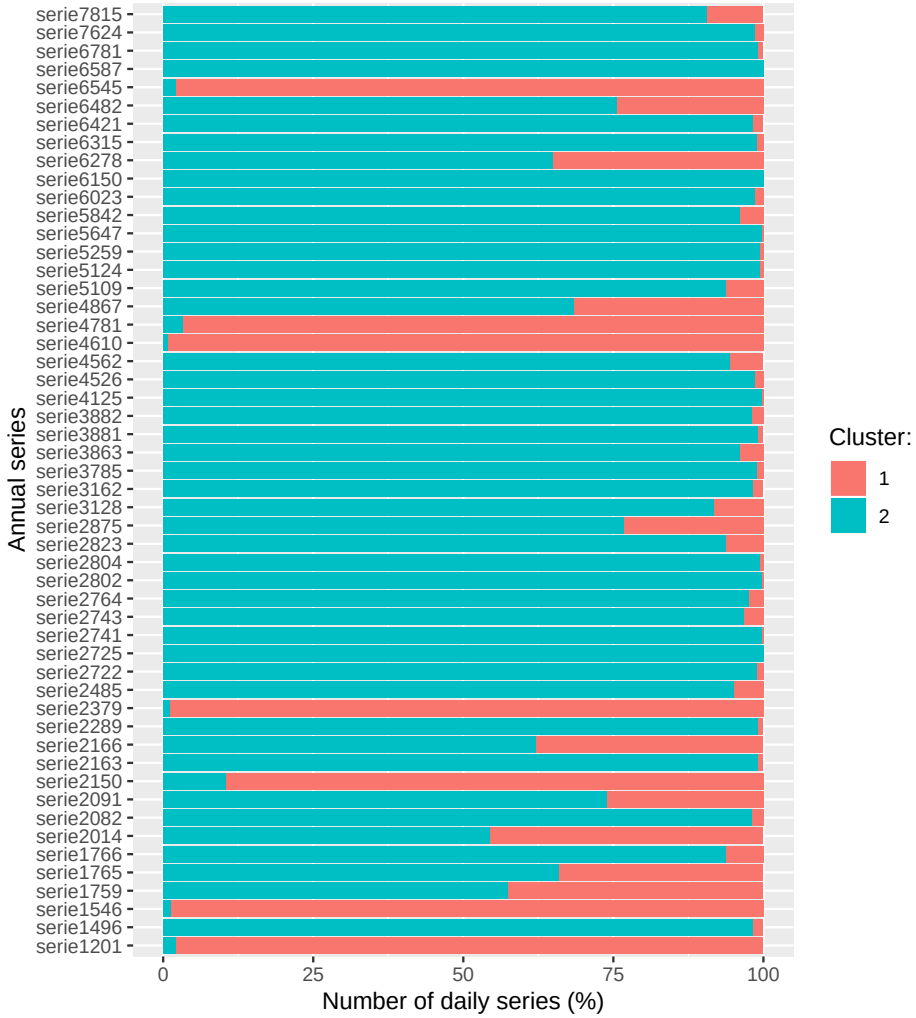

Figure A.13: Hierarchical model annual series membership.

In Figure A.13 it is observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the K-means clustering method, since most clusters belong to a pattern with predominantly diurnal consumption and the patterns identified with predominantly nocturnal consumption are the same in the two approaches.

Table A.2 shows a set of statistical characteristics of the clusters formed:

Table A.2: Hierarchical model clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	5.60	7.16
Median	14.84	18.09
Mean	29.49	46.03
3rd Qu.	36.00	57.73
Max.	981.25	1207.00
IQR	30.40	50.57

Figure A.14 identifies the influence of weekend or holiday days have on the formation of clusters:

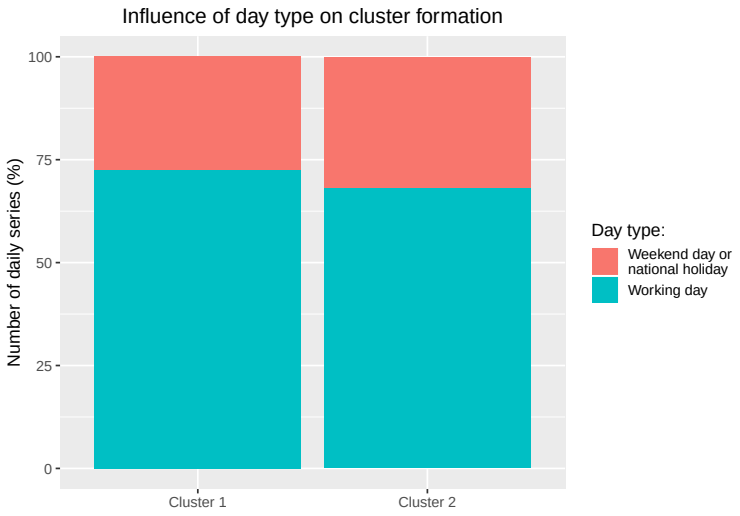

Figure A.14: Hierarchical model influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 2 and around 27% for Cluster 1. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure A.15 allows identifying the influence of day typology in each annual series by cluster type:

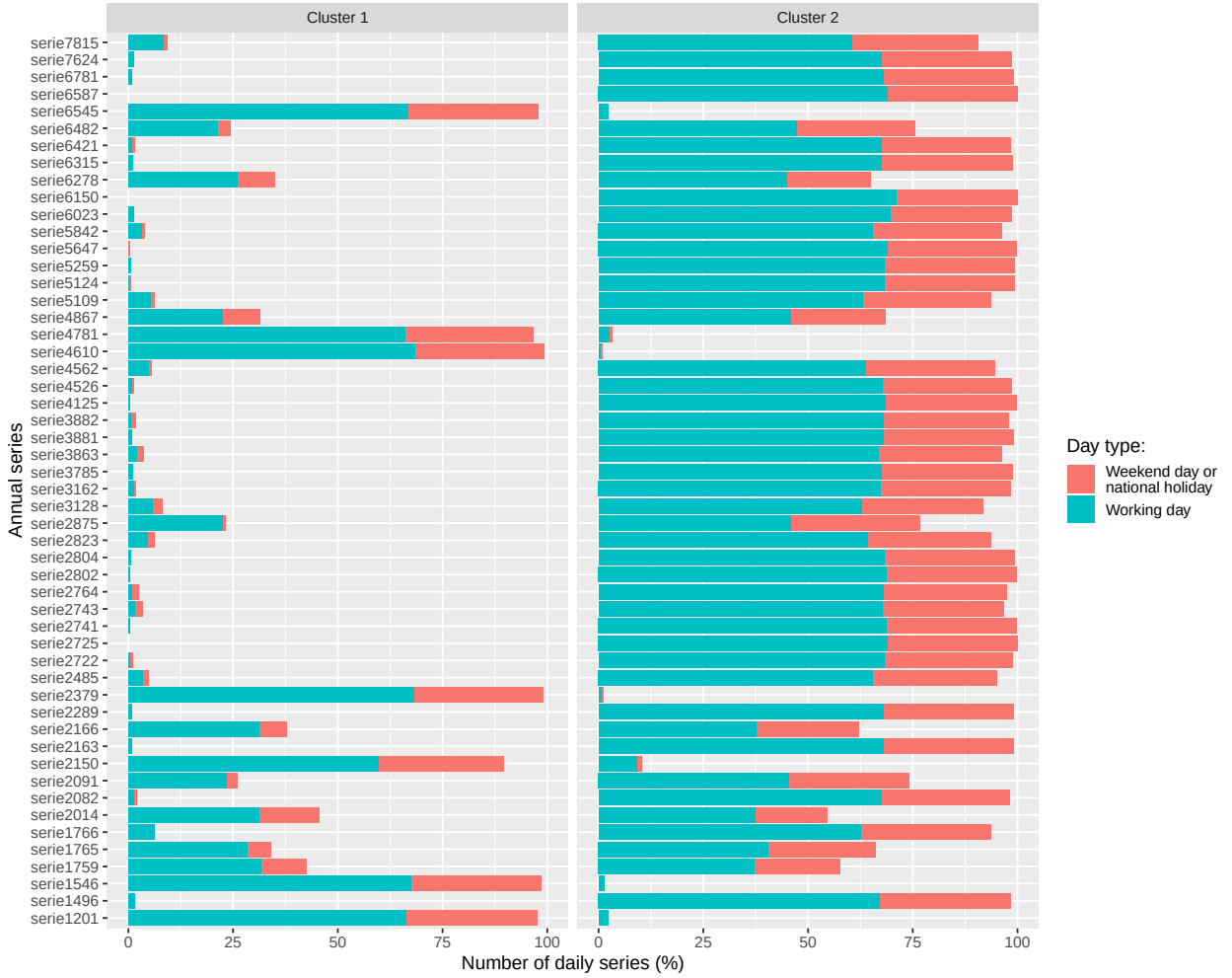

Figure A.15: Hierarchical model influence of day typology on each series by clusters.

As can be seen from Figure A.15, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure A.14, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the K-means model (see section A.1).

A.3 Fuzzy Clustering

In this section we will analyze a clustering model using the Fuzzy approach (see section 3.5.4) with the following components:

- **Distance measure**: Euclidean (see section 3.6.1);
- **Prototype**: Fuzzy-based (see section 3.7.5).

A.3.1 Clustering model internal index evaluation

Figure A.16 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

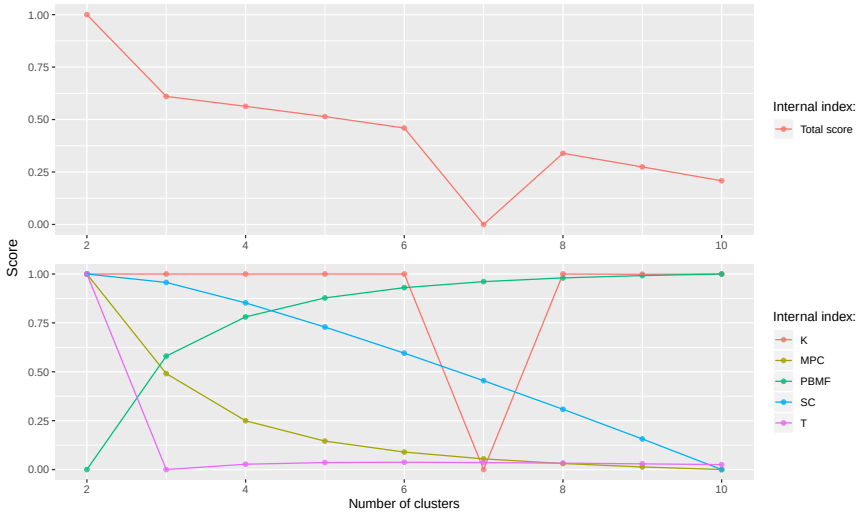

Figure A.16: Internal index evaluation for 1^{st} iteration set of Fuzzy Clustering.

Figure A.16 shows that the best result (Total score) was with the formation of 2 clusters.

Fuzzy clustering model does not require centroid initialization (see section 3.5.4), there is no need to perform an analysis of the best centroid initialization (2^{nd} iteration set) as it did for K-means clustering (see section A.1).

In the next section the clustering model with the formation of 2 clusters will be analyzed.

A.3.2 Clustering model characterization

Figure A.17 shows the visualization of the clusters formed by the model according to the first 3 principal components:

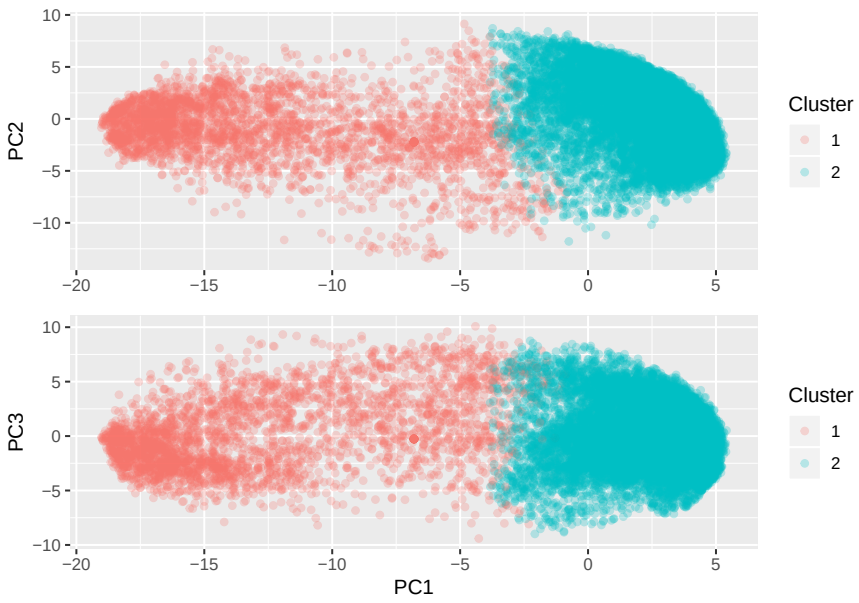

Figure A.17: Clusters formed through the Fuzzy model visualized through the 3 principal components of PCA.

In Figure A.17 it is evident a clear separation between the two formed clusters. This feature demonstrates that in the space formed by the first 3 principal components the use of this soft partition algorithm in the formation of two clusters achieves a sharper separation than the hard partition algorithms (Hierarchical and k-Means).

For this soft partition algorithm, as in previous hard partition algorithms, cluster 1 tends to negative zones according to the main component 1 and cluster 2 tends to positive zones according to this component.

Figure A.18 shows the respective centroids of the clusters formed:

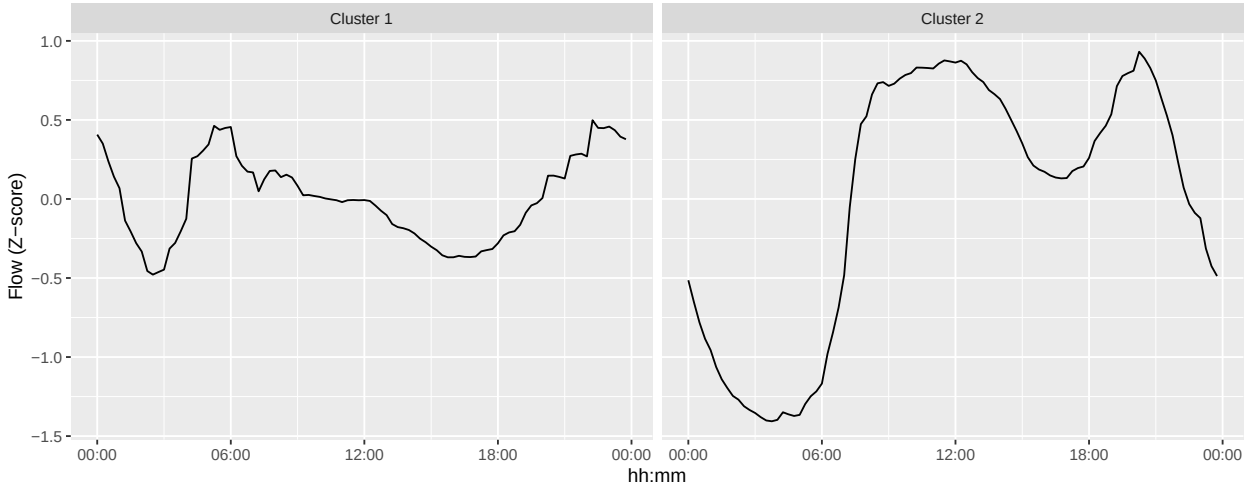

Figure A.18: Fuzzy model centroids.

The centroids representing the clusters according to the Fuzzy prototype definition represent a weighted average considering the weights of the membership matrix of the Fuzzy c-means algorithm. Consequently the presented centroids represent a form of a weighted pattern and not of a real pattern of dataset.

Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around 22:00 and the second peak of consumption occurs around 06:00. The periods of minimum consumption occur at 2:30 and around 17:00.

In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 17:00. The absolute minimum consumption for Cluster 2 occurs around 04:00.

Figure A.19 shows the size of each of the clusters formed:

Figure A.19: Fuzzy model clusters sizes.

Figure A.19 shows that most of the patterns belong to Cluster 2, and Cluster 1 presents only 3208 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure A.20 evaluates the degree of membership of each of the annual series to the formed clusters:

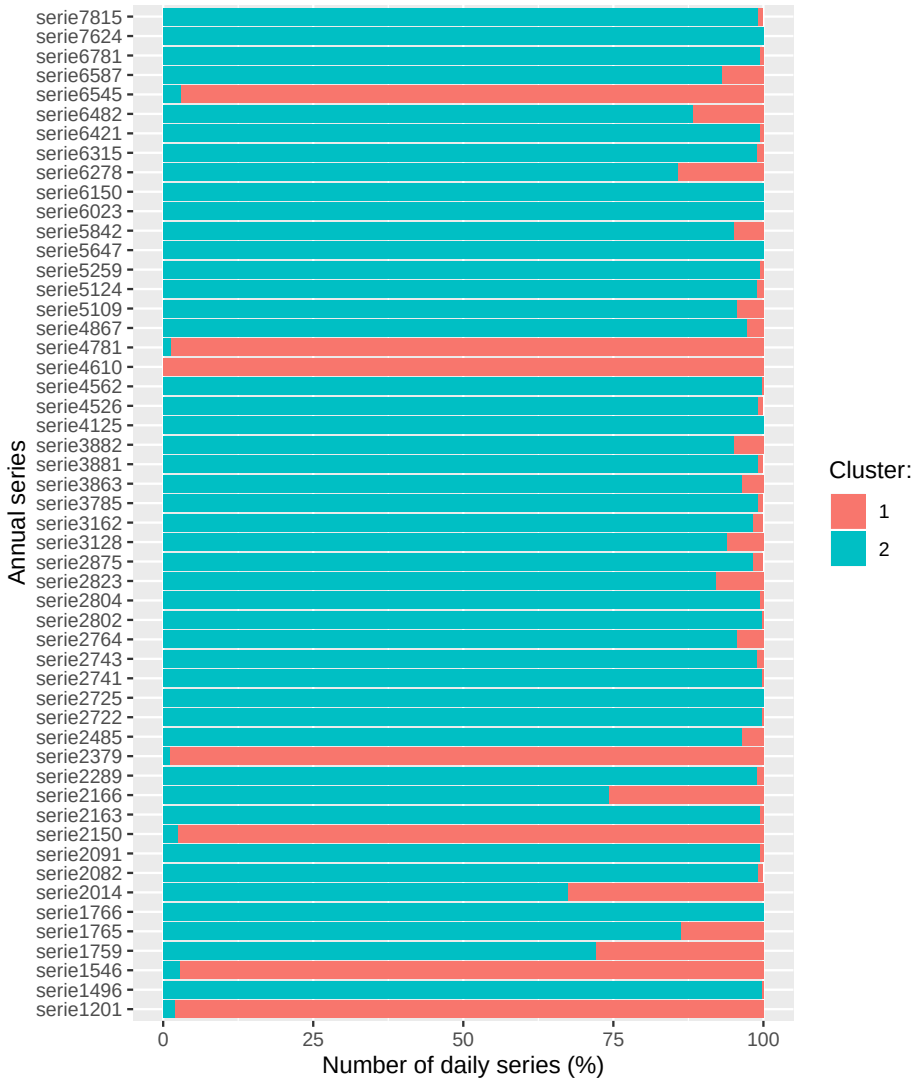

Figure A.20: Fuzzy model annual series membership.

In Figure A.20 it is observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous hard partitioning methods (k-means and hierarchical clustering), since most clusters belong to a daily series with predominantly diurnal consumption and the patterns identified with predominantly nocturnal consumption are the same in the two approaches.

Table A.3 shows a set of statistical characteristics of the clusters formed:

Table A.3: Fuzzy model clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.89	7.32
Median	11.70	19.11
Mean	23.43	46.56
3rd Qu.	26.60	58.66
Max.	981.25	1207.00
IQR	21.71	51.34

Figure A.21 identifies the influence of weekend or holiday days have on the formation of clusters:

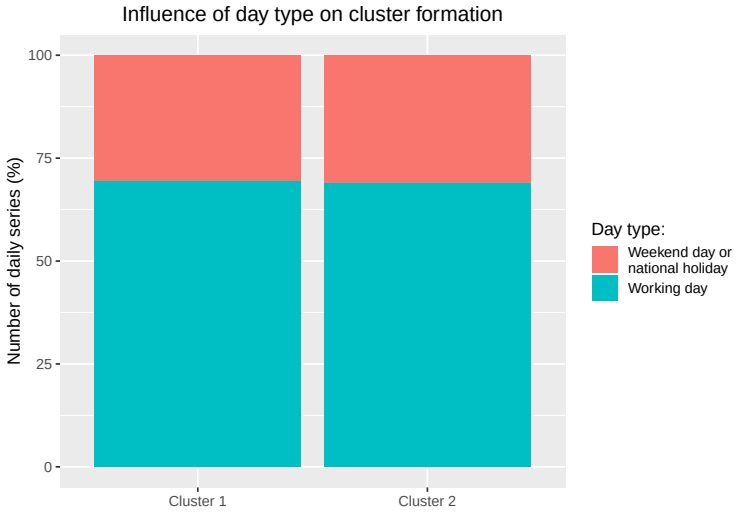

Figure A.21: Fuzzy model influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 2 and Cluster 1. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure A.22 allows identifying the influence of day typology in each annual series by cluster type:

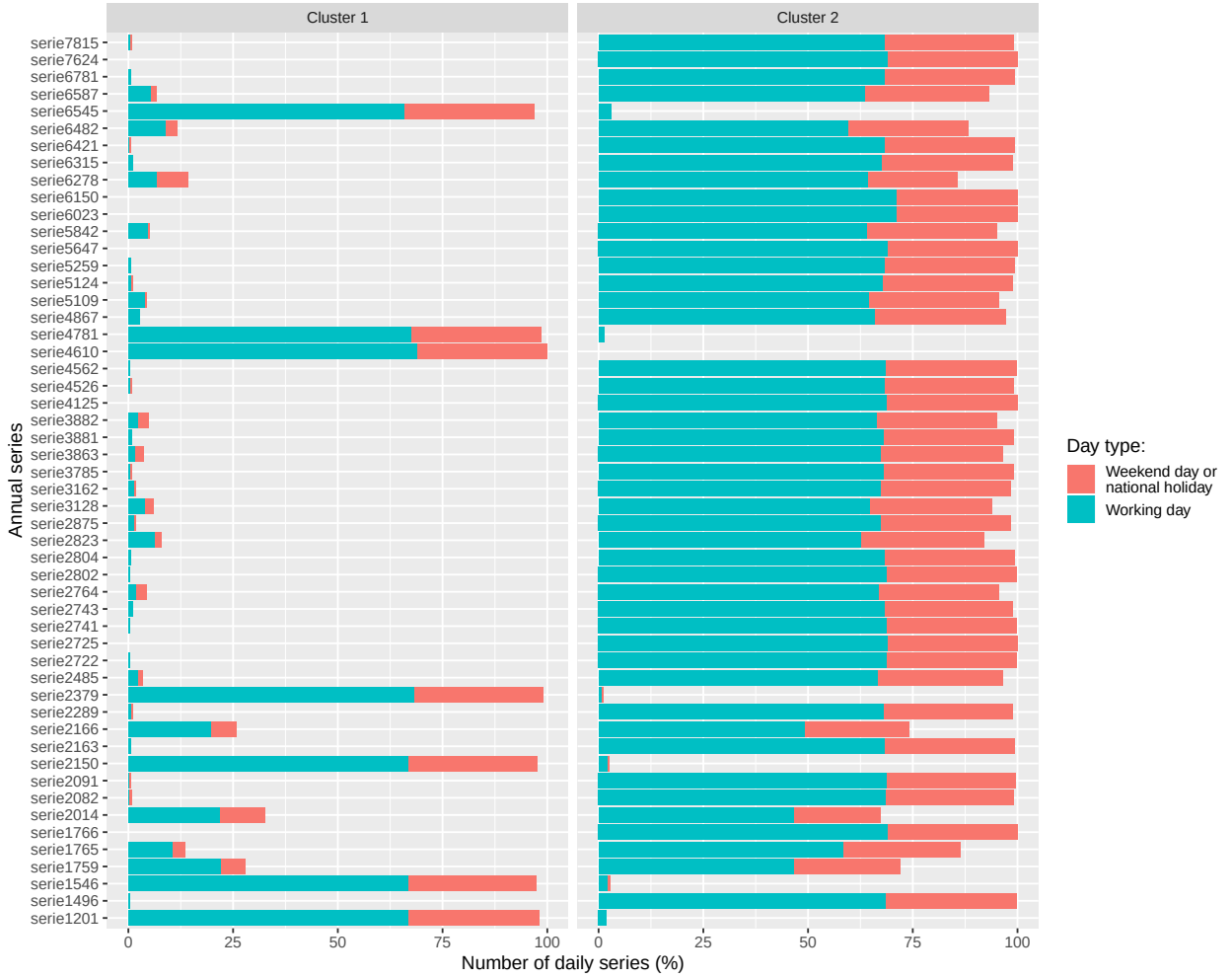

Figure A.22: Fuzzy model influence of day typology on each series by clusters.

As can be seen from Figure A.22, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure A.21, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the previous clustering methods with formation of 2 cluster.

Appendix B

Clustering models with elastic distance measures

B.1 Partitional Clustering with DTW, Mean prototype and 15 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: Mean (see section 3.7.1);
- Comparison time window: 15 minutes (see section 3.6.2).

B.1.1 Clustering model internal index evaluation

Figure B.1 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

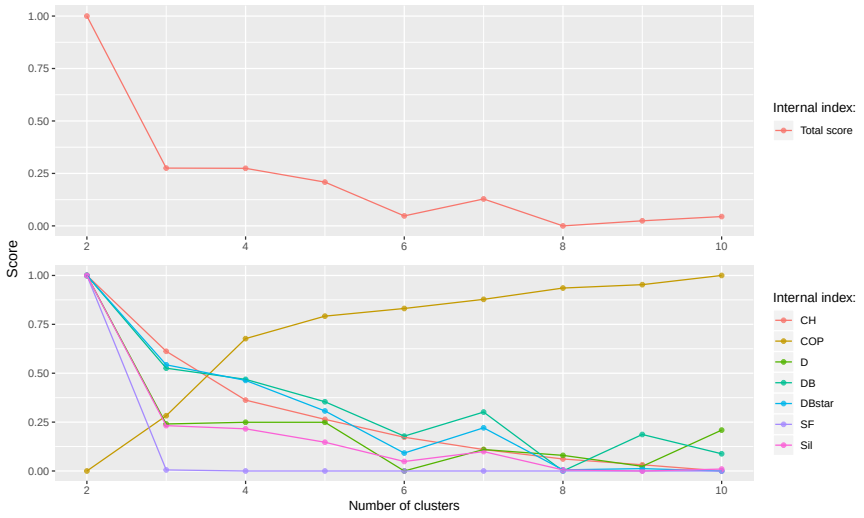

Figure B.1: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, Mean Prototype and 15 minutes time window.

Figure B.1 shows that the best result (Total score) was with the formation of 2 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.2 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

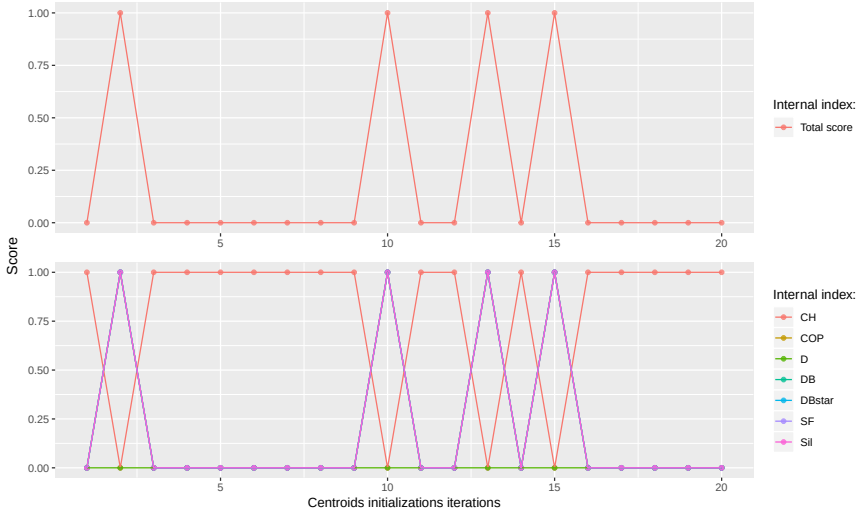

Figure B.2: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, Mean Prototype and 15 minutes time window.

Figure B.2 shows that the 2^{nd}, 10^{th}, 13^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 2^{nd} iteration clustering model with the formation of 2 clusters will be analyzed.

B.1.2 Clustering model characterization

Figure B.3 shows the visualization of the clusters formed by the model according to the first 3 principal components:

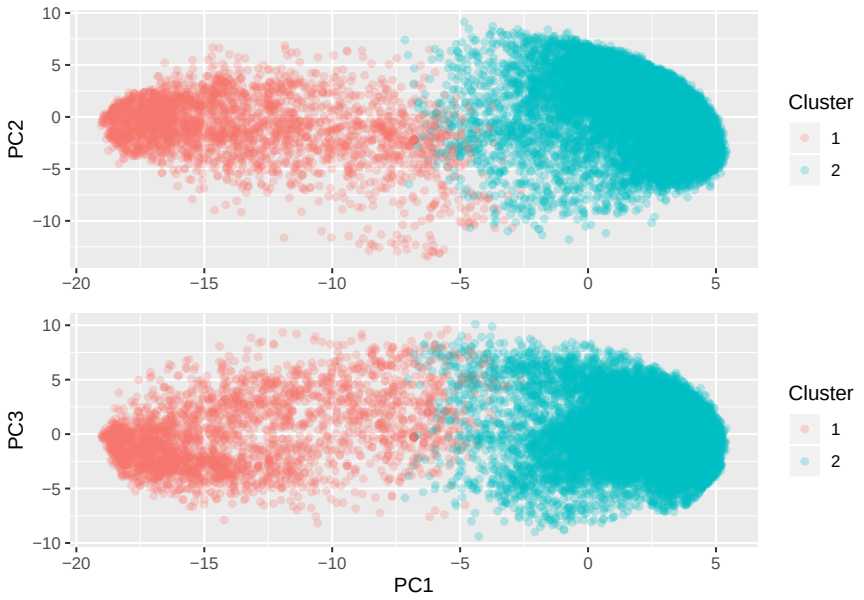

Figure B.3: Clusters formed through the Partition Clustering model with DTW distance, Mean prototype and 15m window visualized through the 3 principal components of PCA.

In Figure B.3 it is evident a clear separation between the two formed clusters. This feature demonstrates that in the space formed by the first 3 principal components, the use of this partition algorithm with euclidean distance and dynamic time warping in the formation of two clusters achieves a sharper separation than the hard partition algorithms with inelastic euclidean distance (hierarchical and k-means). Compared to soft partitioning methods, the results allows to conclude that the performance is slightly worse at the level of cluster separation.

For this partiton algorithm with DTW 15 minutes window constraint and mean centroid, as in previous algorithms, cluster 1 tends to negative zones according to the principal component 1 and cluster 2 tends to positive zones according to this component.

Figure B.4 shows the respective centroids of the clusters formed:

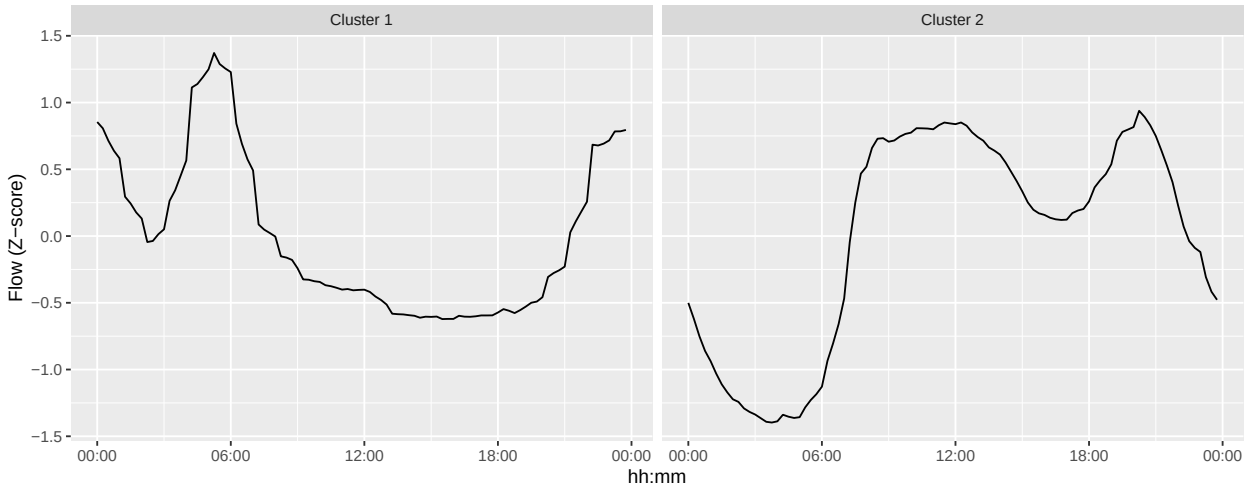

Figure B.4: Partition Clustering model with DTW distance, Mean prototype and 15m window centroids.

Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around midnight and the second peak of consumption occurs around 06:00. The periods of minimum consumption occur at 2:30 and around 16:00.

In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 17:00. The absolute minimum consumption for Cluster 2 occurs around 04:00.

Figure B.5 shows the size of each of the clusters formed:

Figure B.5: Partition Clustering model with DTW distance, Mean prototype and 15m window clusters sizes.

Figure B.5 shows that most of the patterns belong to Cluster 2, and Cluster 1 presents only 2769 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure B.6 evaluates the degree of membership of each of the annual series to the formed clusters:

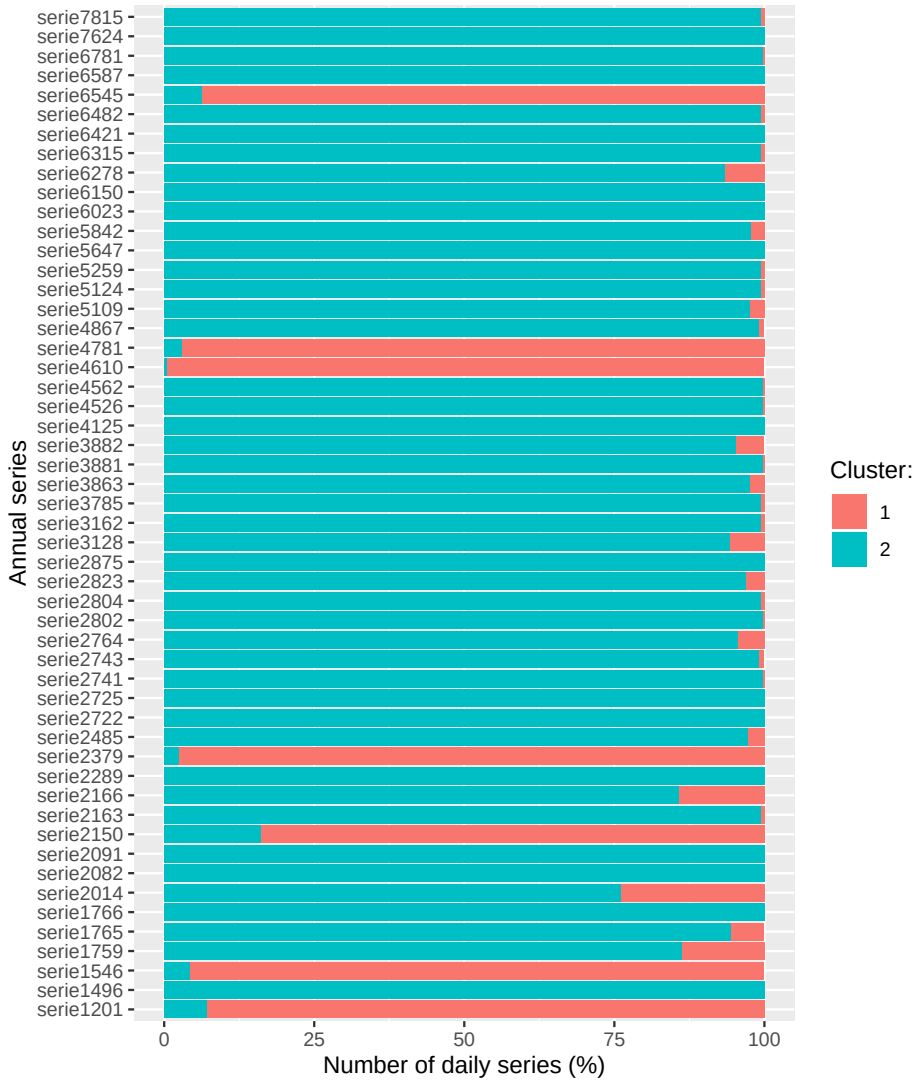

Figure B.6: Partition Clustering model with DTW distance, Mean prototype and 15m window annual series membership.

In Figure B.6 it is observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous hard partitioning methods (k-means and hierarchical clustering), since most clusters belong to a pattern with predominantly diurnal consumption and the patterns identified with predominantly nocturnal consumption are the same in the two approaches.

Table B.1 shows a set of statistical characteristics of the clusters formed:

Table B.1: Partition Clustering model with DTW distance, Mean prototype and 15m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.70	7.34
Median	10.62	19.37
Mean	20.13	46.50
3rd Qu.	23.15	58.80
Max.	981.25	1207.00
IQR	18.45	51.46

Figure B.7 identifies the influence of weekend or holiday days have on the formation of clusters:

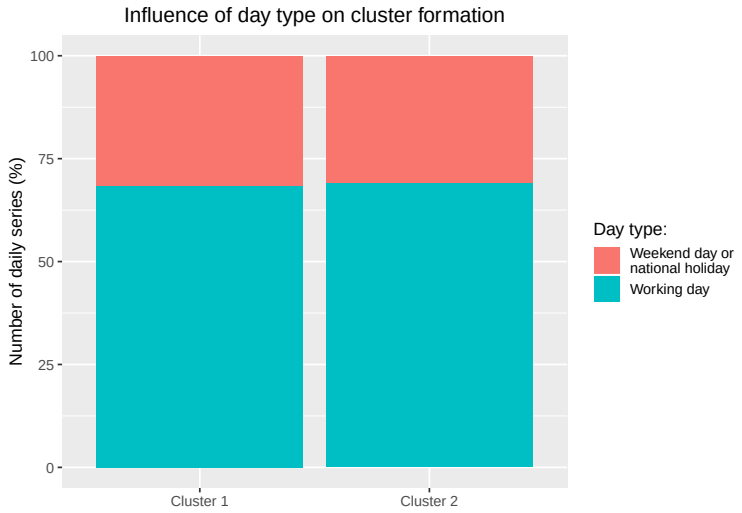

Figure B.7: Partition Clustering model with DTW distance, Mean prototype and 15m window influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 2 and Cluster 1. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure B.8 allows identifying the influence of day typology in each annual series by cluster type:

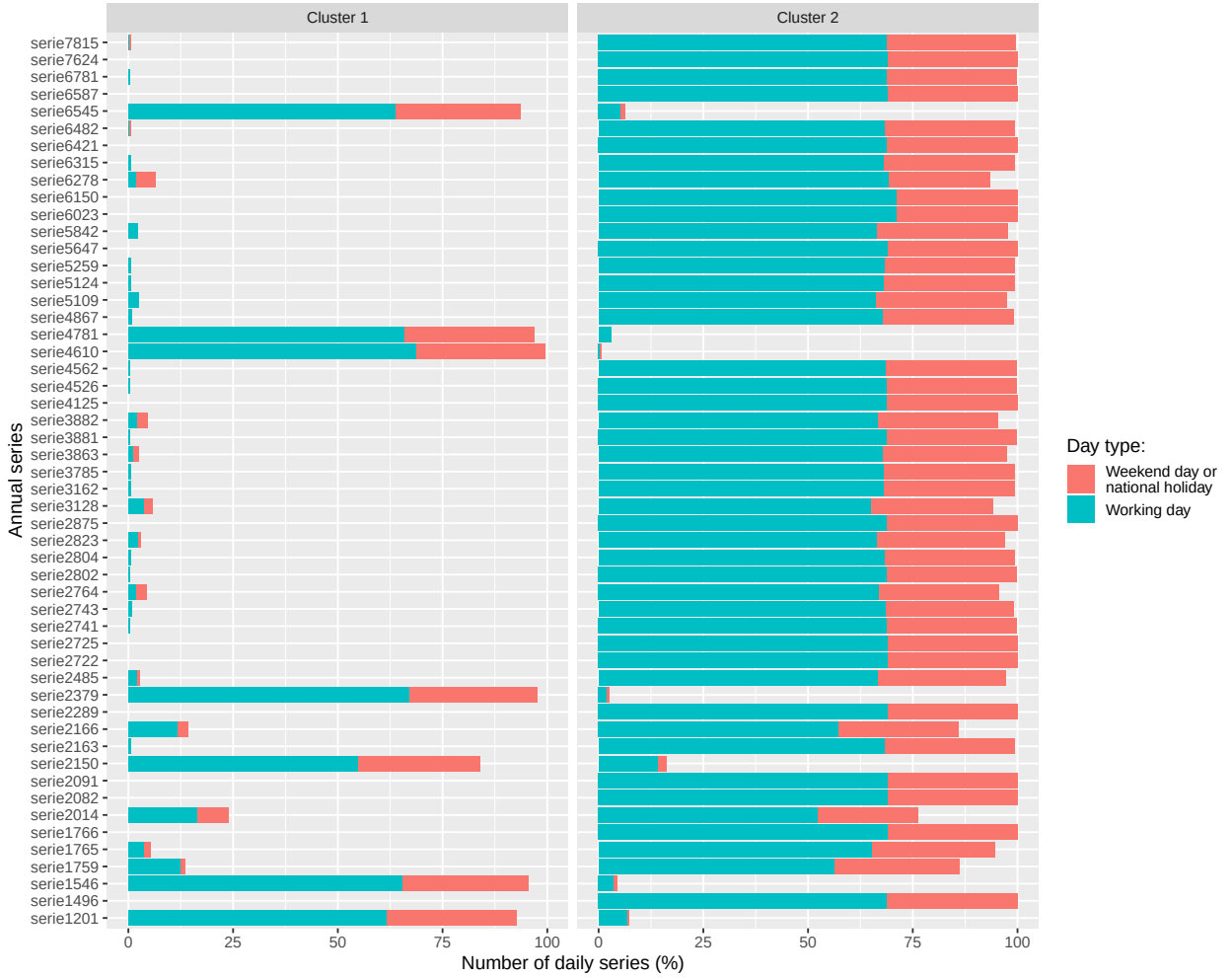

Figure B.8: Partition Clustering model with DTW distance, Mean prototype and 15m window influence of day typology on each series by clusters.

As can be seen from Figure B.8, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure B.7, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the previous clustering methods with formation of 2 clusters.

B.2 Partitional Clustering with DTW, Mean prototype and 30 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: Mean (see section 3.7.1);
- Comparison time window: 30 minutes (see section 3.6.2).

B.2.1 Clustering model internal index evaluation

Figure B.9 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

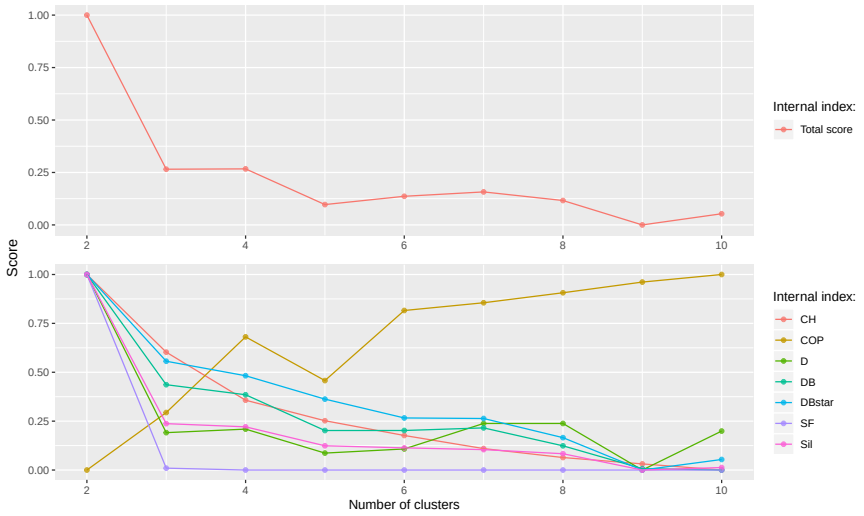

Figure B.9: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, Mean Prototype and 30 minutes time window.

Figure B.9 shows that the best result (Total score) was with the formation of 2 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.10 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

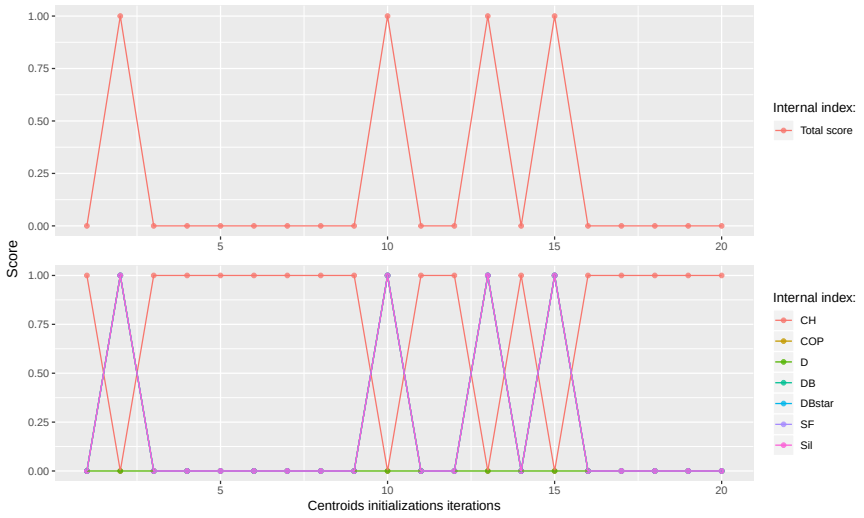

Figure B.10: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, Mean Prototype and 30 minutes time window.

Figure B.10 shows that the 2^{nd}, 10^{th}, 13^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 2^{nd} iteration clustering model with the formation of 2 clusters will be analyzed.

B.2.2 Clustering model characterization

Figure B.11 shows the visualization of the clusters formed by the model according to the first 3 principal components:

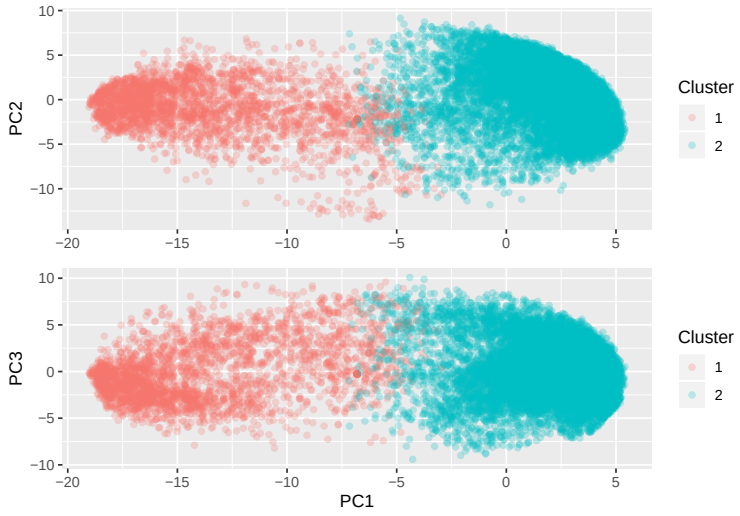

Figure B.11: Clusters formed through the Partition Clustering model with DTW distance, Mean prototype and 30m window visualized through the 3 principal components of PCA.

Through Figure B.11 it is verified that the separation between clusters is identical to that of the partition model with DTW , mean prototype with 15 minutes. For this partiton algorithm, as in previous clustering models, cluster 1 tends to negative zones according to the main component 1 and cluster 2 tends to positive zones according to this component.

Figure B.12 shows the respective centroids of the clusters formed:

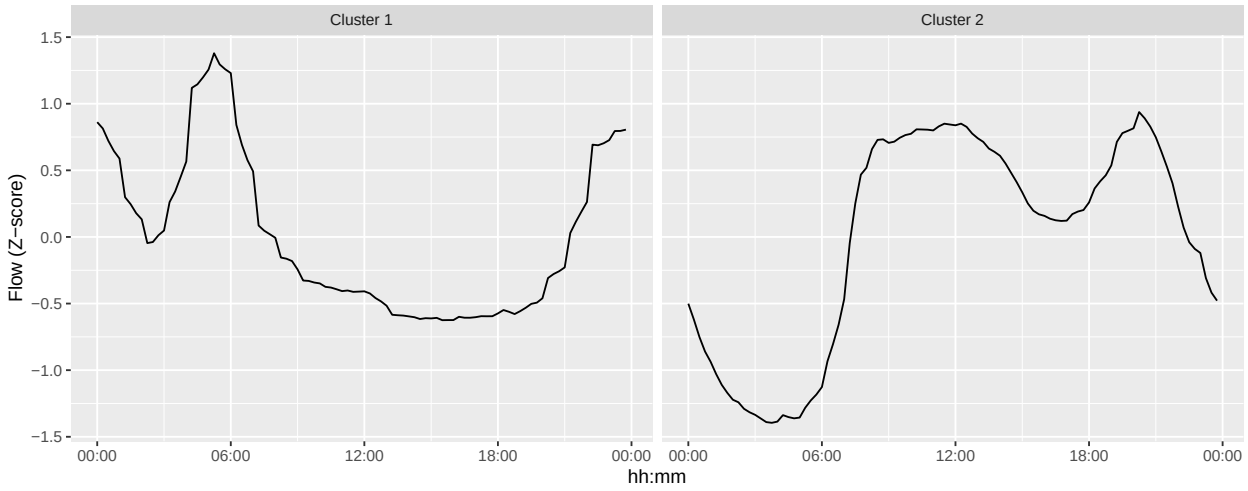

Figure B.12: Partition Clustering model with DTW distance, Mean prototype and 30m window centroids.

Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around midnight and the second peak of consumption occurs around 05:00. The periods of minimum consumption occur at 2:30 and around 16:00. In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 17:00. The absolute minimum consumption for Cluster 2 occurs around 04:00.

Figure B.13: Partition Clustering model with DTW distance, Mean prototype and 30m window clusters sizes.

Figure B.13 shows that most of the patterns belong to Cluster 2, and Cluster 1 presents only 2755 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure B.14 evaluates the degree of membership of each of the annual series to the formed clusters:

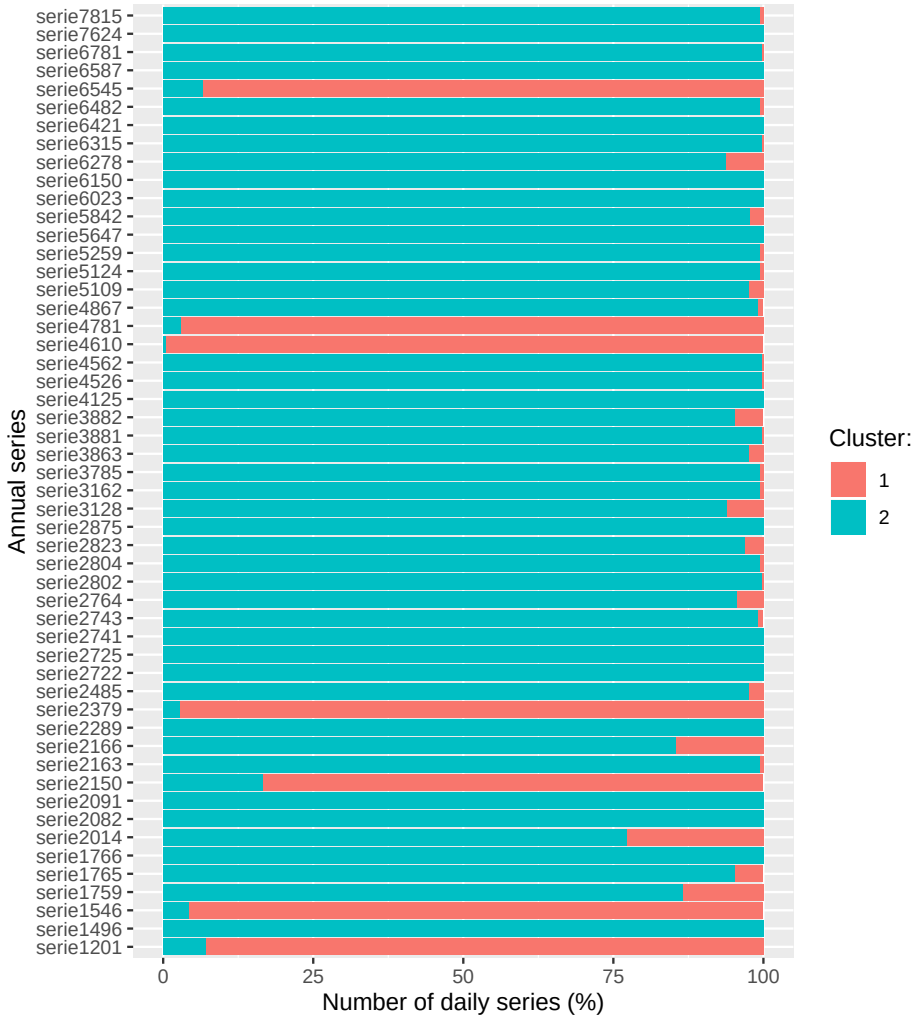

Figure B.14: Partition Clustering model with DTW distance, Mean prototype and 30m window annual series membership.

In Figure B.14 it is observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous hard partitioning methods (k-means and hierarchical clustering), since most clusters belong to a pattern with predominantly diurnal consumption and the patterns identified with predominantly nocturnal consumption are the same in the two approaches.

Table B.2 shows a set of statistical characteristics of the clusters formed:

Table B.2: Partition Clustering model with DTW distance, Mean prototype and 30m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.70	7.34
Median	10.60	19.36
Mean	20.11	46.49
3rd Qu.	23.12	58.78
Max.	981.25	1207.00
IQR	18.42	51.44

Figure B.15 identifies the influence of weekend or holiday days have on the formation of clusters:

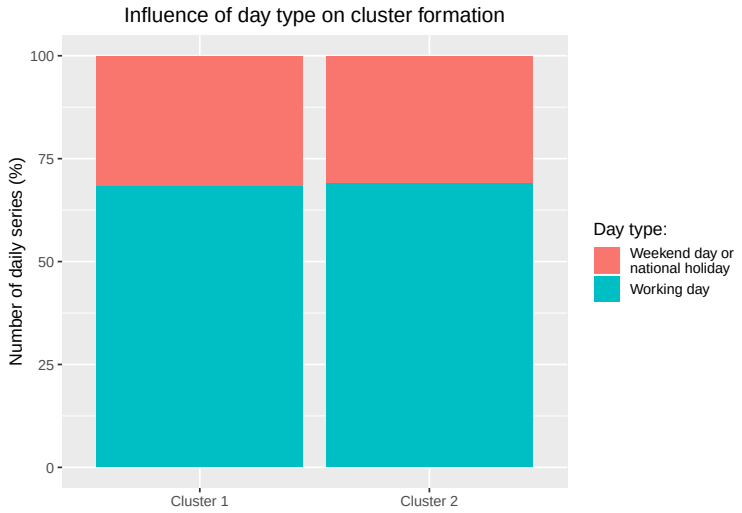

Figure B.15: Partition Clustering model with DTW distance, Mean prototype and 30m window influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 2 and Cluster 1. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure B.16 allows identifying the influence of day typology in each annual series by cluster type:

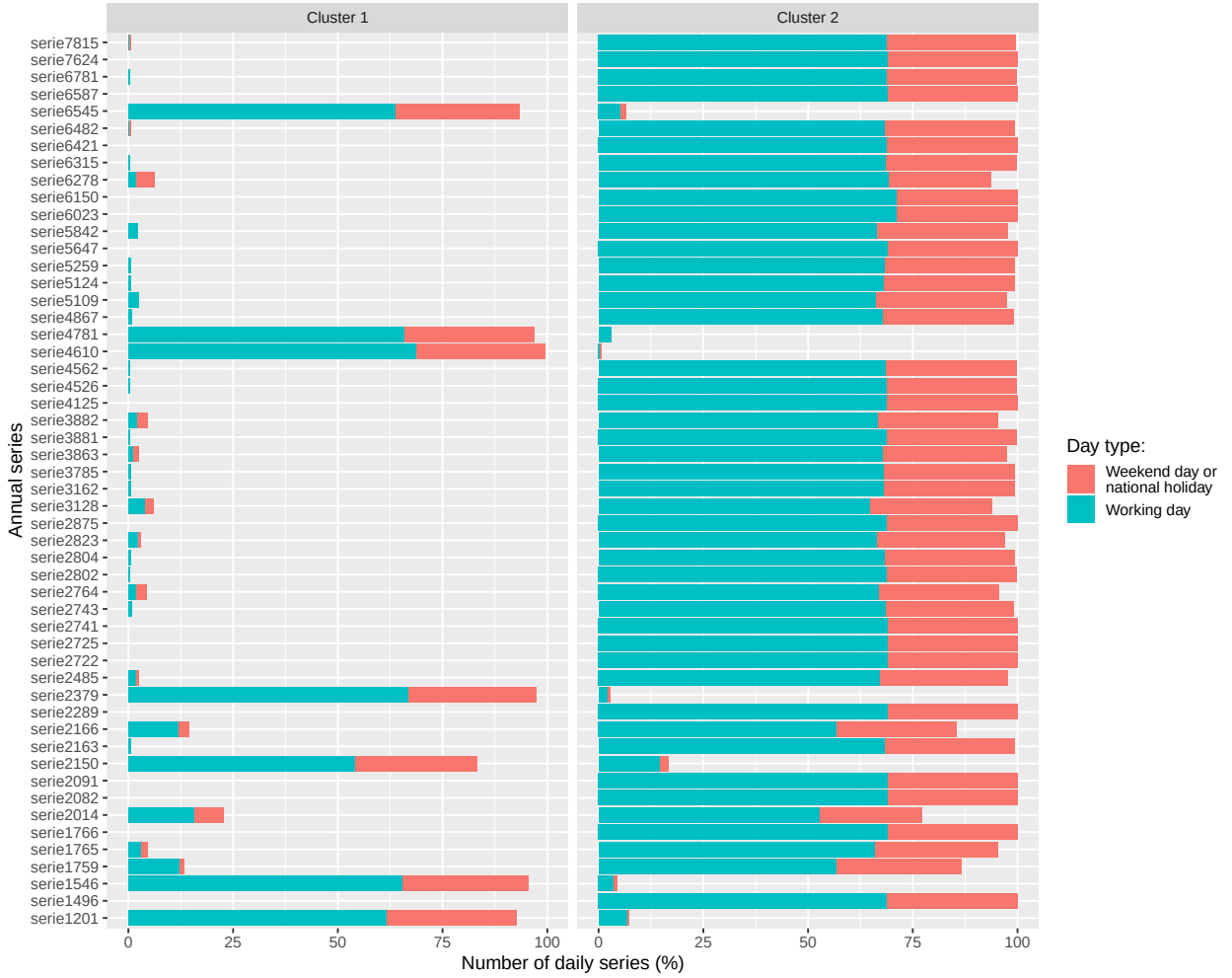

Figure B.16: Partition Clustering model with DTW distance, Mean prototype and 30m window influence of day typology on each series by clusters.

As can be seen from Figure B.16, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure B.15, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the previous clustering methods with formation of 2 clusters.

B.3 Partitional Clustering with DTW, PAM prototype and 30 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 30 minutes (see section 3.6.2).

B.3.1 Clustering model internal index evaluation

Figure B.17 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

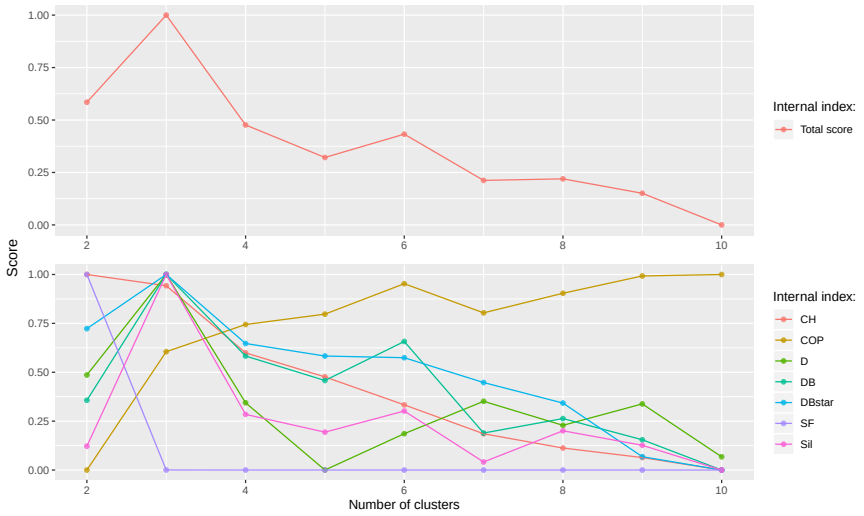

Figure B.17: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, PAM Prototype and 30 minutes time window.

Figure B.17 shows that the best result (Total score) was with the formation of 3 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.18 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

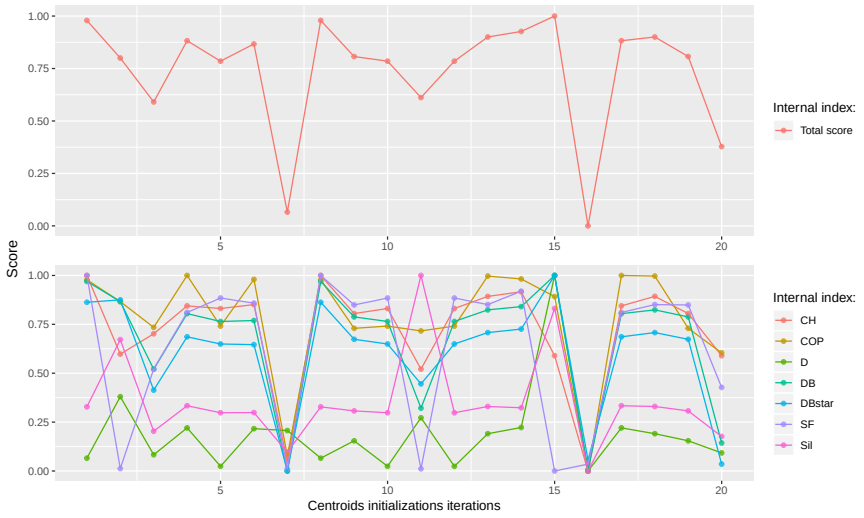

Figure B.18: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, PAM Prototype and 30 minutes time window.

Figure B.18 shows that the 1^{st}, 8^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 3^{rd} iteration clustering model with the formation of 3 clusters will be analyzed.

B.3.2 Clustering model characterization

Figure B.19 shows the visualization of the clusters formed by the model according to the first 3 principal components:

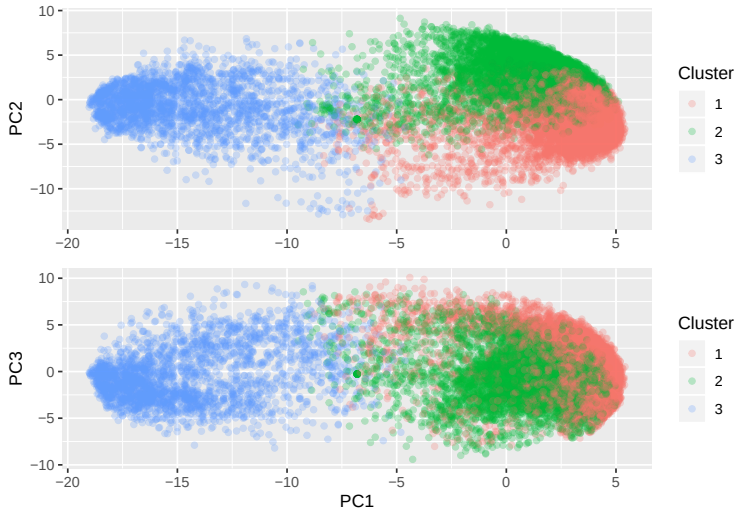

Figure B.19: Clusters formed through the Partition Clustering model with DTW distance, PAM prototype and 30m window visualized through the 3 principal components of PCA.

As can be seen from the figure the results are similar to the Partition model with DTW distance, PAM centroid and 15 minutes window contraint, there is a distinction between cluster 3 and the group formed by clusters 1 and 2, except in zones close to the value of -5 in the first principal component.

For clusters 1 and 2, the projection under principal components 1 and 2 allows to distinguish between the two groups except in areas close to the value of 0 in the principal component 2. Observing clusters 1 and 2 according to the projection on the principal components 1 and 3 it is not possible to clearly distinguish between the two clusters.

Figure B.20 shows the respective centroids of the clusters formed:

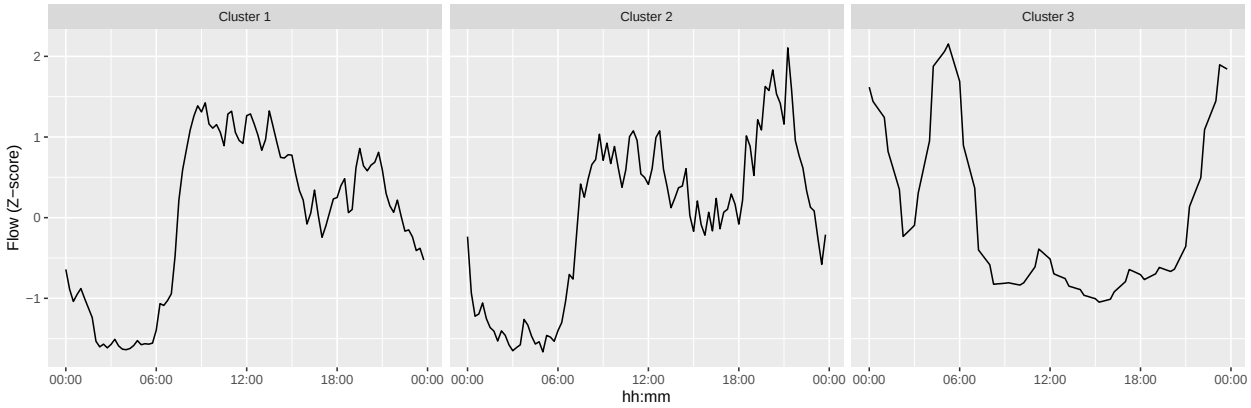

Figure B.20: Partition Clustering model with DTW distance, PAM prototype and 30m window centroids.

Clusters 1 and 2 present higher consumption peaks during the day period, while Cluster 3 shows higher consumption during the night time period.

Cluster 1 shows maximum consumption values at 09:00, 11:00, 12:00 and 14:30. This cluster presents a local minimum near 17:30 and a local maximum around 20:00. From this moment the consumption falls to the minimum value registered at 04:30. This behavior can be either associated with a weekend or working day pattern.

Cluster 2 has local maximum consumption peaks near 08:00, 11:00 and 12:00 and reaches a local minimum around 17:00. From this period consumption increases again until around 21:00 which is the maximum flow value. After this period the consumption drops back down to 03:00 which corresponds to the minimum value of consumption. This behavior represents a typical pattern for working day, given that it is much higher than cluster 1 near dinner, which is when people come home from work.

Although this model does not distinguish between weekend and working day patterns from predominantly daytime consumption patterns. This model can distinguish between daytime consumption patterns with maximum consumption near dinner time (Cluster 2) and predominantly daytime consumption patterns with maximum consumption between 06:00 and 14:30 (Cluster 1).

Cluster 3 shows peak consumption in the midnight period and in the period near 05:00 am. The predominance of this cluster by nocturnal consumption may be due to the use of water is predominantly associated with irrigation of gardens.

Figure B.21 shows the size of each of the clusters formed:

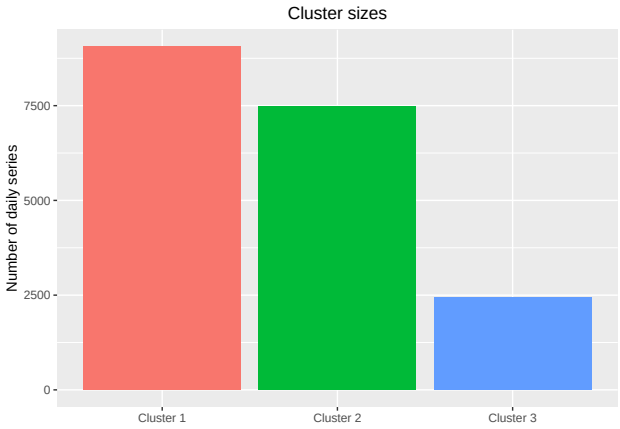

Figure B.21: Partition Clustering model with DTW distance, PAM prototype and 30m window clusters sizes.

Figure B.21 shows that most of the patterns belong to Cluster 1 with 9052 dailly flow patterns, followed by Cluster 2 presents with 7469 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Cluster 3 has 2453 associated daily patterns that represent predominantly nocturnal consumption.

Figure B.22 evaluates the degree of membership of each of the annual series to the formed clusters:

Figure B.22: Partition Clustering model with DTW distance, PAM prototype and 30m window annual series membership.

In Figure B.22 it is observed that in all the annual series the daily patterns belong mostly to Clusters 1 and 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous clustering methods, since most clusters belong to a pattern with predominantly diurnal consumption.

Table B.3 shows a set of statistical characteristics of the clusters formed:

Table B.3: Partition Clustering model with DTW distance, PAM prototype and 30m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**	**Cluster 3 (m^3/h)**
Min.	0.00	0.00	0.00
1st Qu.	10.03	5.62	4.73
Median	28.14	12.70	10.41
Mean	55.78	34.41	19.33
3rd Qu.	67.66	44.56	22.27
Max.	1207.00	981.25	530.50
IQR	57.63	38.94	17.54

Figure B.23 identifies the influence of weekend or holiday days have on the formation of clusters:

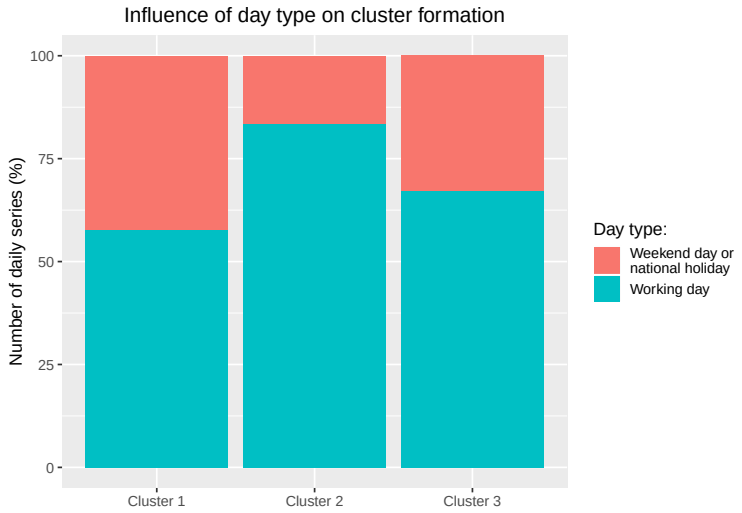

Figure B.23: Partition Clustering model with DTW distance, PAM prototype and 30m window influence of day typology on the formation of clusters.

In the case of cluster 1, the percentage of weekend or holiday patterns is around 44%. These values indicate that this cluster does not allow to identify a distinct behavior between a working day and a weekend or holiday.

As it can be seen, for cluster 2 the percentage of weekend or national holiday patterns is around 12%, proving that this cluster is associated with typical working day behavior.

For Cluster 3 the percentage of weekends and holidays is around 30%. These values indicate that the formed cluster do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure B.24 allows identifying the influence of day typology in each annual series by cluster type:

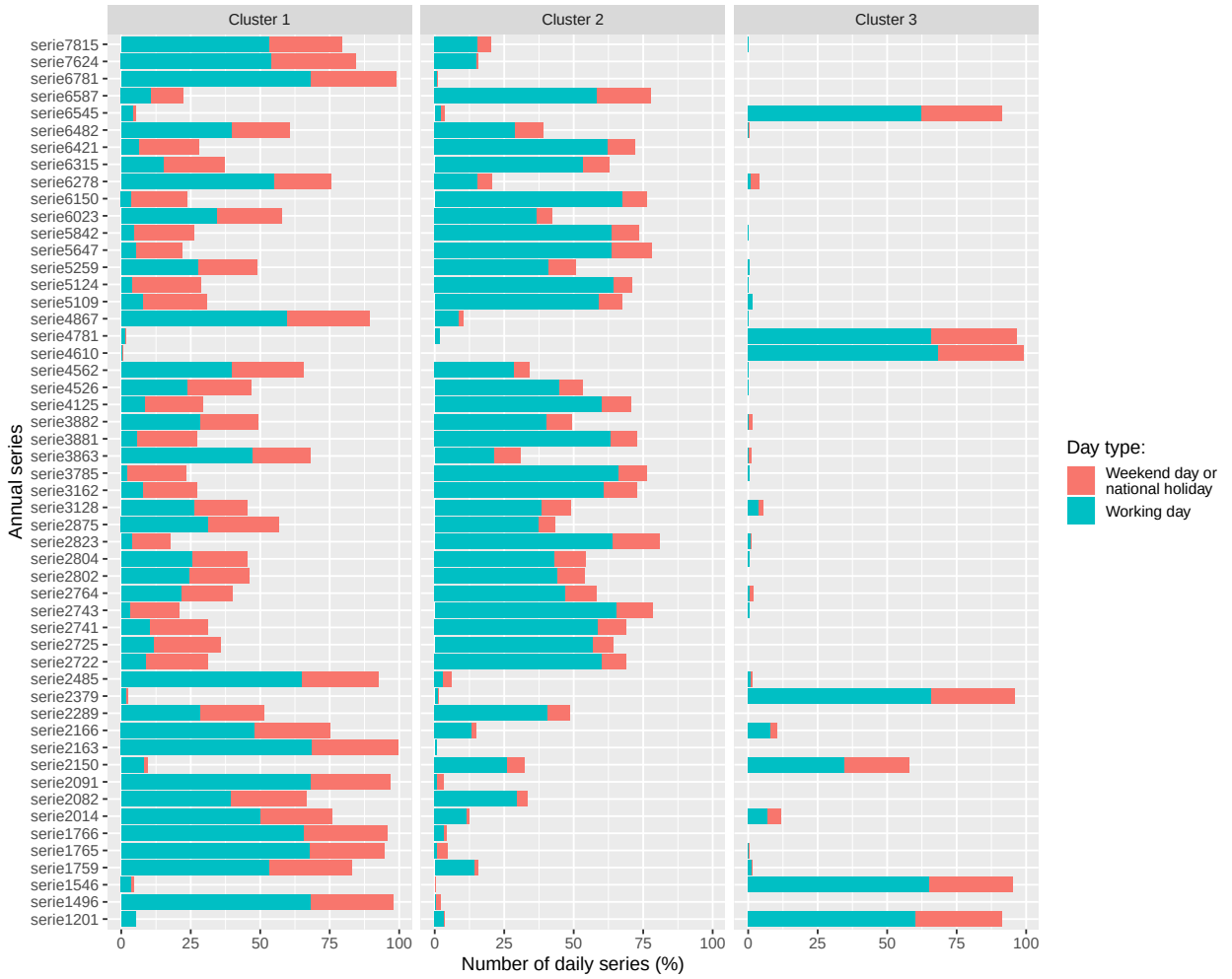

Figure B.24: Partition Clustering model with DTW distance, PAM prototype and 30m window influence of day typology on each series by clusters.

As can be seen in Cluster 2, the annual series show mostly a higher percentage of daily patterns in working days.

In the case of Cluster 1 the highest percentage of patterns in most series is also of working day typology but to a lesser extent than in Cluster 2.

Cluster 3, which represents daily patterns with higher nocturnal consumption, shows that in the annual series in which this cluster is the most representative, the proportions between working days and national holiday / weekend are indicative that typology of the day does not have significant influence on this cluster.

B.4 Partitional Clustering with DTW, DBA prototype and 15 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: DBA (see section 3.7.3);
- Comparison time window: 15 minutes (see section 3.6.2).

B.4.1 Clustering model internal index evaluation

Figure B.25 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

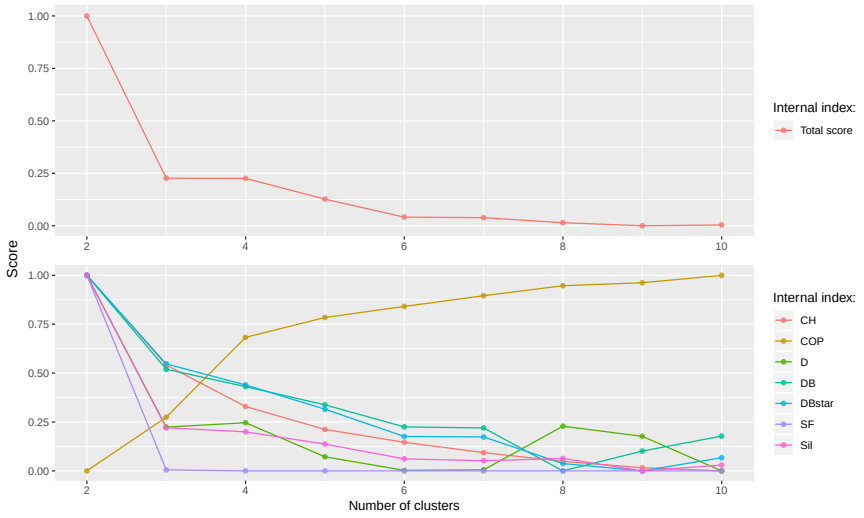

Figure B.25: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, DBA Prototype and 15 minutes time window.

Figure B.25 shows that the best result (Total score) was with the formation of 2 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.26 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

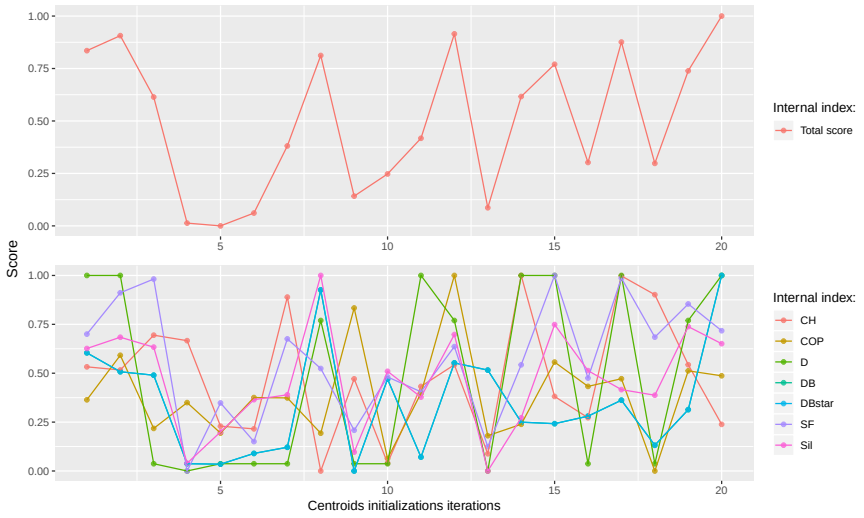

Figure B.26: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, DBA Prototype and 15 minutes time window.

Figure B.26 shows that the 20^{th} iteration provided the best performance in the internal indexes evaluation. In the next section the 20^{th} iteration clustering model with the formation of 2 clusters will be analyzed.

B.4.2 Clustering model characterization

Figure B.27 shows the visualization of the clusters formed by the model according to the first 3 principal components:

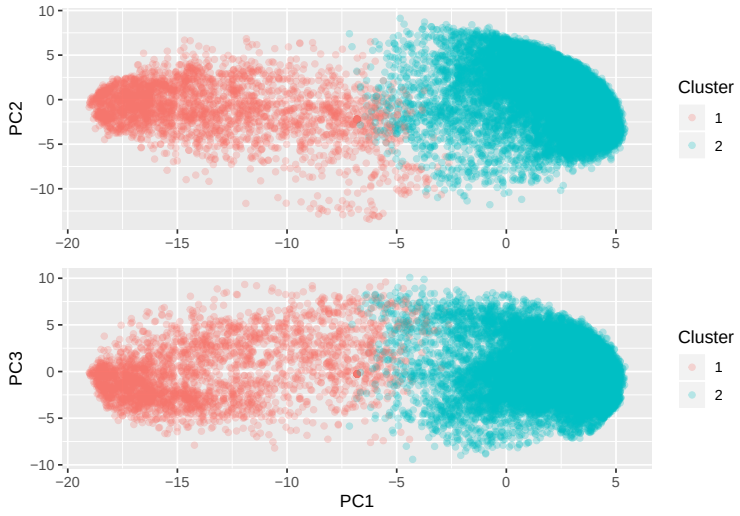

Figure B.27: Clusters formed through the Partition Clustering model with DTW distance, DBA prototype and 15m window visualized through the 3 principal components of PCA.

For this partiton clustering model with DTW 15 minutes window constraint and DBA centroid, Cluster 1 tends to negative zones according to the principal component 1 and Cluster 2 tends to positive zones according to this component, as in previous clustering models that formed 2 clusters.

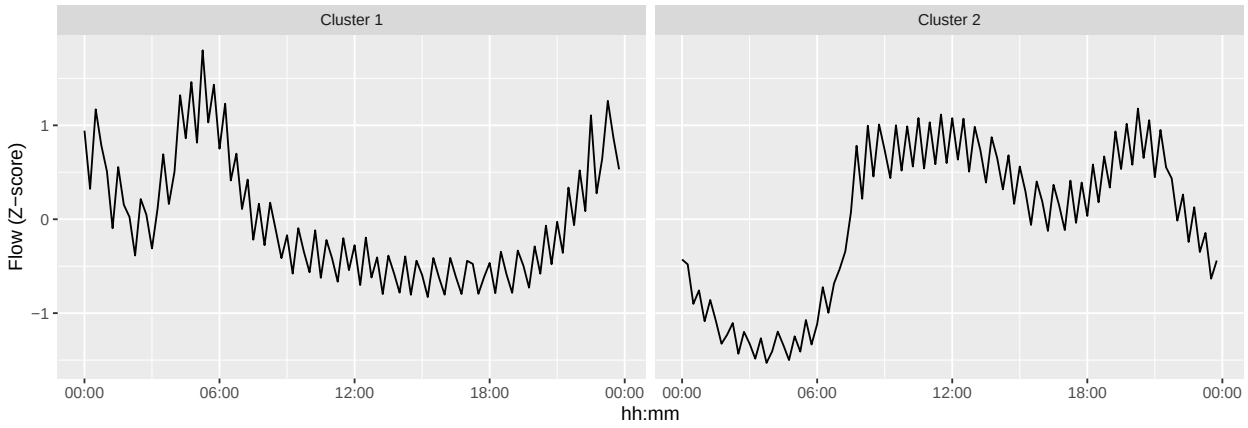

Figure B.28: Partition Clustering model with DTW distance, DBA prototype and 15m window centroids.

In Figure B.28 each centroid representing the clusters in this section is obtained by computing a mean at each point of the centroid taking into account the time points of the series that belong to the cluster and fit into the pre-defined time window. Consequently the presented centroids patterns represent a form of a averaged pattern and not of a real pattern of dataset. Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around 23:00 and the second peak of consumption occurs around 05:00. The periods of minimum consumption occur at 2:30 and around 14:00. In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 17:00. The absolute minimum consumption for Cluster 2 occurs around 04:00.

Figure B.29: Partition Clustering model with DTW distance, DBA prototype and 15m window clusters sizes.

Figure B.29 shows that most of the patterns belong to Cluster 2, and Cluster 1 presents only 2779 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure B.30 evaluates the degree of membership of each of the annual series to the formed clusters:

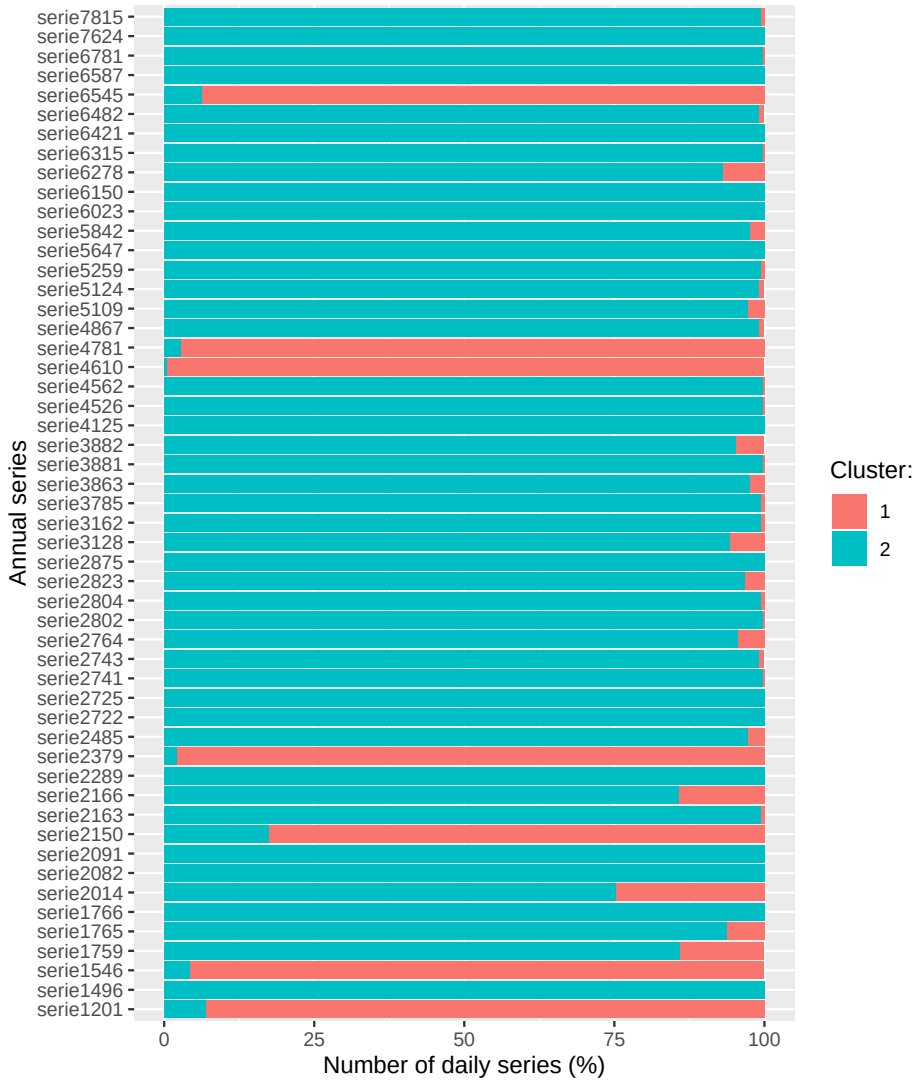

Figure B.30: Partition Clustering model with DTW distance, DBA prototype and 15m window annual series membership.

It was observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous clustering methods, since most clusters belong to a pattern with predominantly diurnal consumption.

Table B.4 shows a set of statistical characteristics of the clusters formed:

Table B.4: Partition Clustering model with DTW distance, DBA prototype and 15m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.72	7.33
Median	10.68	19.34
Mean	20.21	46.51
3rd Qu.	23.27	58.80
Max.	981.25	1207.00
IQR	18.55	51.47

Figure B.31 identifies the influence of weekend or holiday days have on the formation of clusters:

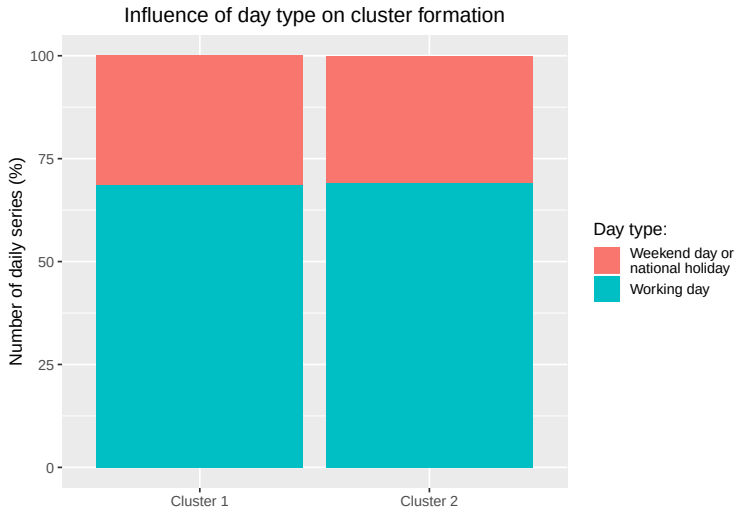

Figure B.31: Partition Clustering model with DTW distance, DBA prototype and 15m window influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 2 and Cluster 1. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.

Figure B.32 allows identifying the influence of day typology in each annual series by cluster type:

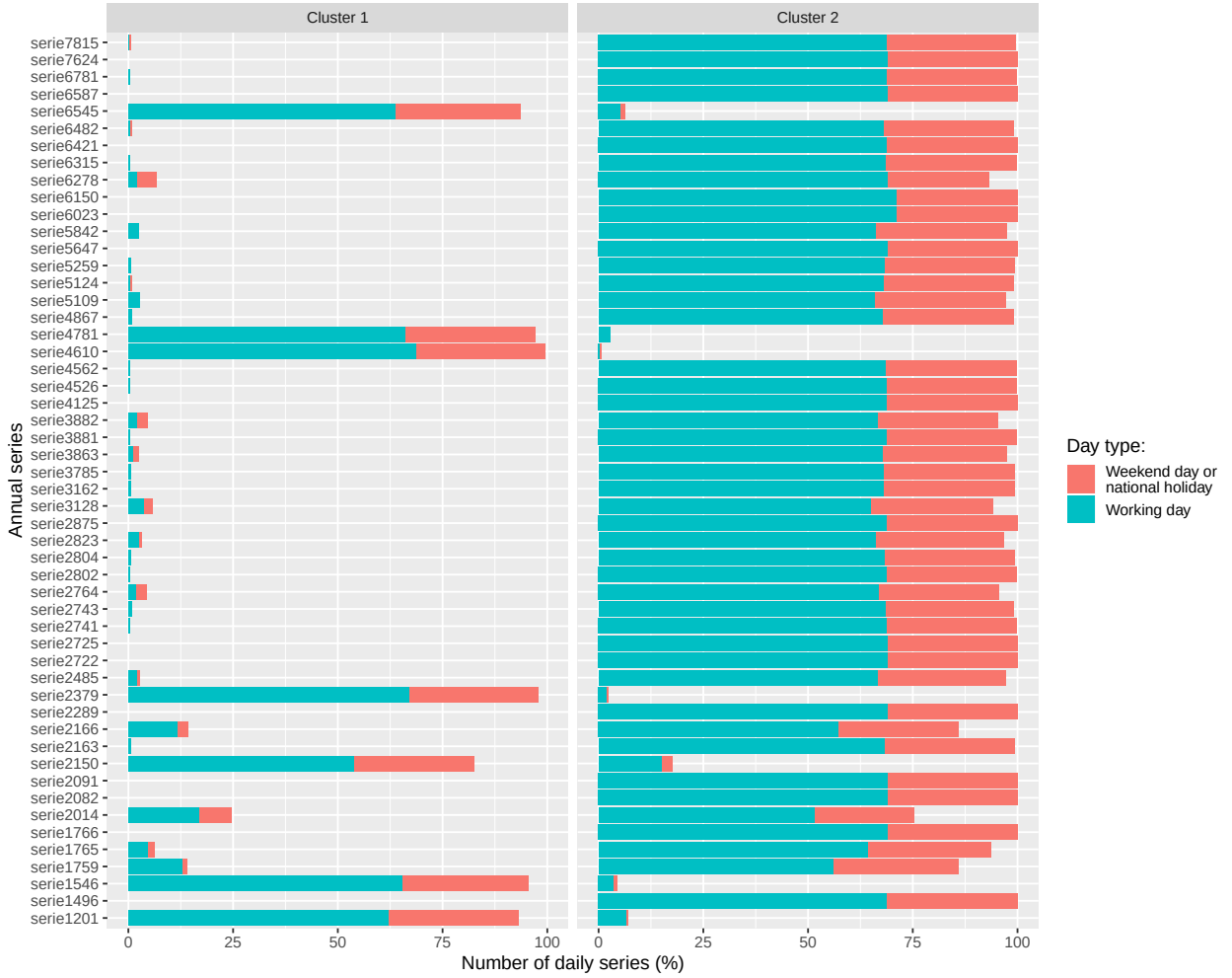

Figure B.32: Partition Clustering model with DTW distance, DBA prototype and 15m window influence of day typology on each series by clusters.

As can be seen from Figure B.32, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure B.31, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the previous clustering methods with formation of 2 clusters.

B.5 Partitional Clustering with DTW, DBA prototype and 30 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: DBA (see section 3.7.3);
- Comparison time window: 30 minutes (see section 3.6.2).

B.5.1 Clustering model internal index evaluation

Figure B.33 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

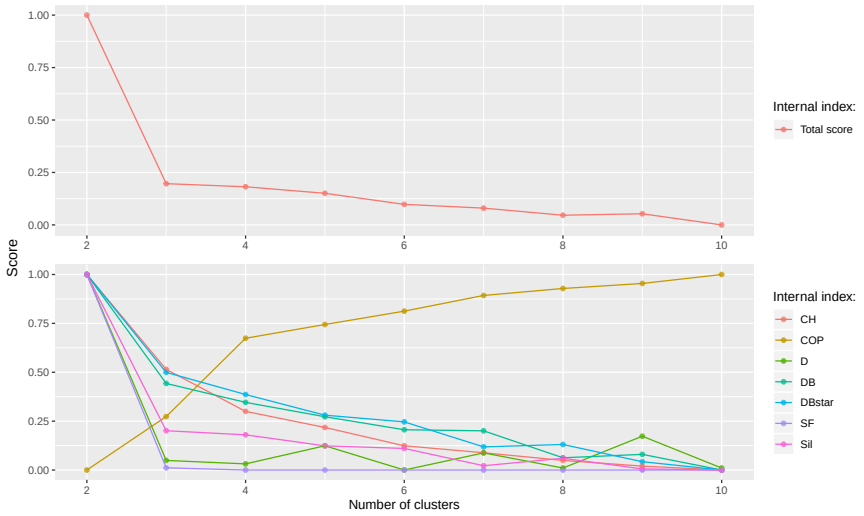

Figure B.33: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, DBA Prototype and 30 minutes time window.

Figure B.33 shows that the best result (Total score) was with the formation of 2 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.34 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

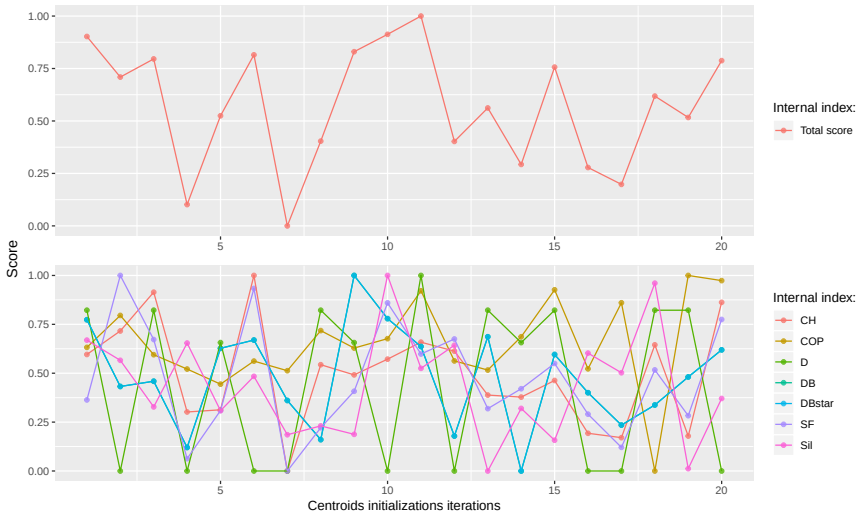

Figure B.34: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, DBA Prototype and 30 minutes time window.

Figure B.26 shows that the 11^{th} iteration provided the best performance in the internal indexes evaluation. In the next section the 11^{th} iteration clustering model with the formation of 2 clusters will be analyzed.

B.5.2 Clustering model characterization

Figure B.35 shows the visualization of the clusters formed by the model according to the first 3 principal components:

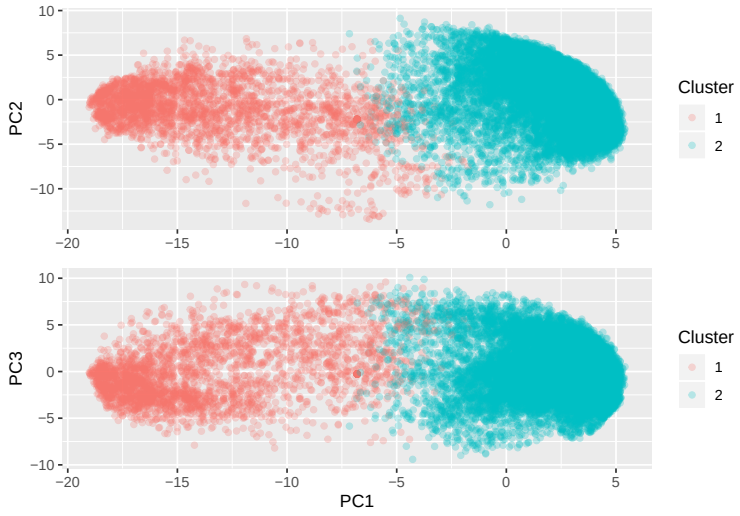

Figure B.35: Clusters formed through the Partition Clustering model with DTW distance, DBA prototype and 30m window visualized through the 3 principal components of PCA.

For this partiton algorithm with DTW 30 minutes window constraint and DBA centroid, Cluster 1 tends to negative zones according to the principal component 1 and Cluster 2 tends to positive zones according to this component, as in previous algorithms that formed 2 clusters.

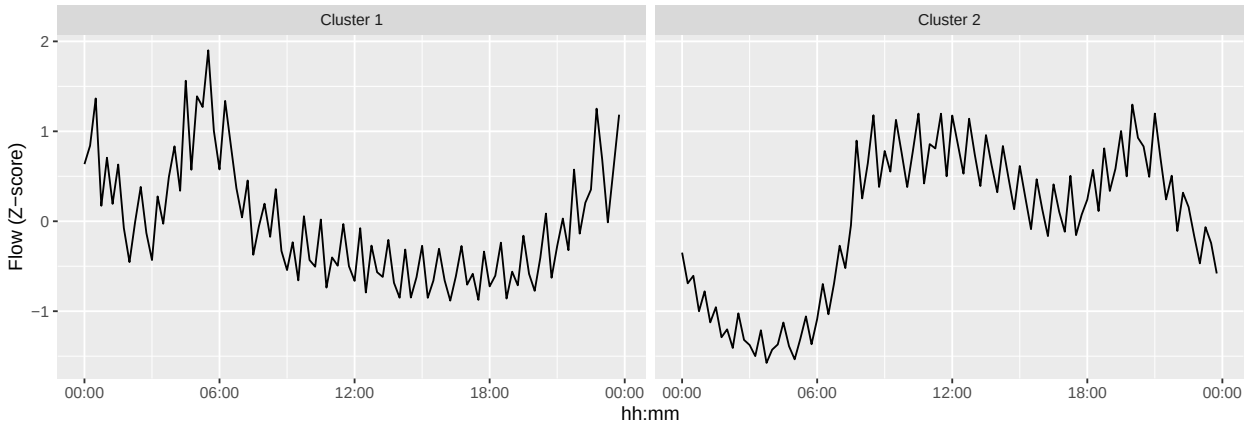

Figure B.36: Partition Clustering model with DTW distance, DBA prototype and 30m window centroids.

In Figure B.36 each centroid representing the clusters in this sub-chapter is obtained by computing a mean at each point of the centroid taking into account the time points of the series that belong to the cluster and fit into the pre-defined time window. Consequently the presented centroids patterns represent a form of a averaged pattern and not of a real pattern of dataset. Cluster 1 presents peak consumption in the night period, the first peak of consumption occurs around 23:00 and the second peak of consumption occurs around 05:00. The periods of minimum consumption occur at 2:30 and around 15:00. In the case of Cluster 2, consumption occurs predominantly during the daytime period with maximum consumption in the period of 12:00 and in the period of 20:00. Among these maximums the cluster prototype shows a local minimum at 16:00. The absolute minimum consumption for Cluster 2 occurs around 04:00.

Figure B.37: Partition Clustering model with DTW distance, DBA prototype and 30m window clusters sizes.

Figure B.37 shows that most of the patterns belong to Cluster 2, and Cluster 2 presents only 2824 daily flow patterns. Indicating that most daily patterns have predominantly peak flows during the daytime period.

Figure B.38 evaluates the degree of membership of each of the annual series to the formed clusters:

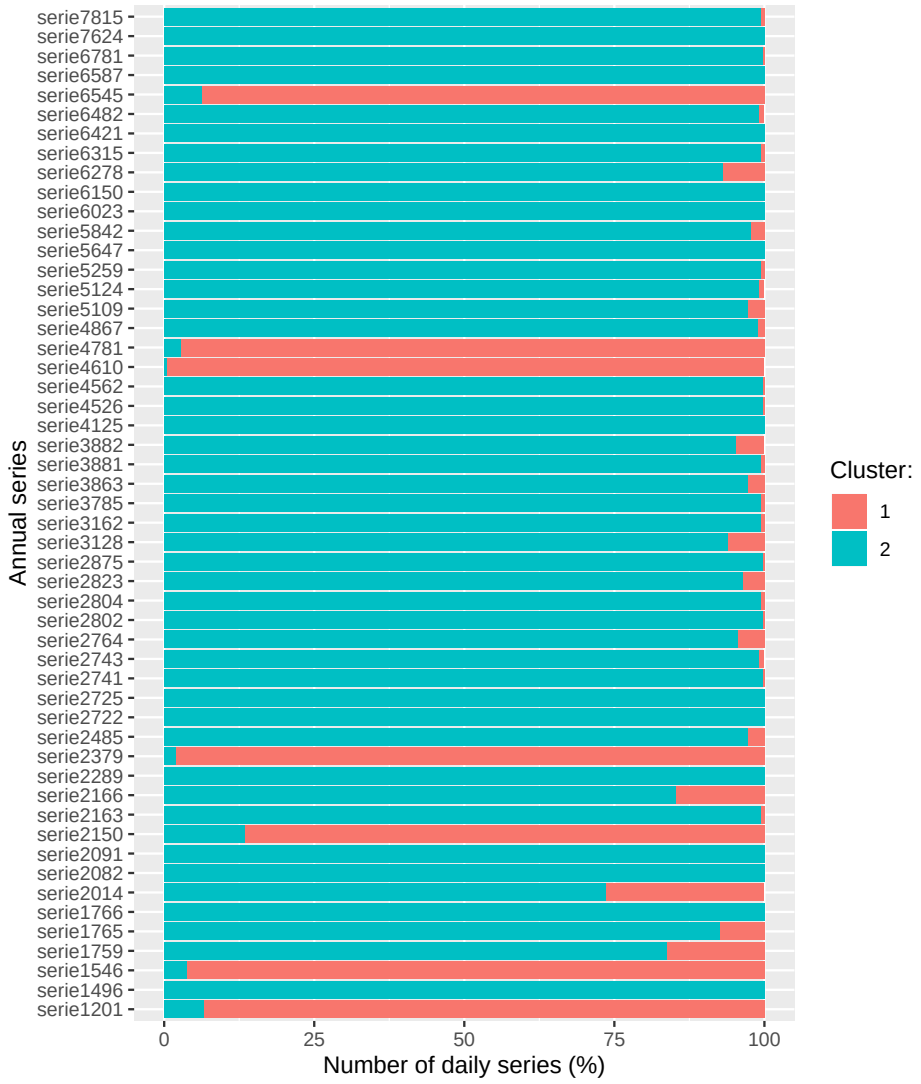

Figure B.38: Partition Clustering model with DTW distance, DBA prototype and 30m window annual series membership.

It was observed that in all the annual series the daily patterns belong mostly to Cluster 2, except the series 6545, 4781, 4610, 2379, 2150, 1546 and 1201. This result is consistent with what was observed in the formation of 2 clusters according to the previous clustering methods, since most clusters belong to a pattern with predominantly diurnal consumption.

Table B.5 shows a set of statistical characteristics of the clusters formed:

Table B.5: Partition Clustering model with DTW distance, DBA prototype and 30m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**
Min.	0.00	0.00
1st Qu.	4.73	7.34
Median	10.74	19.36
Mean	20.27	46.57
3rd Qu.	23.43	58.87
Max.	981.25	1207.00
IQR	18.70	51.53

Figure B.39 identifies the influence of weekend or holiday days have on the formation of clusters:

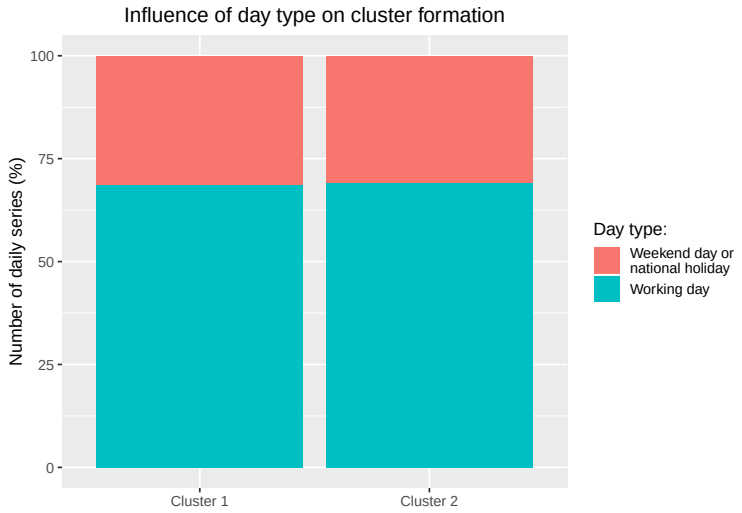

Figure B.39: Partition Clustering model with DTW distance, DBA prototype and 30m window influence of day typology on the formation of clusters.

As can be seen, the percentage of weekends and holidays is around 30% for Cluster 1 and Cluster 2. These values indicate that the formed clusters do not allow to identify a distinct behavior between a working day and a weekend or holiday.h

Figure B.40 allows identifying the influence of day typology in each annual series by cluster type:

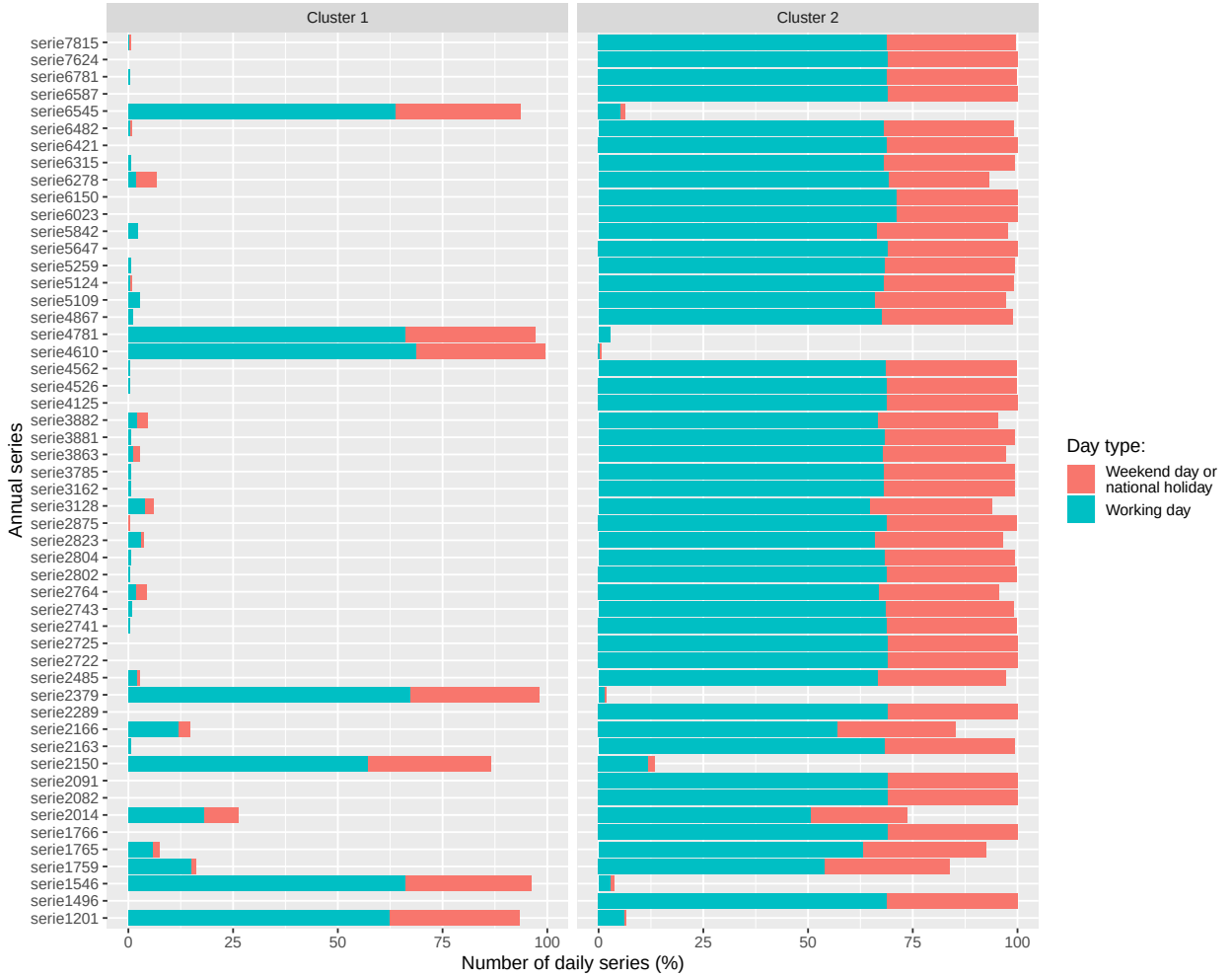

Figure B.40: Partition Clustering model with DTW distance, DBA prototype and 30m window influence of day typology on each series by clusters.

As can be seen from Figure B.40, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in Figure B.39, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day. This result is similar to the analysis carried out for the previous clustering methods with formation of 2 clusters.

B.6 Partitional Clustering with GAK, PAM prototype and 30 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: GAK (see section 3.6.3);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 30 minutes (see section 3.6.2).

B.6.1 Clustering model internal index evaluation

Figure B.41 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

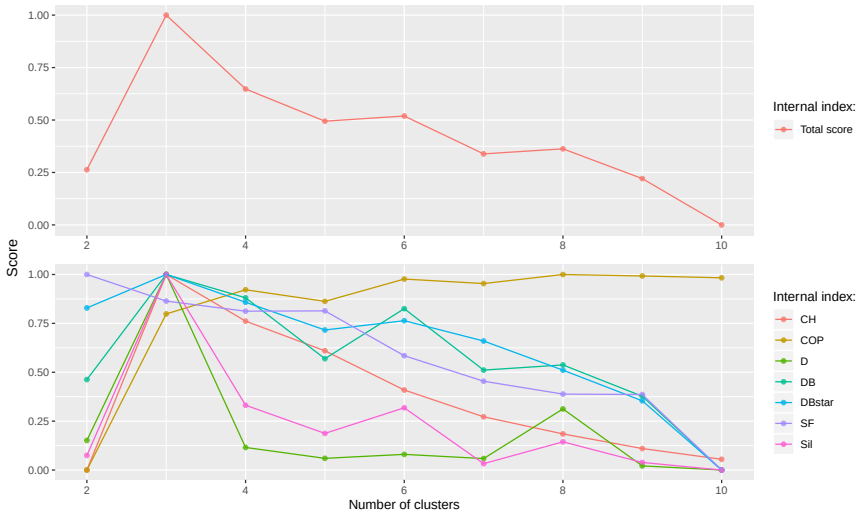

Figure B.41: Internal index evaluation for 1^{st} iteration set of Partitional Clustering with GAK, PAM Prototype and 30 minutes time window.

Figure B.41 shows that the best result (Total score) was with the formation of 3 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure B.42 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

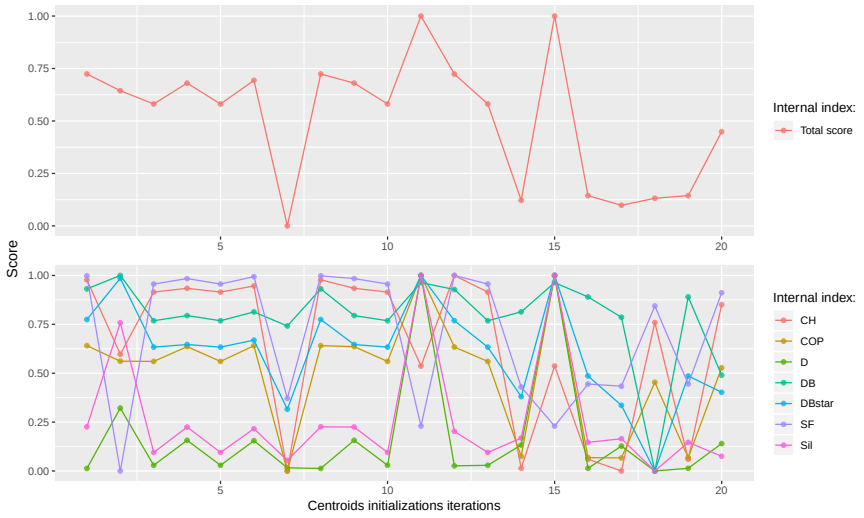

Figure B.42: Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with GAK, PAM Prototype and 30 minutes time window.

Figure B.42 shows that the 11^{th} and 15^{th} iterations provided the best performance in the internal indexes evaluation. In the next section the 15^{th} iteration clustering model with the formation of 3 clusters will be analyzed.

B.6.2 Clustering model characterization

Figure B.43 shows the visualization of the clusters formed by the model according to the first 3 principal components:

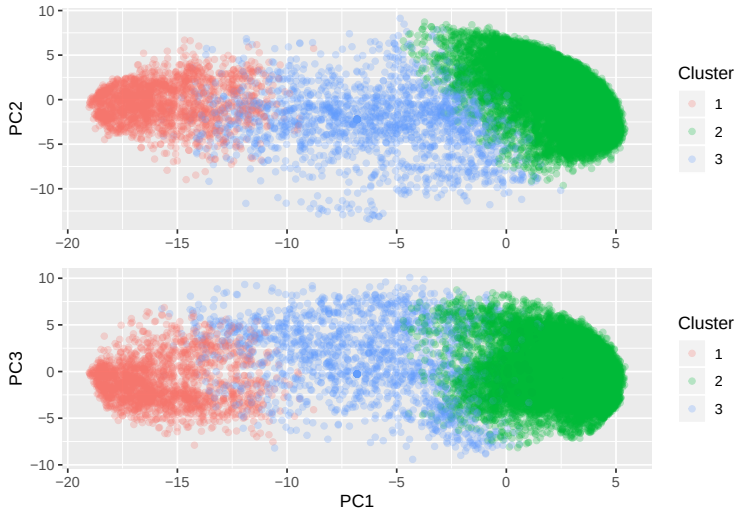

Figure B.43: Clusters formed through the Partition Clustering model with GAK distance, PAM prototype and 30m window visualized through the 3 principal components of PCA.

In Figure B.43 it is possible to see a separation of the clusters, being that Cluster 1 tends to be located tendentially in zones of value inferior to -12 in the principal component 1, Cluster 2 is tended in zones of value superior to -2.5 of the principal component 1. Cluster 3 is located in the intermediate zone between clusters 1 and 2.

The results obtained with the formation of 3 clusters are in line with the results obtained for the partition model with GAK distance, PAM centroid and 15 minutes window. But are also quite different in the location of the clusters compared to the results of other previously presented clustering modes with formatio of 3 clusters.

Figure B.44 shows the respective centroids of the clusters formed:

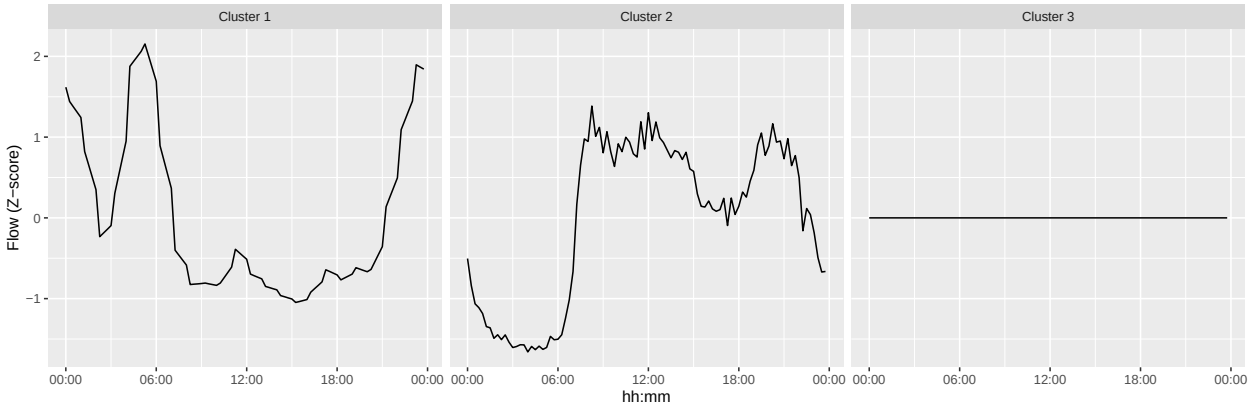

Figure B.44: Partition Clustering model with GAK distance, PAM prototype and 30m window centroids.

Cluster 1 shows peak consumption in the 23:00 period and in the period near 05:00 am. The predominance of this cluster by nocturnal consumption may be due to the use of water is predominantly associated with irrigation of gardens. Cluster 2 has a maximum consumption peak near 08:00, another local maximum at 12:00 and reaches a local minimum around 16:00. From this period consumption increases again until around 20:00 which is a local maximum. After this period the consumption drops back down to 05:00 which corresponds to the minimum value of consumption.The centroid of cluster 3 has a constant flow rate throughout the day. Overall Cluster 2 present higher consumption peaks during the day period, while Cluster 1 shows higher consumption during the night time period. Cluster 3 identifies a group of daily patterns that exhibit a behavior of less flow variation throughout the day compared to the other clusters.

Figure B.45 shows the size of each of the clusters formed. This Figure shows that most of the patterns belong to Cluster 2 with 15326 dailly flow patterns, followed by Cluster 1 presents with 1845 daily flow patterns. Cluster 3 presents 1806 daily flow patterns.

Figure B.45: Partition Clustering model with DTW distance, PAM prototype and 30m window clusters sizes.

Figure B.46 evaluates the degree of membership of each of the annual series to the formed clusters:

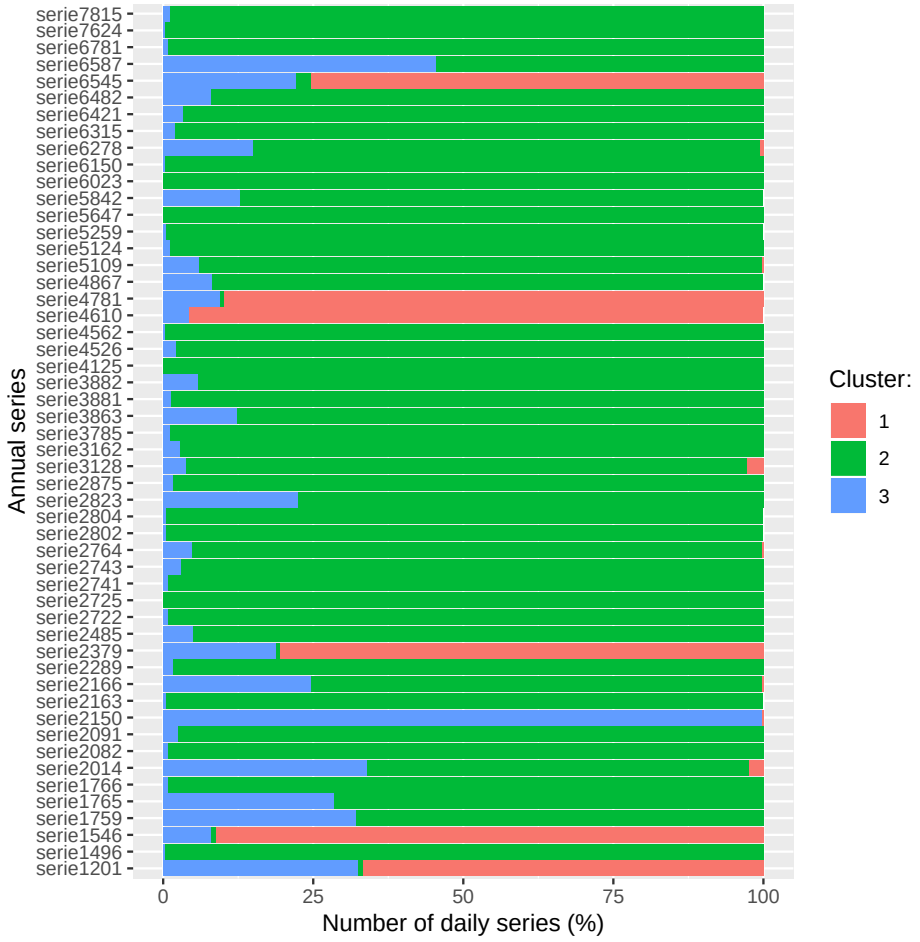

Figure B.46: Partition Clustering model with DTW distance, PAM prototype and 30m window annual series membership.

It was observed that in all the annual series the daily patterns belong mostly to Clusters 2, indicated that most annual series present higher consumption during the daytime period. The exceptions are the series 6545, 4781, 4610, 2379, 1546 and 1201 that belong mostly to Cluster 1 and therefore show higher consumption during the night time.

The annual series 2150 belongs mainly to cluster 3. Other annual series such as 6587, 2166, 2014, 1765, 1759 and 1201 show a high percentage of daily patterns belonging to cluster 3.

Table B.6 shows a set of statistical characteristics of the clusters formed:

Table B.6: Partition Clustering model with GAK distance, PAM prototype and 30m window clusters statistics.

Statistics	**Cluster 1 (m^3/h)**	**Cluster 2 (m^3/h)**	**Cluster 3 (m^3/h)**
Min.	0.00	0.00	0.00
1st Qu.	4.93	7.35	4.97
Median	10.39	18.83	17.85
Mean	18.28	46.17	37.77
3rd Qu.	21.65	57.81	46.40
Max.	312.25	1207.00	981.25
IQR	16.72	50.46	41.43

Figure B.47 identifies the influence of weekend or holiday days have on the formation of clusters:

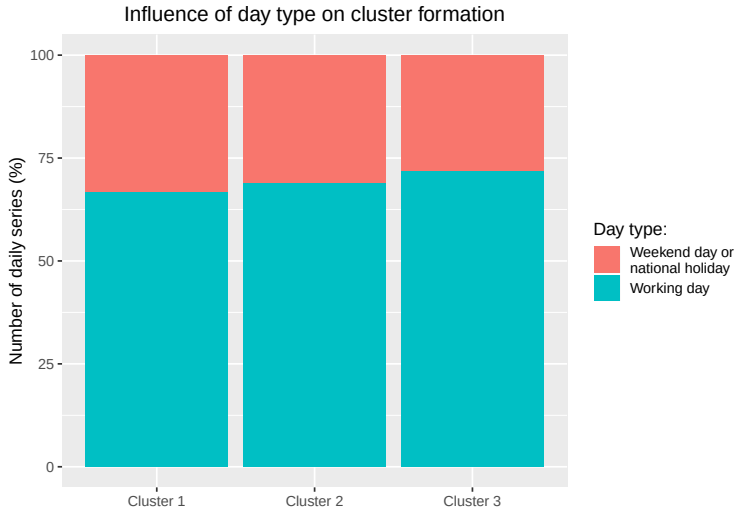

Figure B.47: Partition Clustering model with GAK distance, PAM prototype and 30m window influence of day typology on the formation of clusters.

It is observed that the percentage of weekends and holidays for clusters is about 30%. This distribution indicates that these Clusters do not identify a distinct behavior between working

day and weekend or holiday, since the assignment of the typology of days in a year is of the same order of magnitude.

Figure B.48 allows identifying the influence of day typology in each annual series by cluster type:

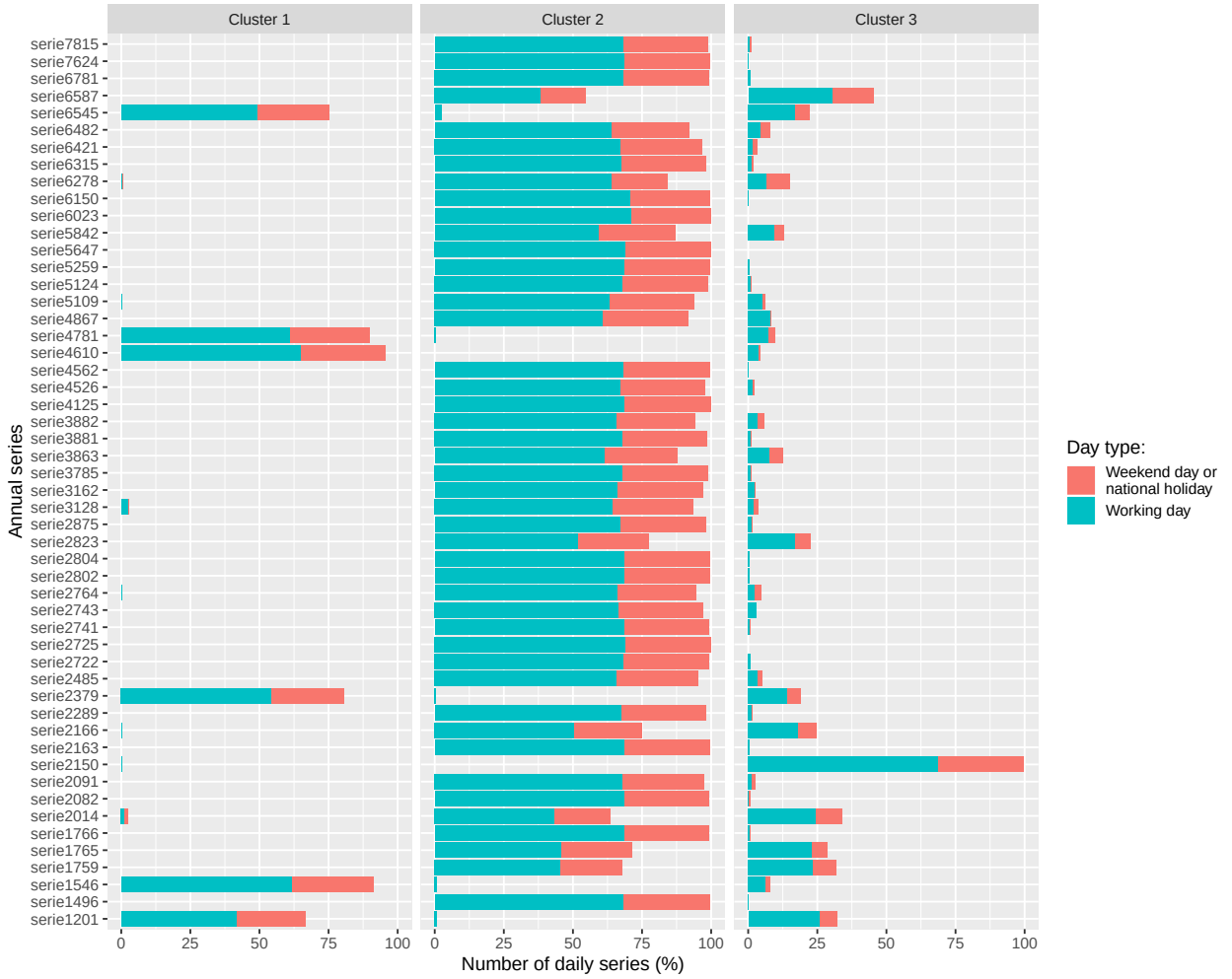

Figure B.48: Partition Clustering model with GAK distance, PAM prototype and 30m window influence of day typology on each series by clusters.

As can be seen from Figure B.48, in the most representative cluster of each annual series it is verified that the proportions of daily patterns belonging to each day typology remains similar to that presented in the graph of the previous section, evidencing that in general there is no influence of the typology of the day in these cases, but in the case of the clusters with less representation for each annual series usually there is influence of the typology of the day.

Appendix C

Cluster 2 - Clustering models with elastic distance measures

C.1 Cluster 2 - Partitional Clustering with DTW, PAM prototype and 15 minutes time window

In this section Cluster 2 subsets will bo analyzed using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: DTW (see section 3.6.2);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 15 minutes (see section 3.6.2).

C.1.1 Clustering model internal index evaluation

Figure C.1 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

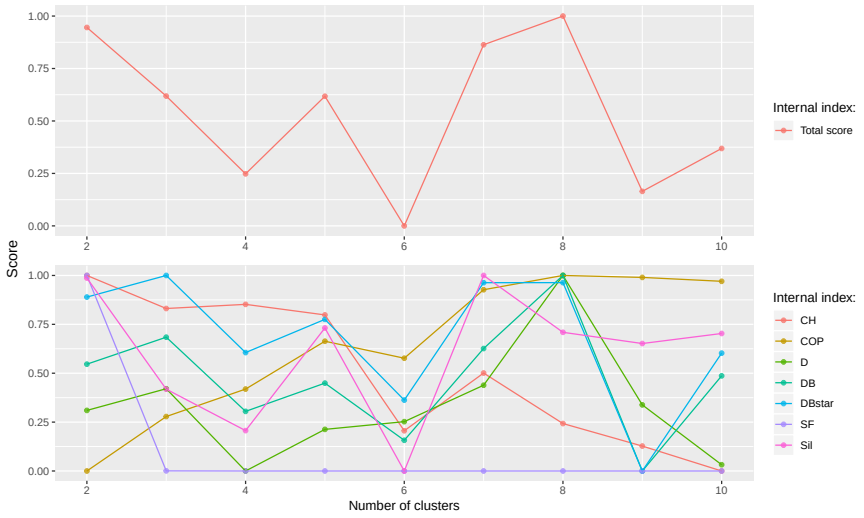

Figure C.1: Cluster 2 - Internal index evaluation for 1^{st} iteration set of Partitional Clustering with DTW, PAM and 15m time window.

Figure C.1 shows that the best result (Total score) was with the formation of 8 clusters. This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure C.2 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 3 clusters with 20 random centroids initializations.

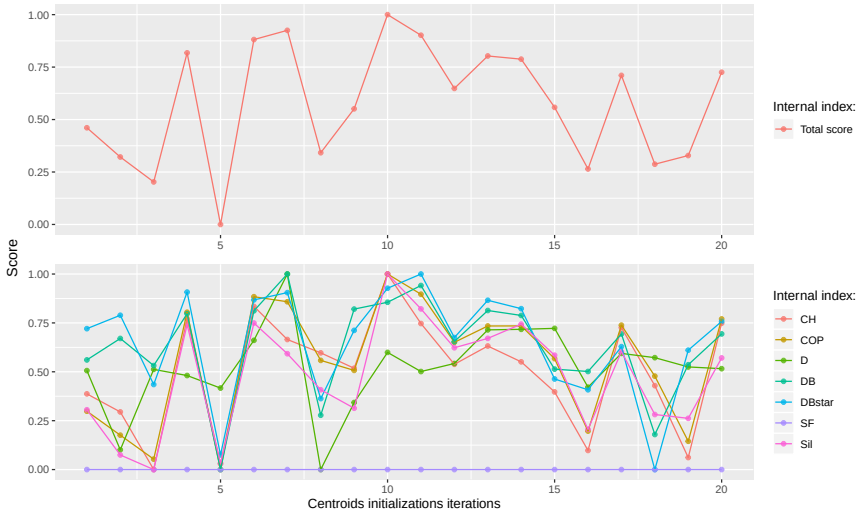

Figure C.2: Cluster 2 - Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with DTW, PAM and 15m time window.

Figure C.2 shows that the 10^{th} iteration provided the best performance in the internal indexes evaluation.

In the next section the 10^{th} iteration clustering model with the formation of 3 clusters will be analyzed.

C.1.2 Clustering model characterization

Figure C.3 shows the visualization of the clusters formed by the model according to the first 3 principal components:

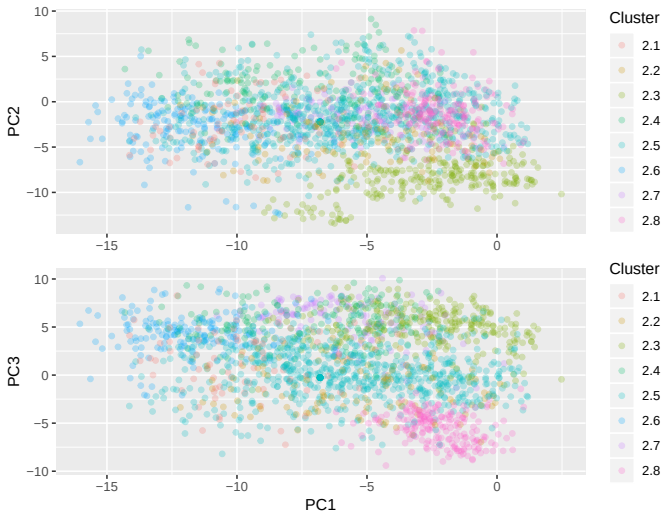

Figure C.3: Cluster 2 - Clusters formed through the Partition Clustering model with DTW, PAM and 15m window visualized through the 3 principal components of PCA.

From Figure C.3 it can be seen that there are no well defined boundaries between clusters formed according to the space represented by the first 3 principal components, this finding may result from the fact that these clusters may not be well represented in time periods with greater weight on principal components (see Figure 4.9).

Figure C.4 shows the respective centroids of the clusters formed:

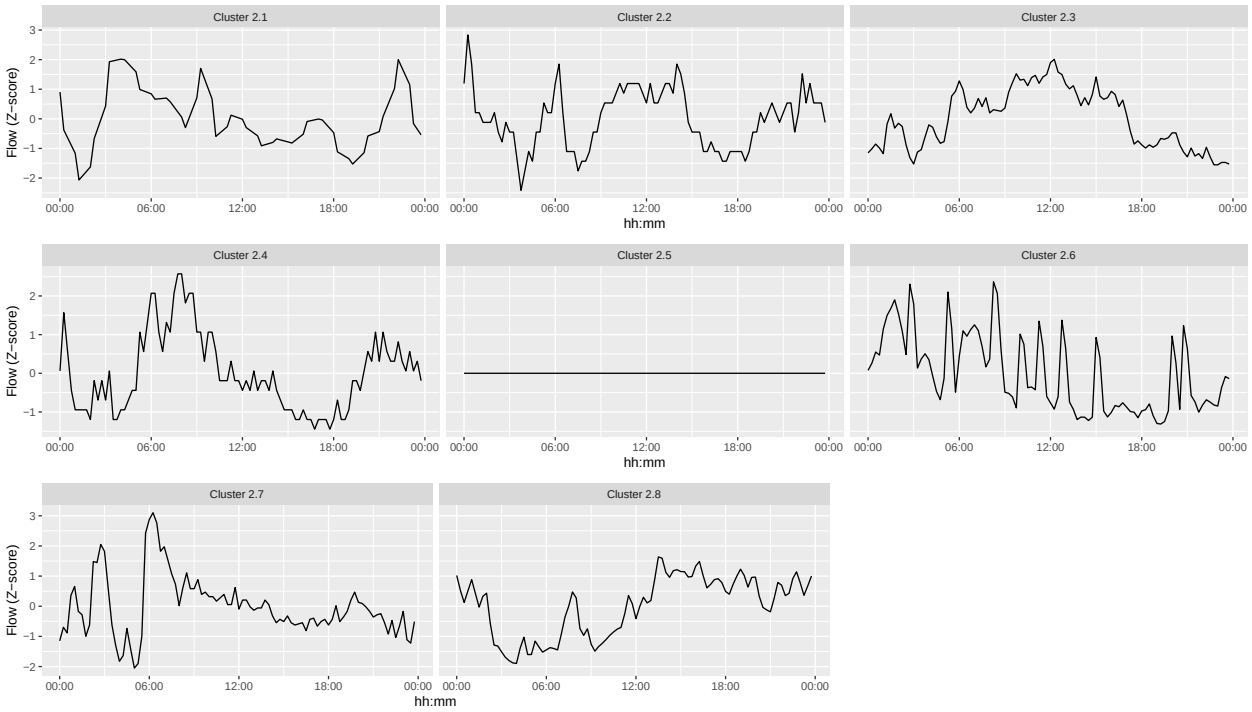

Figure C.4: Cluster 2 - Partition Clustering model with DTW, PAM and 15m window centroids.

Figure C.4 shows that these clusters do not have a significant variability between day and night consumption as seen in cluster 1, 3 and 4 of the Combined Model. Clusters 2.3 and 2.4 are an exception as they have higher daytime consumption compared to nighttime consumption.

Through the centroids represented in Figure C.4 it was possible to characterize the clusters as follows:

- **Cluster 2.1**: shows peak consumption at 04:00, 09:00 and 22:00 and minimum consumption at 01:00 and 19:30. The maximum consumption levels recorded for this cluster are all of the same order of magnitude, which shows that this cluster has water consumption for irrigation (4:00 peak) but is no higher than domestic consumption for other purposes that typically exists in maximum consumption periods recorded at 09:00 and 22:00;
- **Cluster 2.2**: it has peak consumption at 22:00, 00:00, 06:00 and a high water consumption zone between 10:00 and 14:00. Similar to cluster 2.1 the maximum consumptions recorded for this cluster are all roughly the same order of magnitude except the maximum of 00:00, which shows that the watering demand for irrigation is higher than in cluster 2.1;
- **Cluster 2.3**: this cluster has a minimum consumption at 00:00 and from that moment the consumption will grow until 12:00, from that moment the consumption will decrease again. In this cluster the consumption derived from irrigation is not significant, since there is no maximum consumption during the night, and the highest consumption

occurs during the lunch period;

- **Cluster 2.4**: shows consumption peaks at 00:00, 03:00, 06:00, 08:00 and 21:00. The maximum consumption associated with the period of 08:00 is a maximum consumption significantly higher than the remaining local maximums, indicating that in this cluster the consumption associated with irrigation is lower than the domestic consumption used for other purposes;
- **Cluster 2.5**: presents the same centroid as cluster 2 of the Combined Model, meaning cluster 2.5 encompasses daily patterns in which the variation between day and night consumption is still less significant than that recorded in the other subclusters of cluster 2;
- **Cluster 2.6**: this cluster behaves quite differently from the others, the variations in consumption are more instantaneous which can reveal irrigation consumptions throughout the day. This behavior has been detected in this cluster because there may not be significant water consumptions in this cluster for other uses. This fact indicates that instant variations of irrigation consuptions are not diluted in consumptions of other typologies. The behavior represernted by this cluster may not be desirable as it indicates that there is irrigation at times of day with the most sun exposure and some of the water used for these purpose evaporates and is not absorbed by the soil at those times with the most sun exposure;
- **Cluster 2.7**: its consumption tends to decrease between 09:00 and 00:00. In the period from 00:00 to 09:00, it shows local maximums at 01:00 and 03:00 and an absolute maximum at 06:00 associated with irrigation consumptions;
- **Cluster 2.8**: this cluster has a peak consumption at 07:30 which decreases until it reaches a local low at 09:00. Consumption then tends to increase between 09:00 and 14:30, from that moment consumption shows little variability until 01:00, from that moment consumption decreases to the absolute minimum at 04:00.

Figure C.5 shows the size of each of the clusters formed:

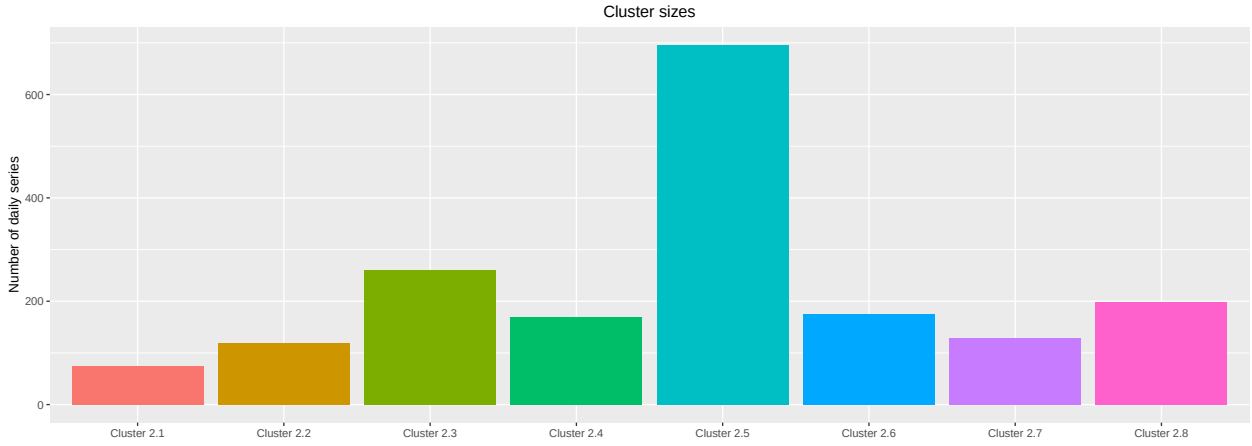

Figure C.5: Cluster 2 - Partition Clustering model with DTW, PAM and 15m window clusters sizes.

Figure C.5 shows that the largest cluster is Cluster 2.5, which has about 696 daily patterns.

The remaining clusters are in the range of 100 to 200 daily patterns. Except for cluster 2.1 which has 75 daily patterns and cluster 2.3 with 260 daily patterns.

Table C.1 shows a set of statistical characteristics of the clusters formed:

Table C.1: Cluster 2 - Partition Clustering model with DTW, PAM and 15m window clusters statistics.

Statistics	**Cluster 2.1 (m^3/h)**	**Cluster 2.2 (m^3/h)**	**Cluster 2.3 (m^3/h)**	**Cluster 2.4 (m^3/h)**	**Cluster 2.5 (m^3/h)**	**Cluster 2.6 (m^3/h)**	**Cluster 2.7 (m^3/h)**	**Cluster 2.8 (m^3/h)**
Min.	0.97	0.00	0.00	0.00	0.00	0.07	0.95	0.00
1st Qu.	6.86	18.40	8.75	10.13	2.35	11.12	22.00	103.61
Median	12.12	60.00	20.80	52.00	5.90	20.04	28.40	128.12
Mean	34.65	62.44	25.95	55.79	17.29	29.02	31.30	110.76
3rd Qu.	19.39	95.42	34.80	76.00	16.19	31.69	34.40	145.99
Max.	881.00	208.75	251.08	376.42	981.25	299.62	766.37	184.24
IQR	12.53	77.02	26.05	65.86	13.84	20.57	12.40	42.38

Figure C.6 identifies the influence of weekend or holiday days have on the formation of clusters:

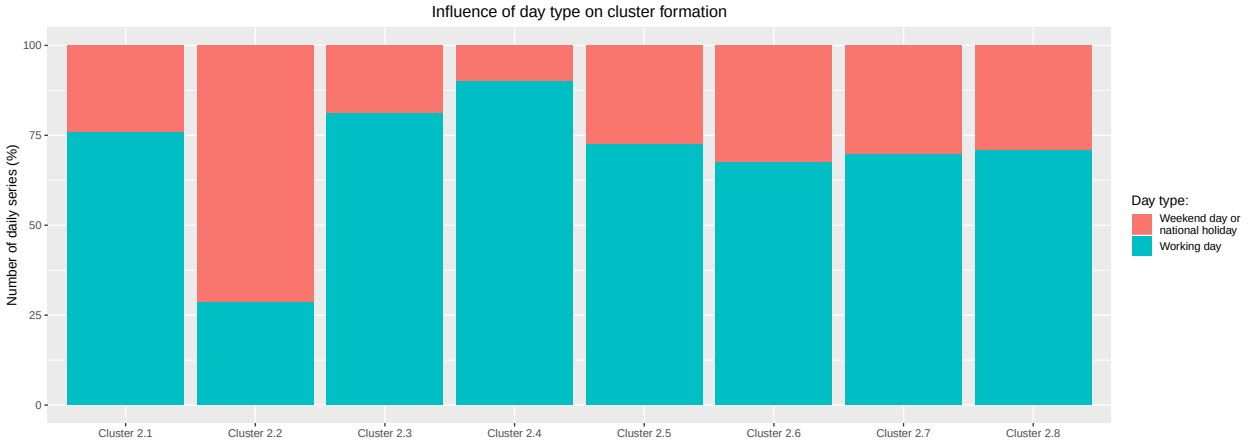

Figure C.6: Cluster 2 - Partition Clustering model with DTW, PAM and 15m window influence of day typology on the formation of clusters.

As it can be seen, most of the clusters are associated with typical workday behavior. Except in the case of cluster 2.2, the percentage of weekend or holiday patterns is around 70%, proving that this cluster is associated with typical weekend or holiday behavior.

Figure C.7 shows the geographic distribution of the clusters formed by the model:

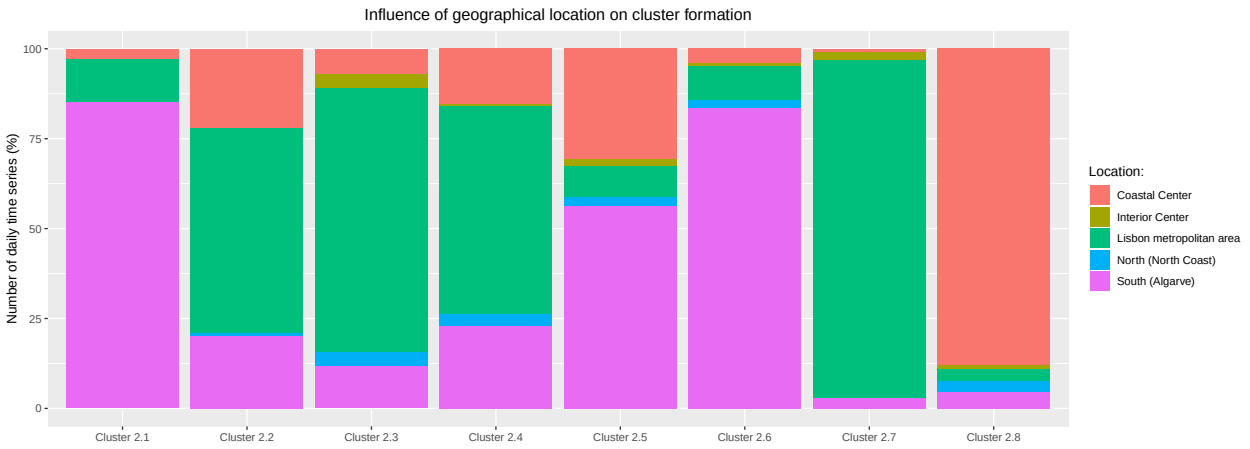

Figure C.7: Cluster 2 - Partition Clustering model with DTW, PAM and 15m window - geographic distribution of the clusters formed.

Figure C.7 shows that Clusters 2.1, 2.5 and 2.6 belong mainly to the South (Algarve) region. In the case of Clusters 2.2, 2.3, 2.4 and 2.7, they mainly belong to the Lisbon metropolitan area. Cluster 2.8 mainly belongs to the Costal Center region. The Interior Center region has little representation in these clusters, and is present in clusters 2.3, 2.4, 2.5, 2.7 and 2.8.

It should be noted that Cluster 2.7 that was previously identified as having poor management of irrigation periods is associated with the South (Algarve) region.

Figure C.8 shows the distribution of wet months and dry months in the clusters formed by the model:

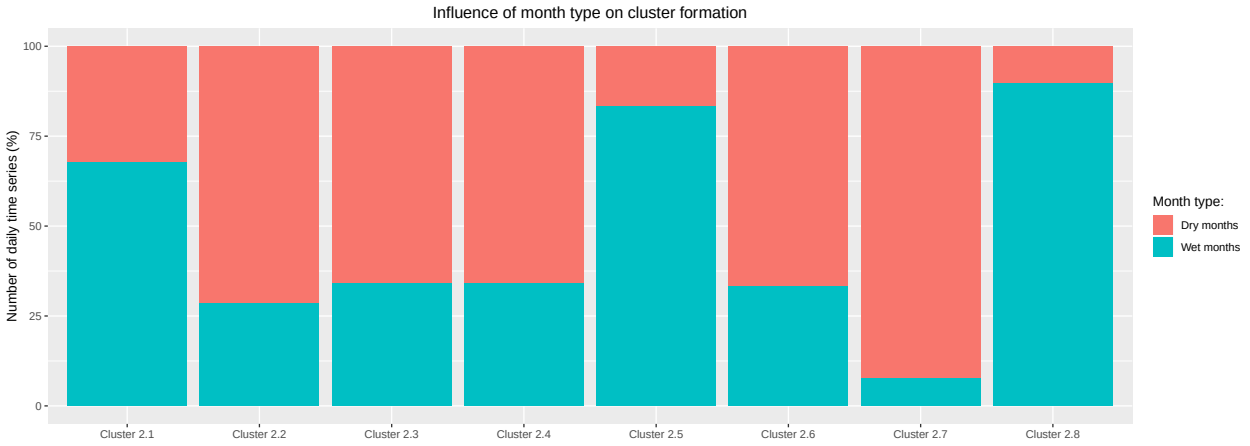

Figure C.8: Cluster2 - Partition Clustering model with DTW, PAM and 15m window - distribution of wet months and dry months in the clusters formed.

Figure C.8 shows that most of these clusters belong mostly to the dry months typology. With the exception of Clusters 2.1, 2.5 and 2.8 which mostly belong to the wet months typology.

C.2 Cluster 2 - Partitional Clustering with GAK, PAM prototype and 15 minutes time window

In this section we will analyze a clustering model using the Partitional Clustering approach (see section 3.5.2) with the following components:

- Distance measure: GAK (see section 3.6.3);
- Prototype: PAM (see section 3.7.2);
- Comparison time window: 15 minutes (see section 3.6.2).

C.2.1 Clustering model internal index evaluation

Figure C.9 shows the internal index validation of the 1^{st} iteration set, which aims to validate the optimal number of clusters to form within the range of 2 to 10 clusters.

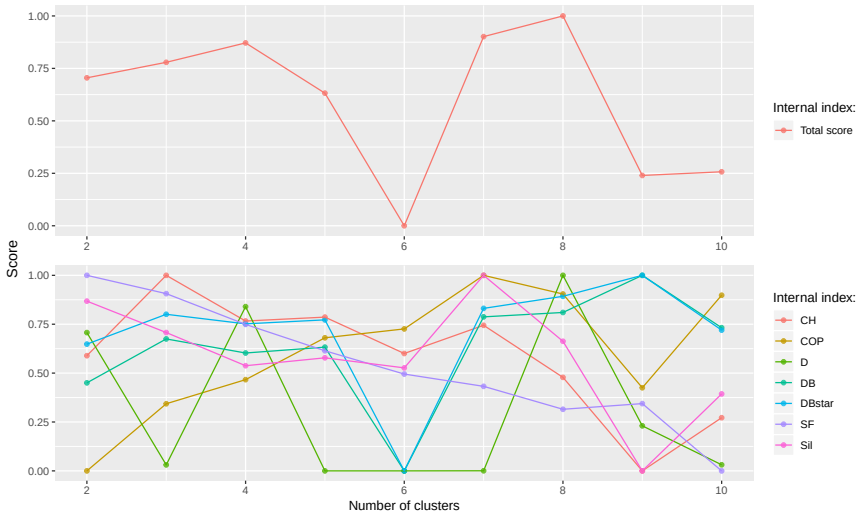

Figure C.9: Cluster 2 - Internal index evaluation for 1^{st} iteration set of Partitional Clustering with GAK, PAM and 15m time window.

Figure C.9 shows that the best result (Total score) was with the formation of 8 clusters.

This clustering approach needs to initially allocate centroids (see section 3.5.2), after setting the number of clusters to be formed it is necessary to run the model with different centroid initializations in order to evaluate which centroids initialization is best according to the internal index measures.

Figure C.10 shows the internal index validation of the 2^{nd} iteration set, which aims to validate the best centroids initialization, running the model to form 2 clusters with 20 random centroids initializations.

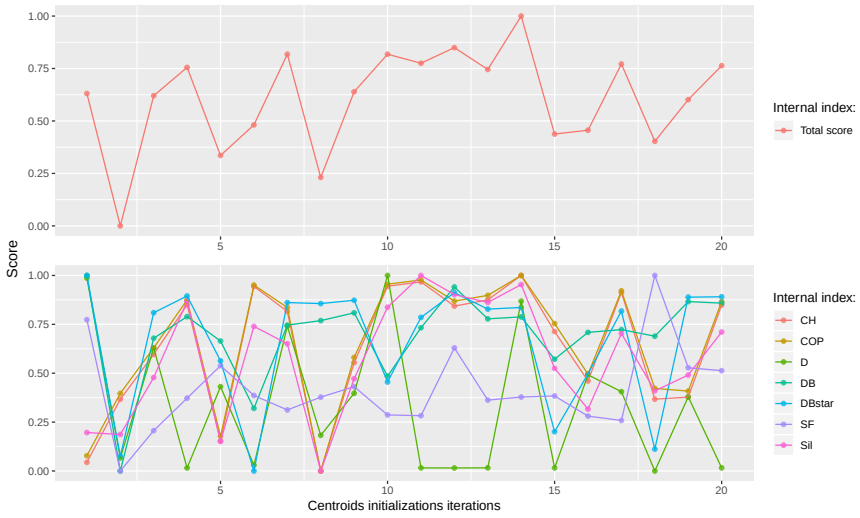

Figure C.10: Cluster 2 - Internal index evaluation for 2^{nd} iteration set of Partitional Clustering with GAK, PAM and 15m time window.

Figure C.10 shows that the 14^{th} iteration provided the best performance in the internal indexes evaluation. In the next section the 14^{th} iteration clustering model with the formation of 3 clusters will be analyzed.

C.2.2 Clustering model characterization

Figure C.11 shows the visualization of the clusters formed by the model according to the first 3 principal components:

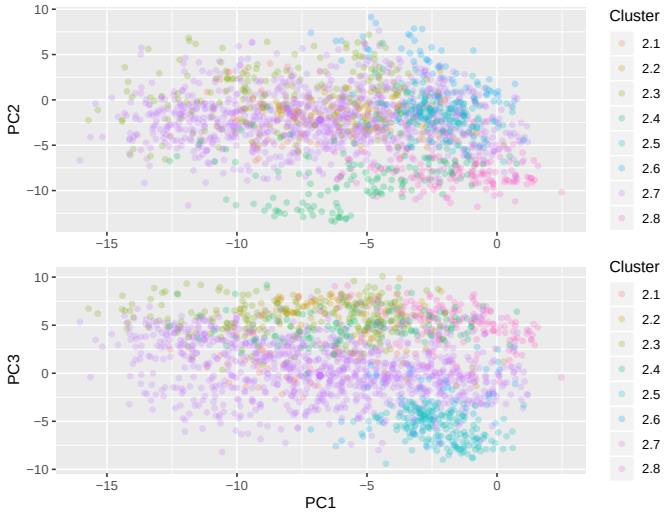

Figure C.11: Cluster 2 - Clusters formed through the Partition Clustering model with GAK, PAM and 15m window visualized through the 3 principal components of PCA.

From Figure C.11 it can be seen that there are no well defined boundaries between clusters formed according to the space represented by the first 3 principal components, this finding may result from the fact that these clusters may not be well represented in time periods with greater weight on principal components (see Figure 4.9).

Figure C.12 shows the respective centroids of the clusters formed:

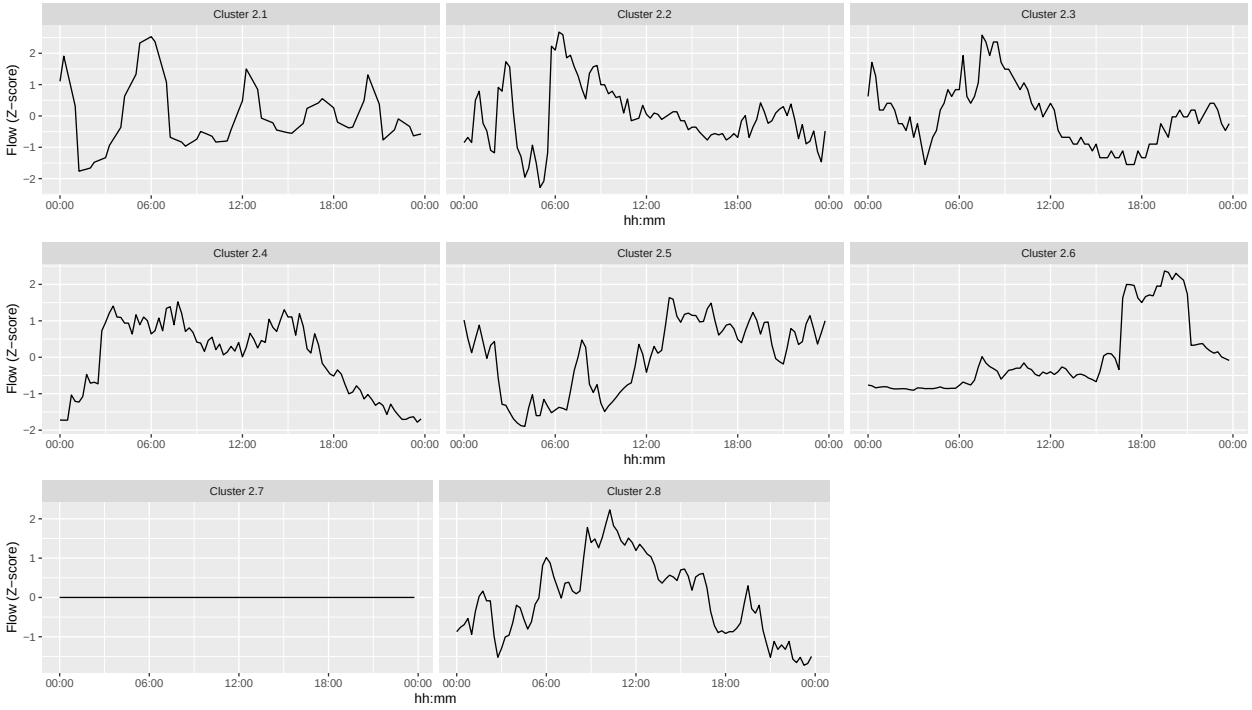

Figure C.12: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window centroids.

Figure C.12 shows that these clusters do not have a significant variability between day and night consumption as seen in cluster 1, 3 and 4 of the Combined Model. Clusters 2.3 and 2.4 are an exception as they have higher daytime consumption compared to nighttime consumption.

Through the centroids represented in Figure C.12 it was possible to characterize the clusters as follows:

- **Cluster 2.1**: shows peak consumption at 06:00, 12:00, 17:00 and 20:00 and minimum consumption at 01:30, 08:00, 10:30, 15:00, and 21:00. The maximum consumption levels recorded for 06:00 period is significantly higher than the remaining consumption maximums, which shows that in this cluster the use of water for irrigation is significant;
- **Cluster 2.2**: its consumption tends to decrease between 09:00 and 00:00. In the period from 00:00 to 09:00, it shows local maximums at 01:00 and 03:00 and an absolute maximum at 06:00 associated with irrigation consumptions;
- **Cluster 2.3**: shows consumption peaks at 00:00, 06:00, 08:30 and 22:30. The maximum consumption associated with the period of 08:30 is a maximum consumption

significantly higher than the remaining local maximums, indicating that in this cluster the consumption associated with irrigation is lower than the domestic consumption used for other purposes;

- **Cluster 2.4**: shows consumption peaks at 03:30, 07:30, 15:00. The maximum consumption levels recorded for this cluster are all of the same order of magnitude, which shows that this cluster has water consumption for irrigation (03:30 peak) but is not higher than domestic consumption for other purposes that typically exists in maximum consumption periods recorded at 07:30 and 15:00;
- **Cluster 2.5**: this cluster has a peak consumption at 07:30 which decreases until it reaches a local low at 09:00. Consumption then tends to increase between 09:00 and 14:30, from that moment consumption shows little variability until 01:00, from that moment consumption decreases to the absolute minimum at 04:00;
- **Cluster 2.6**: this cluster behaves quite differently from the others, it has a low consumption between 00:00 and 16:30, from this period consumption increases rapidly and only decreases after 21:00;
- **Cluster 2.7**: presents the same centroid as cluster 2 of the Combined Model, meaning cluster 2.7 encompasses daily patterns in which the variation between day and night consumption is still less significant than that recorded in the other subclusters of cluster 2;
- **Cluster 2.8**: this cluster has a minimum consumption at 00:00 and from that moment the consumption will grow until 09:30, from that moment the consumption will decrease again. In this cluster the consumption derived from irrigation is not significant, since there is no maximum consumption during the night, and the highest consumption occurs during the morning period.

Figure C.13 shows the size of each of the clusters formed:

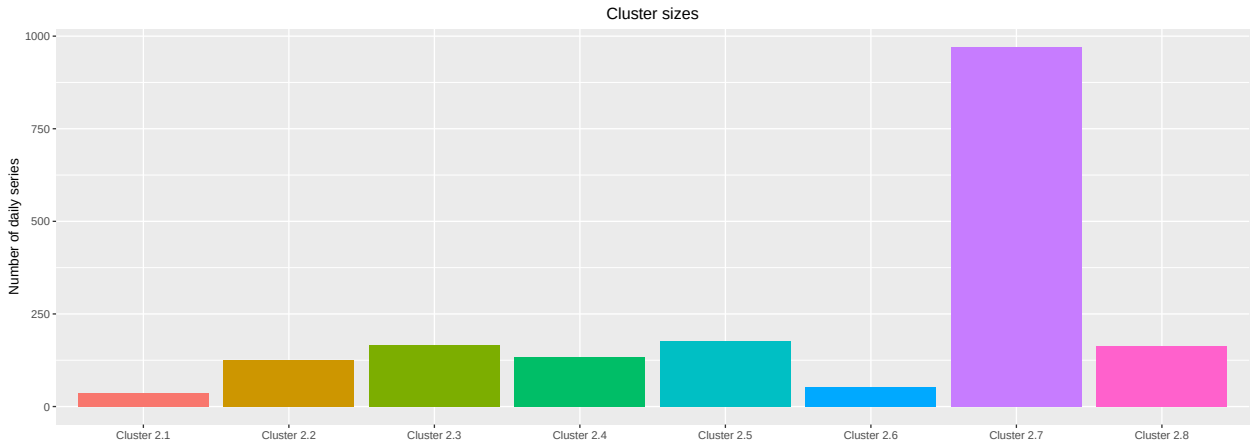

Figure C.13: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window clusters sizes.

Figure C.13 shows that the largest cluster is Cluster 2.7, which has about 971 daily patterns. The remaining clusters are in the range of 100 to 200 daily patterns. Except for Cluster 2.1 which has 36 daily patterns and cluster 2.6 with 53 daily patterns.

Compared to Figure C.5 of the Partitional Clustering with DTW model, PAM prototype and 15 minutes time window, it appears that Cluster 2.7 of the present model has more daily patterns than Cluster 2.5 of the Partitional Clustering with model. DTW, PAM prototype and 15 minutes time window. In the case of the remaining clusters of this model it is found that in general they aggregate less daily patterns than the clusters of the Partitional Clustering with DTW, PAM prototype and 15 minutes time window. Indicating that in the present model Cluster 2.7 aggregates more dailly patterns and the remaining Clusters are less representative than those of Figure C.5.

Table C.2 shows a set of statistical characteristics of the clusters formed:

Table C.2: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window clusters statistics.

Statistics	**Cluster 2.1 (m^3/h)**	**Cluster 2.2 (m^3/h)**	**Cluster 2.3 (m^3/h)**	**Cluster 2.4 (m^3/h)**	**Cluster 2.5 (m^3/h)**	**Cluster 2.6 (m^3/h)**	**Cluster 2.7 (m^3/h)**	**Cluster 2.8 (m^3/h)**
Min.	0.31	0.95	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	2.41	22.40	15.30	9.23	112.88	1.10	3.30	12.70
Median	3.94	28.40	53.03	17.45	132.32	6.12	9.35	31.20
Mean	15.16	31.97	56.84	27.07	121.71	23.18	24.08	35.67
3rd Qu.	6.76	34.40	76.00	23.30	147.54	19.91	25.80	48.80
Max.	530.50	766.37	376.42	881.00	184.24	340.00	981.25	503.75
IQR	4.35	12.00	60.70	14.07	34.66	18.80	22.50	36.10

Figure C.14 identifies the influence of weekend or holiday days have on the formation of clusters:

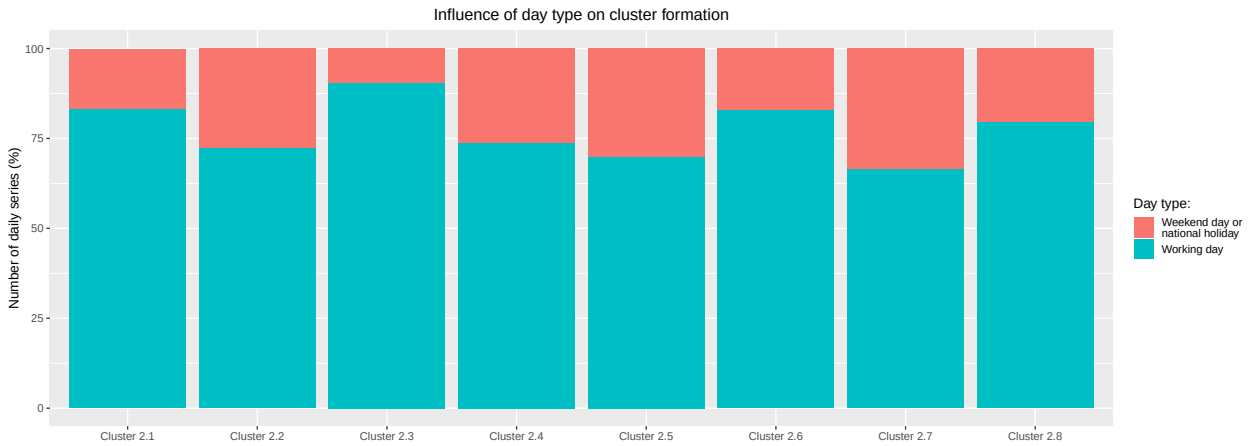

Figure C.14: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window influence of day typology on the formation of clusters.

As it can be seen, all clusters are associated with typical workday behavior.

Figure C.15 shows the geographic distribution of the clusters formed by the model:

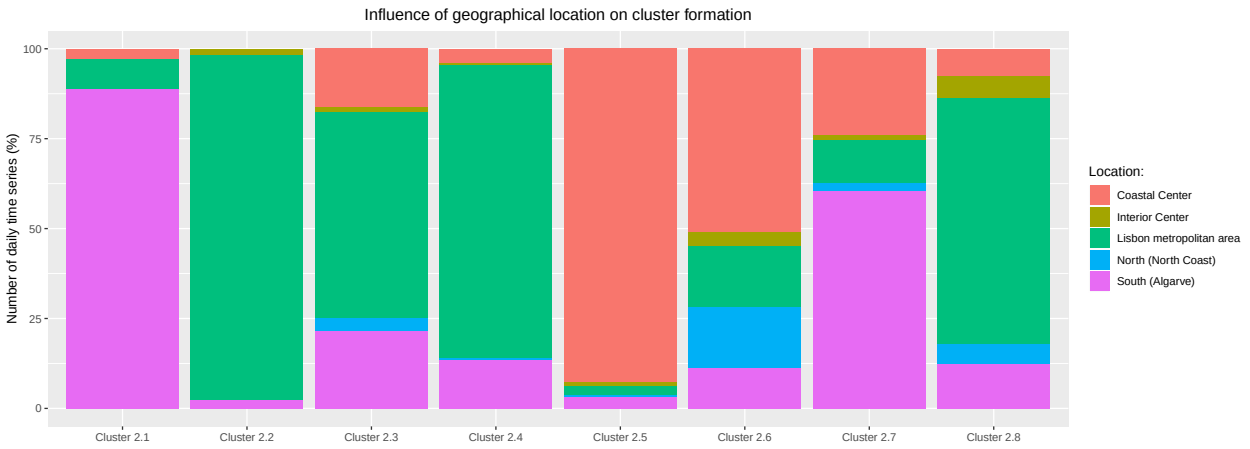

Figure C.15: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window - geographic distribution of the clusters formed.

Figure C.15 shows that Clusters 2.1 and 2.7 belong mainly to the South (Algarve) region. In the case of Clusters 2.2, 2.3, 2.4 and 2.8, they mainly belong to the Lisbon metropolitan area. Clusters 2.5 and 2.6 mainly belongs to the Costal Center region. The Interior Center region has little representation in these clusters, and is present in all clusters, except for the Cluster 2.1.

Figure C.16 shows the distribution of wet months and dry months in the clusters formed by the model:

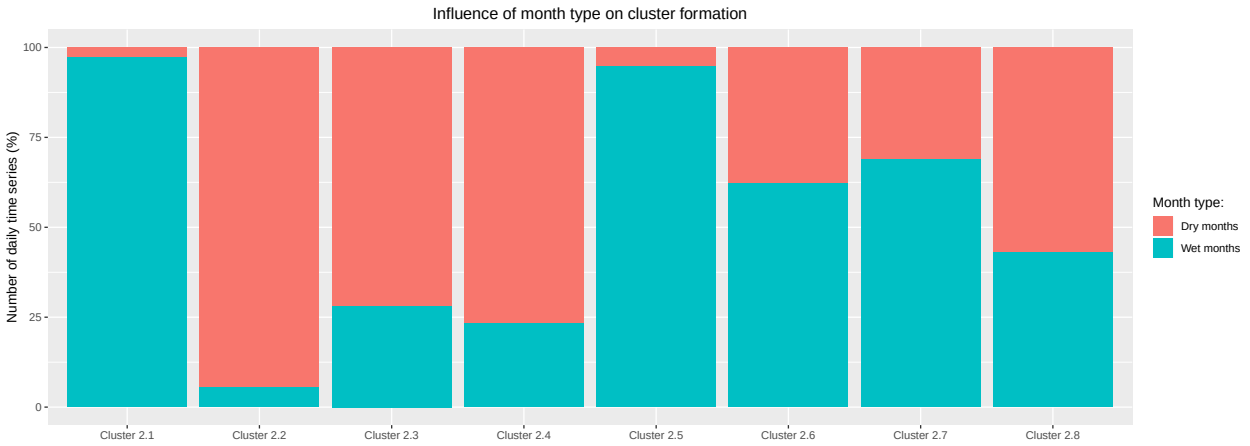

Figure C.16: Cluster 2 - Partition Clustering model with GAK, PAM and 15m window - distribution of wet months and dry months in the clusters formed.

Figure C.16 shows that Clusters 2.1, 2.5, 2.6 and 2.7 belong mostly to the dry months typology. Clusters 2.2, 2.3, 2.4 and 2.8 belong mainly to the wet months typology.

References

AGHABOZORGI, Saeed, Ali SEYED SHIRKHORSHIDI and Teh YING WAH, 2015. Time-series clustering - A decade review. *Information Systems* [online]. B.m.: Elsevier, **53**, 16–38. Available at: doi:10.1016/j.is.2015.04.007

ALEGRE, H, J M BAPTISTA, E Cabrera JR., F CUBILLO, P DUARTE, W HIRNER, W MERKEL and R PARENA, 2006. *Performance Indicators for Water Supply Services (Manual of Best Practice)*. B.m.: IWA Publishing. ISBN 1843390515.

ALLAIRE, JJ, Yihui XIE, Jonathan MCPHERSON, Javier LURASCHI, Kevin USHEY, Aron ATKINS, Hadley WICKHAM, Joe CHENG, Winston CHANG and Richard IANNONE, 2019. *Rmarkdown: Dynamic documents for r* [online]. Available at: `https://CRAN.R-project.org/package=rmarkdown`

ANDREOPOULOS, Bill, Aijun AN, Xiaogang WANG and Michael SCHROEDER, 2009. *A roadmap of clustering algorithms: finding a match for a biomedical application*. 2009.

APA, 2012. *Programa nacional para o uso eficiente da água - implementação 2012-2020*. B.m.: Agência Portugues do Ambiente, I.P.

ARBELAITZ, Olatz, Ibai GURRUTXAGA, Javier MUGUERZA, Jesús M. PÉREZ and Iñigo PERONA, 2013. An extensive comparative study of cluster validity indices. *Pattern Recognition* [online]. **46**(1), 243–256. Available at: doi:10.1016/j.patcog.2012.07.021

AUGUIE, Baptiste, 2017. *GridExtra: Miscellaneous functions for "grid" graphics* [online]. Available at: `https://CRAN.R-project.org/package=gridExtra`

BAGNALL, Anthony and Gareth JANACEK, 2005. Clustering time series with clipped data. *Machine Learning* [online]. **58**(2-3), 151–178. Available at: doi:10.1007/s10994-005-5825-6

BANERJEE, Arindam and Joydeep GHOSH, 2001. Clickstream Clustering Using Weighted Longest Common Subsequences. In: *In proceedings of the web mining workshop at the 1st siam conference on data mining*. pp. 33–40.

BARRELA, Rui, 2015. *Data reconstruction of flow time series in water distribution networks*. B.m. Master's thesis. Instituto Superior Técnico.

BENJAMINI, Yoav, 1988. Opening the Box of a Boxplot. *The American Statistician* [online]. B.m.: Taylor & Francis, **42**(4), 257–262. Available at: doi:10.1080/00031305.1988.10475580

BEZDEK, James C, 1981. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers. ISBN 0306406713.

BRADLEY, P S, Usama FAYYAD and Cory REINA, 1998. Scaling Clustering Algorithms to Large Databases. In: *Proceedings of the fourth international conference on knowledge discovery and data mining* [online]. B.m.: AAAI Press, pp. 9–15. KDD'98. Available at: `http://dl.acm.org/citation.cfm?id=3000292.3000295`

CALIŃSKI, Tadeusz and Harabasz JA, 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods* [online]. **3**, 1–27. Available at: doi:10.1080/03610927408827101

CANDELIERI, A., D. SOLDI, D. CONTI and F. ARCHETTI, 2014. Analytical leakages localization in water distribution networks through spectral clustering and support vector MACHINES. The icewater approach. *Procedia Engineering* [online]. **89**, 1080–1088. Available at: doi:10.1016/j.proeng.2014.11.228

C. DUNN, Joseph, 1973. A fuzzy relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Cybernetics and Systems*. **3**, 32–57.

CHEIFETZ, Nicolas, Zineb NOUMIR, Allou SAMÉ, Anne Claire SANDRAZ, Cédric FÉLIERS and Véronique HEIM, 2017. Modeling and clustering water demand patterns from real-world smart meter data. *Drinking Water Engineering and Science* [online]. **10**(2), 75–82. Available at: doi:10.5194/dwes-10-75-2017

CHEN-PAN, Liao, 2012. *File:Boxplot vs PDF.svg - Wikimedia Commons* [online]. 2012. [accessed. 2019-09-01]. Available at: `https://commons.wikimedia.org/wiki/File:Boxplot_vs_PDF.svg`

CHU, Selina, Eamonn KEOGH, David HART and Michael PAZZANI, 2013. Iterative Deepening Dynamic Time Warping for Time Series. In: [online]. pp. 195–212. Available at: doi:10.1137/1.9781611972726.12

COMINOLA, A., A. MORO, L. RIVA, M. GIULIANI and A. CASTELLETTI, 2016. Profiling residential water users' routines by eigenbehavior modelling. *8th International Congress on Environmental Modelling and Software*.

CORPORATION, Microsoft and Steve WESTON, 2018. *DoParallel: Foreach parallel adaptor for the 'parallel' package* [online]. Available at: `https://CRAN.R-project.org/package=doParallel`

CORRADINI, A, 2001. Dynamic time warping for off-line recognition of a small gesture vocabulary. In: *Proceedings ieee iccv workshop on recognition, analysis, and tracking of faces and gestures in real-time systems* [online]. pp. 82–89. Available at: doi:10.1109/RATFG.2001.938914

CUTURI, Marco, 2011. Fast Global Alignment Kernels. In: *Proceedings of the 28th international conference on machine learning, icml 2011*. pp. 929–936.

DAVE, Rajesh N., 1996. Validating fuzzy partitions obtained through c-shells

clustering. *Pattern Recognition Letters* [online]. **17**(6), 613–623. Available at: doi:10.1016/0167-8655(96)00026-8

ERSAR, 2013. *Non-revenue water in water supply systems corresponds annually to 167 millions euros, Press note*. Lisbon: ERSAR.

ESTER, Martin, Hans-Peter KRIEGEL, Jorg SANDER and Xiaowei XU, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: . B.m.: AAAI Press, pp. 226–231.

FARLEY, M, 2001. Leakage Management and Control. 1–98.

FRIGGE, Michael, David C HOAGLIN and Boris IGLEWICZ, 1989. Some Implementations of the Boxplot. *The American Statistician* [online]. B.m.: Taylor & Francis, **43**(1), 50–54. Available at: doi:10.1080/00031305.1989.10475612

GENTLE, J. E., L. KAUFMAN and P. J. ROUSSEUW, 2006. *Finding Groups in Data: An Introduction to Cluster Analysis.* [online]. B.m.: Wiley-Interscience. 2. ISBN 0471735787. Available at: doi:10.2307/2532178

GIORGINO, Toni, 2009. Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software* [online]. **31**(7), 1–24. Available at: doi:10.18637/jss.v031.i07

GURRUTXAGA, Ibai, Iñaki ALBISUA, Olatz ARBELAITZ, José I. MARTÍN, Javier MUGUERZA, Jesús M. PÉREZ and Iñigo PERONA, 2010. SEP/COP: An efficient method to find the best partition in hierarchical clustering based on a new cluster validity index. *Pattern Recognition* [online]. **43**(10), 3364–3373. Available at: doi:10.1016/j.patcog.2010.04.021

HAN, Jiawei, Micheline KAMBER and Jian PEI, 2011. *Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)* [online]. B.m.: Morgan Kaufmann. ISBN 9380931913. Available at: `https://www.amazon.com/Data-Mining-Concepts-Techniques-Management/dp/0123814790?SubscriptionId=0JYN1NVW651KCA56C102&tag=techkie-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0123814790`

HASTIE, T, R TIBSHIRANI and J FRIEDMAN, 2009. *The elements of statistical learning. Elements (Vol. 1), New York City, NY*. ISBN 9780387848570.

HAUTAMAKI, Ville, Pekka NYKANEN and Pasi FRANTI, 2009. Time-series clustering by approximate prototypes. In: [online]. pp. 1–4. D. ISBN 9781424421756. Available at: doi:10.1109/icpr.2008.4761105

KAMPSTRA, Peter, 2007. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software* [online]. **28**. Available at: doi:10.18637/jss.v028.c01

KAUFMANN, Leonard and Peter ROUSSEEUW, 1987. Clustering by Means of Medoids. *Data Analysis based on the L1-Norm and Related Methods*. 405–416.

KEOGH, Eamonn and Shruti KASETTY, 2002. On the need for time series data mining

benchmarks. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02* [online]. 102. Available at: doi:10.1145/775047.775062

KEOGH, Eamonn and M PAZZANI, 1998. An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback. *In: 4th International Conference on Knowledge Discovery and Data Mining.* [online]. 239–241. Available at: `https://www.aaai.org/Papers/KDD/1998/KDD98-041.pdf`

KIM, Minho and R. S. RAMAKRISHNA, 2005. New indices for cluster validity assessment. *Pattern Recognition Letters* [online]. **26**(15), 2353–2363. Available at: doi:10.1016/j.patrec.2005.04.007

KINGDOM, Bill, Roland LIEMBERGER and Philippe MARIN, 2006. *The Challange of reducing non-revenue water in developing countries* [online]. 8. Washington, DC: The World Bank. Available at: `http://siteresources.worldbank.org/INTWSS/Resources/WSS8fin4.pdf`

KRISHNAPURAM, Raghu, Anupam JOSHI, Olfa NASRAOUI and Liyu YI, 2001. Low-complexity fuzzy relational clustering algorithms for Web mining. *IEEE Transactions on Fuzzy Systems* [online]. **9**(4), 595–607. Available at: doi:10.1109/91.940971

KWON, S H, 1998. Cluster validity index for fuzzy clustering. *Electronics Letters* [online]. **34**(22), 2176–2177. Available at: doi:10.1049/el:19981523

L. DAVIES, David and Don BOULDIN, 1979. A Cluster Separation Measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* [online]. **PAMI-1**, 224–227. Available at: doi:10.1109/TPAMI.1979.4766909

LEGENDRE, Pierre, 2012. *Numerical ecology.* Amsterdam Boston: Elsevier. ISBN 9780444538680.

LIN, Jessica, Michail VLACHOS, Eamonn KEOGH and Dimitrios GUNOPULOS, 2004. Iterative Incremental Clustering of Time Series [online]. 106–122. Available at: doi:10.1007/978-3-540-24741-8_8

LOUREIRO, Dália, 2010. *Consumption analysis methodologies for an efficient management of water distribution systems.* B.m. PhD thesis. Instituto Superior Técnico.

LOUREIRO, Dália, Conceição AMADO, André MARTINS, Diogo VITORINO, Aisha MAMADE and Sérgio Teixeira COELHO, 2016a. Water distribution systems flow monitoring and anomalous event detection: A practical approach. *Urban Water Journal* [online]. **13**(3), 242–252. Available at: doi:10.1080/1573062X.2014.988733

LOUREIRO, Dália, Aisha MAMADE, Marta CABRAL, Conceição AMADO and Dídia COVAS, 2016b. A Comprehensive Approach for Spatial and Temporal Water Demand Profiling to Improve Management in Network Areas. *Water Resources Management* [online]. B.m.: Water Resources Management, **30**(10), 3443–3457. Available at: doi:10.1007/s11269-016-1361-3

LOUREIRO, Dália, Margarida REBELO, Aisha MAMADE, Paula VIEIRA and Rita

RIBEIRO, 2015. Linking water consumption smart metering with census data to improve demand management. *Water Science and Technology: Water Supply* [online]. **15**(6), 1396–1404. Available at: doi:10.2166/ws.2015.086

MACQUEEN, J, 1967. Some methods for classification and analysis of multivariate observations. In: *In 5-th berkeley symposium on mathematical statistics and probability*. pp. 281–297.

MAECHLER, Martin, Peter ROUSSEEUW, Anja STRUYF and Mia HUBERT, 2019. *Cluster: "Finding groups in data": Cluster analysis extended rousseeuw et al.* [online]. Available at: `https://CRAN.R-project.org/package=cluster`

MAHALAKSHMI, G., S. SRIDEVI and S. RAJARAM, 2016. A survey on forecasting of time series data. *2016 International Conference on Computing Technologies and Intelligent Data Engineering, ICCTIDE 2016* [online]. Available at: doi:10.1109/ICCTIDE.2016.7725358

MAMADE, Aisha, 2013. *PROFILING CONSUMPTION PATTERNS USING EXTENSIVE MEASUREMENTS A spatial and temporal forecasting approach for water distribution systems*. B.m. Master's thesis. Instituto Superior Técnico.

MARQUES, Ana, 2018. *Mathematical modeling of garden watering demand*. B.m. Master's thesis. Instituto Superior Técnico.

MATHWORKS, 2019. *Standardized z-scores - MATLAB zscore - MathWorks Benelux* [online]. 2019. [accessed. 2019-08-20]. Available at: `https://nl.mathworks.com/help/stats/zscore.html`

MATTEUCCI, Matteo, 2019a. *Clustering - Fuzzy C-means* [online]. 2019. [accessed. 2019-08-20]. Available at: `http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html`

MATTEUCCI, Matteo, 2019b. *Clustering - Hierarchical* [online]. 2019. [accessed. 2019-08-20]. Available at: `http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html`

MATTEUCCI, Matteo, 2019c. *Clustering - K-means* [online]. 2019. [accessed. 2019-08-20]. Available at: `http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html`

MINNEN, David, Thad STARNER, Irfan ESSA and Charles ISBELL, 2006. Discovering Characteristic Actions from On-Body Sensor Data Georgia Institute of Technology. *Work*. 0–7.

MINNEN, D, C L ISBELL, I ESSA and T STARNER, 2007. *Discovering multivariate motifs using subsequence density estimation and greedy mixture learning* [online]. 2007. Available at: `https://www.scopus.com/inward/record.uri?eid=2-s2.0-36348977475&partnerID=40&md5=0ea92d1e1c7e9e2a07b7481e5227028e`

MOUNCE, S R, W R FURNASS, E GOYA, M HAWKINS and J B BOXALL, 2016. Clustering and classification of aggregated smart meter data to better understand

how demand patterns relate to customer type. *Proceedings of the 14th International Conference of Computing and Control for the Water Industry - CCWI 2016*. 1–9.

MÜLLER, Kirill and Hadley WICKHAM, 2019. *Tibble: Simple data frames* [online]. Available at: `https://CRAN.R-project.org/package=tibble`

PAKHIRA, Malay K., Sanghamitra BANDYOPADHYAY and Ujjwal MAULIK, 2004. Validity index for crisp and fuzzy clusters. *Pattern Recognition* [online]. **37**(3), 487–501. Available at: doi:10.1016/j.patcog.2003.06.005

PANDEY, Padmakar, Akash CHAKRABORTY and G. C. NANDI, 2019. Efficient Neural Network Based Principal Component Analysis Algorithm. *2018 Conference on Information and Communication Technology, CICT 2018* [online]. B.m.: IEEE, 1–5. Available at: doi:10.1109/INFOCOMTECH.2018.8722348

PANUCCIO, Antonello, Manuele BICEGO and Vittorio MURINO, 2002. A Hidden Markov Model-Based Approach to Sequential Data Clustering. In: Terry CAELLI, Adnan AMIN, Robert P W DUIN, Dick DE RIDDER and Mohamed KAMEL, eds. *Structural, syntactic, and statistical pattern recognition: Joint iapr international workshops sspr 2002 and spr 2002 windsor, ontario, canada, august 6–9, 2002 proceedings* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 734–743. ISBN 978-3-540-70659-5. Available at: doi:10.1007/3-540-70659-3_77

PAPARRIZOS, John and Luis GRAVANO, 2015. K-shape: Efficient and accurate clustering of time series. *Proceedings of the ACM SIGMOD International Conference on Management of Data* [online]. **2015-May**(1), 1855–1870. Available at: doi:10.1145/2723372.2737793

PATIL, Kalpak, Naresh Kumar NAGWANI and Sarsij TRIPATHI, 2018. A Parametric Study of Partitioning and Density Based Clustering Techniques for Boxplot Generation. *2018 3rd International Conference for Convergence in Technology, I2CT 2018* [online]. B.m.: IEEE, 1–5. Available at: doi:10.1109/I2CT.2018.8529468

PETITJEAN, François, Alain KETTERLIN and Pierre GANÇARSKI, 2011. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition* [online]. **44**(3), 678–693. Available at: doi:10.1016/j.patcog.2010.09.013

RABINER, L, S LEVINSON, A ROSENBERG and J WILPON, 1979. Speaker-independent recognition of isolated words using clustering techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing* [online]. **27**(4), 336–349. Available at: doi:10.1109/TASSP.1979.1163259

RAI, Pradeep and Shubha SINGH, 2010. A Survey of Clustering Techniques. *International Journal of Computer Applications* [online]. **7**(12), 1–5. Available at: doi:10.5120/1326-1808

RANI, Sangeeta and Geeta SIKKA, 2012. Recent Techniques of Clustering of Time Series Data: A Survey. *International Journal of Computer Applications* [online]. **52**(15), 1–9. Available at: doi:10.5120/8282-1278

RATANAMAHATANA, Chotirat and E KEOGH, 2004. Everything you know about dynamic

time warping is wrong. In: .

R CORE TEAM, 2019. *R: A language and environment for statistical computing* [online]. Vienna, Austria: R Foundation for Statistical Computing. Available at: `https://www.R-project.org/`

ROUSSEEUW, Peter, 1987. Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Comput. Appl. Math. 20, 53-65. *Journal of Computational and Applied Mathematics* [online]. **20**, 53–65. Available at: doi:10.1016/0377-0427(87)90125-7

ROUSSEEUW, Peter and Christophe CROUX, 1993. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* [online]. **88**, 1273–1283. Available at: doi:10.1080/01621459.1993.10476408

SAITTA, Sandro, Benny RAPHAEL and Ian SMITH, 2007. A Bounded Index for Cluster Validity. In: [online]. pp. 174–187. Available at: doi:10.1007/978-3-540-73499-4_14

SAKOE, H and S CHIBA, 1971. A dynamic programming approach to continuous speech recognition. *Proceedings of the Seventh International Congress on Acoustics*. 65–69.

SAKOE, H and S CHIBA, 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing* [online]. **26**(1), 43–49. Available at: doi:10.1109/TASSP.1978.1163055

SARDA-ESPINOSA, Alexis, 2017. Comparing Time-Series Clustering Algorithms in R Using the dtwclust Package [online]. 1–41. Available at: `https://cran.r-project.org/web/packages/dtwclust/vignettes/dtwclust.pdf`

SARDA-ESPINOSA, Alexis, 2019. *Dtwclust: Time series clustering along with optimizations for the dynamic time warping distance* [online]. Available at: `https://CRAN.R-project.org/package=dtwclust`

SEGHOUANE, Abd Krim, Navid SHOKOUHI and Inge KOCH, 2019. Sparse Principal Component Analysis with Preserved Sparsity Pattern. *IEEE Transactions on Image Processing* [online]. B.m.: IEEE, **28**(7), 3274–3285. Available at: doi:10.1109/TIP.2019.2895464

SELA, Lina, Michael ALLEN, Ami PREIS, Mudasser IQBAL and Andrew J WHITTLE, 2015. Environmental Modelling & Software Automated sub-zoning of water distribution systems. *Environmental Modelling and Software* [online]. **65**, 1–14. Available at: doi:10.1016/j.envsoft.2014.11.025

SENIN, Pavel, 2016. *Z-normalization | SAX-VSM* [online]. 2016. [accessed. 2019-08-20]. Available at: `https://jmotif.github.io/sax-vsm_site/morea/algorithm/znorm.html`

SHEIKHOLESLAMI, Gholamhosein, Surojit CHATTERJEE and Aidong ZHANG, 1998. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: . pp. 428–439.

SILVA, Maria José, 2016. *Modelação da Incerteza e Deteção de Outliers para Melhoria do*

Diagnóstico de Perdas em Sistemas de Abastecimento de Água Matemática e Aplicações. B.m. Master's thesis. Instituto Superior Técnico.

SOLOMON, Nick, 2019. *Thesisdown: An updated r markdown thesis template using the bookdown package*.

SPINU, Vitalie, Garrett GROLEMUND and Hadley WICKHAM, 2018. *Lubridate: Make dealing with dates a little easier* [online]. Available at: `https://CRAN.R-project.org/package=lubridate`

STRATEGIC ALLIANCE FOR WATER LOSS REDUCTION, 2017. *Strategic Alliance for Water Loss Reduction - District Metered Areas (DMAs)* [online]. 2017. [accessed. 2019-06-23]. Available at: `http://www.waterloss-reduction.com/index.php/en/solutions/district-metered-areas-dmas`

TANG, Yuangang, Fuchun SUN and Zengqi SUN, 2005. Improved validation index for fuzzy clustering. *Proceedings of the American Control Conference* [online]. B.m.: IEEE, **2**, 1120–1125. Available at: doi:10.1109/ACC.2005.1470111

VLACHOS, Michail, Dimitrios GUNOPULOS and Gautam DAS, 2004. INDEXING TIME-SERIES UNDER CONDITIONS OF NOISE. In: *Data mining in time series databases* [online]. pp. 67–100. Available at: doi:10.1142/9789812565402_0004

VLACHOS, Michail, Jessica LIN, Eamonn KEOGH and Dimitrios GUNOPULOS, 2003. A Wavelet-Based Anytime Algorithm for K-Means Clustering of Time Series. In: *In proc. Workshop on clustering high dimensionality data and its applications*. pp. 23–30.

VLACHOS, M, G KOLLIOS and D GUNOPULOS, 2002. Discovering similar multidimensional trajectories. In: *Proceedings 18th international conference on data engineering* [online]. pp. 673–684. Available at: doi:10.1109/ICDE.2002.994784

VUORI, V and J LAAKSONEN, 2002. A comparison of techniques for automatic clustering of handwritten characters. In: *Object recognition supported by user interaction for service robots* [online]. pp. 168–171 vol.3. Available at: doi:10.1109/ICPR.2002.1047821

WANG, Weina and Yunjie ZHANG, 2007. On fuzzy cluster validity indices. *Fuzzy Sets and Systems* [online]. **158**(19), 2095–2117. Available at: doi:10.1016/j.fss.2007.03.004

WANG, Wei, Jiong YANG and Richard MUNTZ, 1997. STING: A statistical information grid approach to spatial data mining. In: . B.m.: Morgan Kaufmann, pp. 186–195.

WANG, Xiaozhe, Kate SMITH and Rob HYNDMAN, 2006. Characteristic-Based Clustering for Time Series Data. *Data Min. Knowl. Discov.* [online]. Hingham, MA, USA: Kluwer Academic Publishers, **13**(3), 335–364. Available at: doi:10.1007/s10618-005-0039-x

WARREN LIAO, T., 2005. Clustering of time series data - A survey. *Pattern Recognition* [online]. **38**(11), 1857–1874. Available at: doi:10.1016/j.patcog.2005.01.025

WICKHAM, Hadley, Winston CHANG, Lionel HENRY, Thomas Lin PEDERSEN, Kohske TAKAHASHI, Claus WILKE, Kara WOO and Hiroaki YUTANI, 2019a. *Ggplot2: Create elegant data visualisations using the grammar of graphics* [online]. Available at: `https:`

//CRAN.R-project.org/package=ggplot2

WICKHAM, Hadley, Romain FRANÇOIS, Lionel HENRY and Kirill MÜLLER, 2019b. *Dplyr: A grammar of data manipulation* [online]. Available at: https://CRAN.R-project.org/package=dplyr

WICKHAM, Hadley and Lionel HENRY, 2019. *Tidyr: Easily tidy data with 'spread()' and 'gather()' functions* [online]. Available at: https://CRAN.R-project.org/package=tidyr

WICKHAM, Hadley, Jim HESTER and Romain FRANCOIS, 2018. *Readr: Read rectangular text data* [online]. Available at: https://CRAN.R-project.org/package=readr

WILLIAMSON, D, R A PARKER and Juliette KENDRICK, 1989. The box plot: A simple visual method to interpret data. *Annals of internal medicine* [online]. **110**, 916–921. Available at: doi:10.1059/0003-4819-110-11-916

XIE, Yihui, 2019a. *Bookdown: Authoring books and technical documents with r markdown* [online]. Available at: https://CRAN.R-project.org/package=bookdown

XIE, Yihui, 2019b. *Knitr: A general-purpose package for dynamic report generation in r* [online]. Available at: https://CRAN.R-project.org/package=knitr

XIE, Yihui, 2019c. *Tinytex: Helper functions to install and maintain 'tex live', and compile 'latex' documents* [online]. Available at: https://CRAN.R-project.org/package=tinytex

ZAHID, N., M. LIMOURI and A. ESSAID, 1999. A new cluster-validity for fuzzy clustering. *Pattern Recognition* [online]. **32**(7), 1089–1097. Available at: doi:10.1016/S0031-3203(98)00157-5

ZHU, Hao, 2019. *KableExtra: Construct complex table with 'kable' and pipe syntax* [online]. Available at: https://CRAN.R-project.org/package=kableExtra