Water demand estimation and outlier detection from smart meter data using classification and Big Data methods

D. García[*, **], D. Gonzalez[**], J. Quevedo[*], V. Puig[*], J. Saludes[*]

[*] Center of Research in Supervision, Safety and Automatic Control (CS^{2}AC) of the Universitat Politècnica de Catalunya (UPC), Rambla Sant Nebridi, 22 08222 Terrassa, Barcelona, Spain
(E-mail: *diego.garcia@upc.edu*; *joseba.quevedo@upc.edu*; *vicenc.puig@upc.edu*; *jordi.saludes@upc.edu*)
[**] CETaqua, Carretera d'Esplugues, 75 08940 Cornellà de Llobregat, Barcelona, Spain
(E-mail: *dgarciava@agbar.es*; *dgonzalezv@cetaqua.com*)

Abstract
Automatic Meter Reading (AMR) systems are being deployed in many cities to obtain insight into the status and the behavior of District Metering Area (DMA) with more granularity. Until now, the water consumption readings of the population were taken one per month or one each two-months. In contrast, AMR systems provide hourly readings for households and more frequent readings for big consumers. On the one hand, this paper aims at predicting water demand and detect suspicious behaviors – e.g. a leak, a smart meter break down or even a fraud – by extracting water consumption patterns. On the other hand, the main contribution of this paper, a software framework, based on Big Data techniques, is presented to tackle the barriers of traditional data storage and data analysis since the volume of AMR data collected by Water Utilities is enormous and it is continuously growing because this technology is expanding

Keywords
Smart meters; water demand; clustering; big data;

1. INTRODUCTION

Water distribution networks aim at providing final clients with water from different sources. In order to manage efficiently these large complex networks, a big quantity of parameters must be measured (e.g. flows, pressures, demands, reservoirs' levels, etc.) throughout the whole network.

These networks must fulfill the water demand of citizens and industries. Therefore, demand forecast is an important and essential tool to anticipate and develop plans as to decide the best way of satisfying this demand based on different criteria, e.g. water source, tariff, energy consumption, etc.

Several research works propose methods for predicting demands and classifying them based on different approaches, including statistical models and machine learning models (e.g. Solanas et al. (2010, 2012); Aksela et al., 2011; McKenna et al., 2014).

When it comes to managing the network efficiently and preventing severe problems such as flooding, leakages or intrusions, the uncertainty in large-scale critical systems such as water networks poses a big risk. For instance, the World Bank has estimated the total cost of Non-Revenue Water (NRW) to utilities worldwide at US$14 billion per year.

For these reasons, among others, Automatic Meter Reading (AMR) systems are being deployed in many cities to get a real-time insight into the status and the behavior of District Metering Area (DMA). Until now, the water consumption readings of the population were taken one per month or one each two-months. In contrast, new smart meters provide hourly readings for households and more frequent readings for big consumers.

The number of parameters being measured (some of them with high frequencies), is becoming a big deal for traditional data warehouses. In contrast, Big Data provides new ways to process large quantities of data in parallel and in reasonable time. Thus, it allows extracting values, causes or events from the historical data that might have been overlooked. So far, historical data from SCADA systems is gathered in increasingly large databases due to the growing number of sensors. But, when new information does not lead to more insight, keeping historical data without clear purposes is just a waste of space and money.

The contributions of this work are twofold. On the one hand, we develop a Big Data technologies based framework (Hadoop[1] and Spark[2]) that provides an unsupervised classification of the demand patterns from smart meters data. Thus, Water Utilities can forecast the demand of a particular client, a group, or the DMA as a whole.

On the other hand, this framework features outlier detection (an outlier being related to a breakdown in a smart meter, a leakage or an unsuited smart meter). Thus, the NRW of Water Utilities will decrease. In addition, new fee systems based on the consumer's behavior could be implemented.

This framework has been implemented as a software application using Spark, a large-scale data processing engine. This Big Data engine is running over a reliable, scalable, distributed computing cluster of processing nodes, based on Hadoop. Hence, this architecture allows scaling up to thousands of processing nodes without modifying any software implementation.

2. METHODS

The methodology, depicted in Figure 1, is based on four general steps which are detailed below.

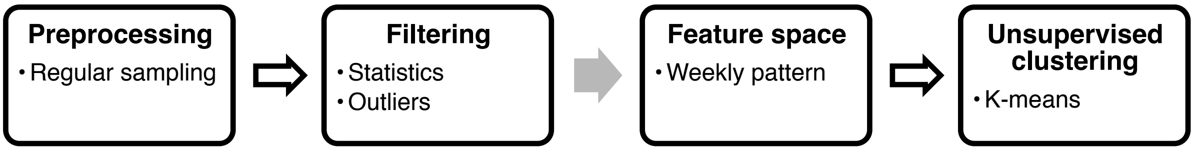

Figure 1. Steps to extract consumption patterns.

Preprocessing

Although each smart meter collects hourly data, in practice the interval is not exactly an hour: one observed demand is registered at 10:46 and the next one at 11:49. And these readings are not aligned to sharp o'clock hours (i.e. 10:00, 11:00, etc.). Moreover, some readings are missing mainly due to communication problems, among others. In this work, we have assumed the data collected is valid. The problem of data validation/reconstruction has already been addressed in (Garcia et al., 2014)

In this preprocessing stage a linear interpolation is applied to regularize the sampling interval and to align the sampling time to o'clock hours. Future work will consider better interpolation methods, but when only a few samples are missing a linear interpolation is enough.

Given a vector of observed demands of length N, $\boldsymbol{x} = (x_1, x_2, \dots, x_N)$, first we apply a linear interpolation method obtaining $\hat{\boldsymbol{x}} = (\hat{x}_1, \hat{x}_2, \dots, \hat{x}_p)$ with a regular sampling time.

[1] Hadoop provides, among other features, the MapReduce paradigm (Dean et al., 2008).

[2] Spark is a large-scale data processing engine (Matei et al., 2010).

$$\hat{x}_i^{(y)} = x_0^{(x)} + (x_1^{(y)} - x_0^{(y)}) \frac{\hat{x}_i^{(x)} - x_0^{(x)}}{x_1^{(x)} - x_0^{(x)}} \qquad (1)$$

where x_0 and $x_1$are the nearest irregular observations to $\hat{x}_i$ satisfying $x_0^{(x)} < x_i^{(x)} < x_1^{(x)}$.

Given the hourly demand vector $\hat{\boldsymbol{x}}$, we obtain the hourly consumption vector $\boldsymbol{z}$ applying the differences $z_t = (\hat{x}_{t+1} - \hat{x}_t)/T$ where T is the sampling time.

Filtering

This stage filters useful information and discards the useless one to be passed to the next stages. As mentioned before, a few missing observations can be estimated by a linear interpolation. We have set a maximum missing data threshold per week of 10%. Hence, only weeks with at least a 90% of data are processed by the following stages, otherwise they are discarded.

Once applied the previous filter, the following statistical indicators are estimated over the hourly consumption vectors $\boldsymbol{z}$: maximum, minimum, mean and variance. These indicators are used to discard smart meters which are always reading a constant consumption equal to zero, probably placed in empty houses. In addition, smart meters with negative readings are discarded, because backflow should not happen in the final points of DMAs.

Feature space

Different techniques for representing and reducing the dimensionality of time series has been proposed in the literature (Lin, J. et al., 2012), e.g. the Gaussian Mixture Models (McKenna et al., 2014) for representing water demands. In this paper, a feature vector $\boldsymbol{\omega_i} = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{i168})$ of length 168 (which is the number of hours of a week) represents the weekly pattern for a given smart meter i (see Figure 2 below) where each component k is given by

$$\omega_{ik} = \frac{\sum_{h(j)=k} z_j}{M_k}, \qquad (k \in [1{,}168]) \qquad (2)$$

where M_k is the number of observations that satisfies $h(j) = k$ and $h(j)$ returns the hour of the week of the datum's timestamp j.

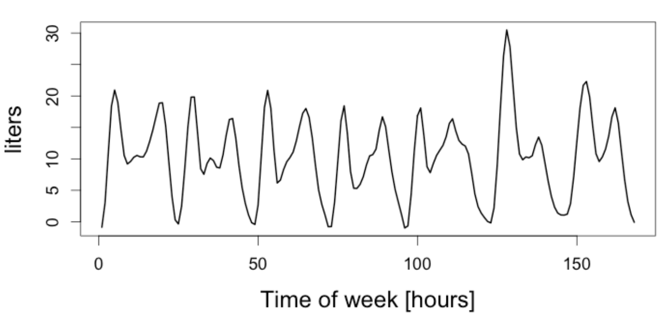

Figure 2. Weekly pattern demand (from Monday to Sunday) of sensor 6.

Unsupervised Clustering

The dataset used in this paper is composed by a set of smart meters and the associated observed demands. No additional information is available. Hence, as e.g. neither who the consumer is, nor the smart meter's diameter or activity are known.

The unsupervised clustering method applied in this framework is the *k*-means (Hartigan et al., 1979). This algorithm aims at partitioning *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

Therefore, the unknown parameter *k* must be estimated previously. As we will see in the application detailed in Section 4, a plot of the within groups sum of squares by number of clusters helps to determine the appropriate number of clusters.

3. FRAMEWORK

The volume of raw data generated by DMAs with AMR zones handled by Water Utilities' data centers is too big to be analyzed (e.g. R, Matlab, etc.) and stored (e.g. Oracle, MySQL, etc.) by means of traditional technologies. Even if with ad hoc and expensive solutions could handle the growth (increasing number of smart meters deployed each year) at the beginning, it would be unsustainable in the short-term. Thus, we propose a framework based on Big Data technologies to achieve a robust horizontal scalability independent of the data volume. Furthermore, all the technologies applied are open source in order to reduce the investment in expensive licenses.

The proposed Big Data framework, depicted in Figure 3, is compound by three modules (in columns). The *Storage* system is supported by Hadoop with the Hadoop Distributed File System (HDFS), where the raw data is collected. The *Processing* module applies the methodology detailed before and is based on a large-scale data processing engine called Spark. The Preprocessing and Feature space stages are implemented based on built-in Spark functions, but the linear interpolator is provided by Breeze. The Unsupervised clustering module, the *k*-means algorithm, is provided by MLlib, a scalable machine learning library on top of Spark. Finally, the results obtained from the *Processing* stage are saved in a distributed database called Cassandra.

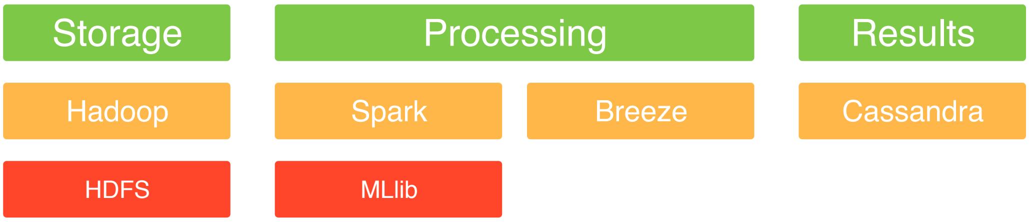

Figure 3. Big Data framework.

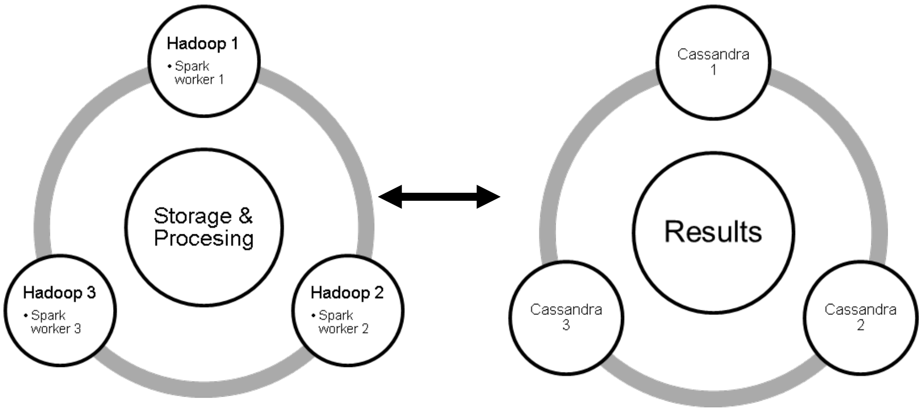

Figure 4. Big Data architecture.

Two clusters form the hardware architecture: one is used for the storage and processing stage and another one for saving the results and interacting with the user. This architecture is depicted in Figure 4 above.

4. APPLICATION

We present some results based on the Alicante city DMA (Spain). Alicante is a coastal city with a population of around 300 thousand people. The dataset used in this work is the hourly sampling observational readings from 51,117 smart meters of a one-year period (from July 2013 to July 2014). This dataset has 317,705,562 observed readings that corresponds to a size of 14 Gigabytes.

Figure 5 shows the mean (the left one) and the maximum (the right one) histograms, binned by 5 liters per hour. The right plot shows a big peak at the first bin [0-5] l/h. They are probably smart meters installed in empty houses or off-line water distribution pipes, or they could even be fraud (bypassing the smart meter).

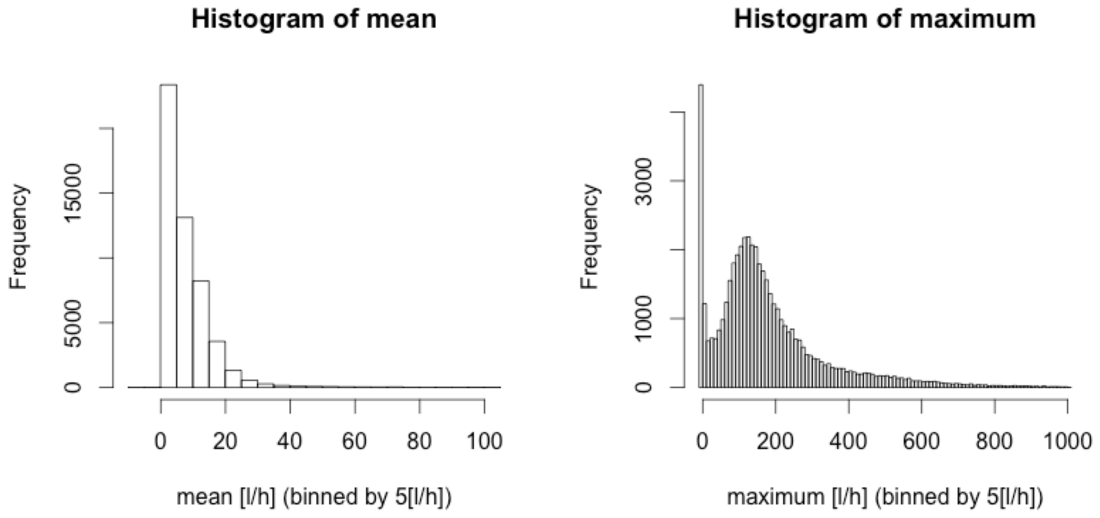

Figure 5. Mean and maximum statistics binned by 5[l/h].

The statistics filters listed in Table 1 above are applied discarding 6,914 smart meters. After applying this filter, weekly patterns are extracted following the methodology described previously (see Figure 2).

Table 1. Statistics filters applied.

Filter	Formula	Smart meters
Total consumption must be positive	Total>=0	78
Minimum hourly consumption must be positive	Minimum>=0	2,476
Maximum hourly consumption must be positive and other than zero	Maximum>0	4,360
Smart meters discarded		6,914

Due to the limitation of space for this paper and the difficulty to visualize thousands of results, we have considered 100 smart meters for illustrative purposes, shown in Figure 6, out of 51,117.

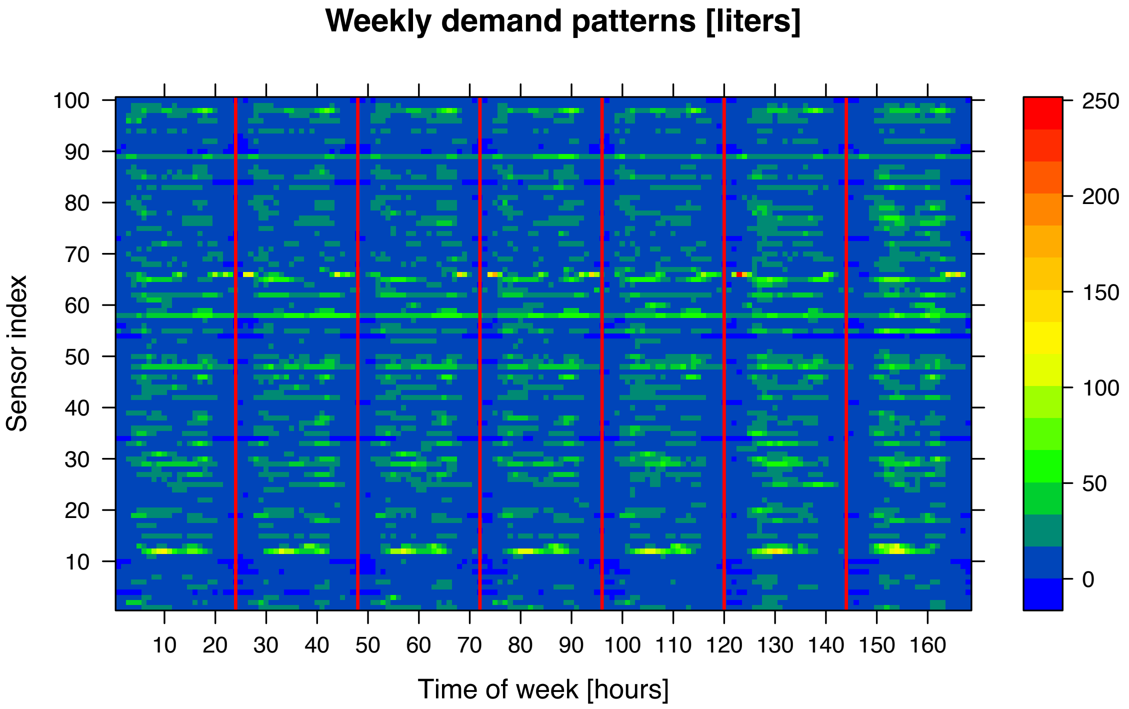

Figure 6. Sample of weekly average demands.

As it has been pointed out, the *k*-means algorithm requires the number of clusters input parameter. Hence, the within groups sum of squares is obtained (see Figure 7 below). Notice that after considering nine clusters the sum begins to be stable.

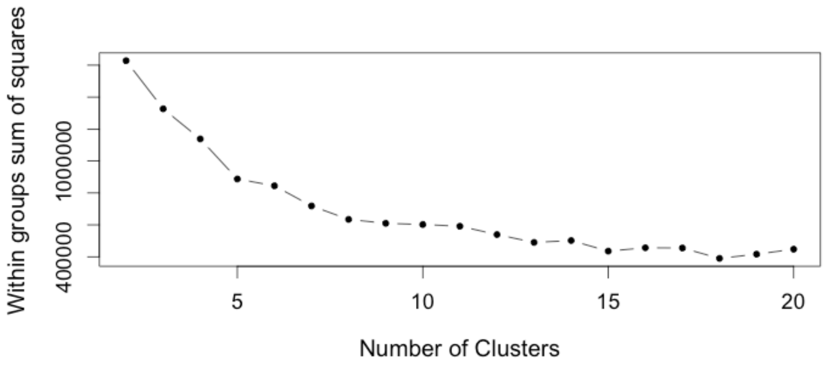

Figure 7. Within groups sum of squares by number of clusters.

The clusters obtained are shown in Figure 8. On the one hand, clusters 2, 3, 4, 5 and 6 have several members which allow a Water Utility to: forecast the demand of a client or the DMA, improve a leak detection model or detect a pattern change different to the pattern expected to the activity declared in the contract.

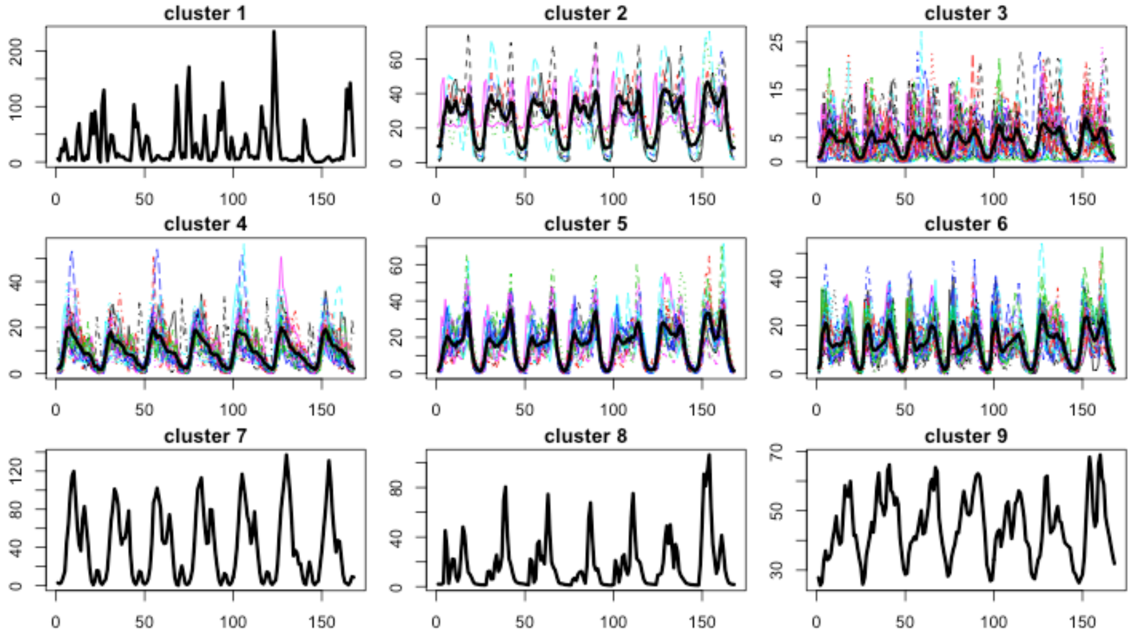

Figure 8. Weekly demand patterns classified (each cluster center is represented by a thick black line). The horizontal axis represents the time index (in hours) and the vertical axis represents the average consumption (in liters).

On the other hand, clusters 1, 7, 8 and 9 shows outliers (they have only one member) which must be analyzed with more detail. Cluster 1 is an irregular consumption pattern with several peaks between 100 to 200 l/h that could be generated by some irrigation or washing system. Therefore the Water Utility could change the type of fee assigned or could verify if the real activity of the client is not the one declared by the client in the contract (fraud). The pattern of cluster 7 shows a pattern with a regular daily shape but with an average consumption clearly higher in relation to the rest of clusters, thus the Water Utility can change the type of fee assigned or even give advice to this client in order to achieve a responsible water consumption. Cluster 9 shows the water demand pattern with the least number of consecutive low consumptions hours (valleys), probably due to a leak.

5. CONCLUSIONS

This paper presents a methodology to classify demand patterns based on AMR data. Weekly patterns extracted are classified using the *k*-means algorithm. The framework implemented, based on the methodology using Big Data techniques, is totally independent of the data volume and scalable to be adaptive to any expansion of AMR deployments in new zones. Finally, the proposed framework has been applied to the Alicante DMA, using a year-period dataset from 51,117 smart meters, thus obtaining a set of clusters. The results obtained show clusters compound by several water demand patterns that allow a Water Utility to forecast the water demand of a client or the DMA, and therefore to manage the network efficiently. The results obtained also show some clusters compound only by individual smart meters with different behavior that could be explained by other variables (not available for this work) such as consumer's activity, but could be generated by a breakdown in the meter, a leak or fraud.

As future research, this DMA pattern classification approach will be improved in order to be used for nodal estimation aimed at client demand monitoring or leak location. Moreover, the problem of placing new AMR in other DMAs will be addressed using as a starting point the results obtained using the DMA demand pattern classification approach presented here.

Acknowledgements
This work is supported by AGAUR Doctorate Industrial 2013-DI-041. We would also like to thank to "Aguas de Alicante", the Water Utility Company, for the dataset provided.

References

J. L. Solanas and M. R. Cussó (2010), "Multivariate consumption profiling (MCP) for intelligent meter systems: a methodology to define categories and levels", *Water Sci. Technol. Water Supply*, vol. 10, no. 5, p. 710, Dec. 2010.

J. L. Solanas and M. R. Cussó (2012), "MCP methodology for intelligent water metering (IWM): assessment of low flow consumption", *Water Sci. Technol. Water Supply*, vol. 12, no. 3, p. 270, May 2012.

A. W. Moore and M. C. Nechyba (1997), "Learning to recognize time series: combining ARMA models with memory-based learning", *Proc. 1997 IEEE Int. Symp. Comput. Intell. Robot. Autom. CIRA'97. 'Towards New Comput. Princ. Robot. Autom.*, pp. 246–251.

Lin, J., Williamson, S., Borne, K., & DeBarr, D. (2012), "Pattern recognition in time series". Advances in Machine Learning and Data Mining for Astronomy, 1, 617-645.

Karthikeyani Visalakshi Thangavel (2009), "Impact of Normalization in Distributed K-Means Clustering," *Int. J. Soft Comput.*, vol. 4, pp. 168–172.

A. Z. Mamade (2013), "PROFILING CONSUMPTION PATTERNS USING EXTENSIVE MEASUREMENTS: A spatial and temporal forecasting approach for water distribution systems". I. S. Técnico. General methodology Tools and applications.

X. Zhang, J. Liu, Y. Du, and T. Lv (2011), "A novel clustering method on time series data," *Expert Syst. Appl.*, vol. 38, no. 9, pp. 11891–11900, Sep. 2011.

S. A. McKenna, F. Fusco, and B. J. Eck (2014), "Water Demand Pattern Classification from Smart Meter Data," *Procedia Eng.*, vol. 70, pp. 1121–1130.

Hartigan, J., & Wong, M. (1979). "Algorithm AS 136: A k-means clustering algorithm". Applied Statistics, 28(1), pp. 100–108.

Aksela, K., & Aksela, M. (2011). "Demand estimation with automated meter reading in a distribution network". Journal of Water Resources Planning and Management, pp. 456–467.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), pp. 107-113.

Matei, Z., Chowdhury, M., J. Franklin, M., Shenker, S., & Stoica, I. (2010). "Spark: cluster computing with working sets". In HotCloud.

García, D., Quevedo, J., Puig, V., Saludes, J., Espin, S., Roquet, J., & Valero, F. (2014). "Automatic Validation of Flowmeter Data in Transport Water Networks: Application to the ATLLc Water Network". Intelligent Data Engineering and Automated Learning – IDEAL 2014 SE - 15 (Vol. 8669, pp. 118–125). Springer International Publishing. doi:10.1007/978-3-319-10840-7_15