AUTOMATIC, DYNAMIC PARAMETRIZATION OF CONSUMPTION PATTERNS IN WATER SUPPLY

A PARAMETRIZAÇÃO AUTOMÁTICA E DINÂMICA DE PADRÕES DE CONSUMO EM ABASTECIMENTO DE ÁGUA

Margarida Azeitona [a,*], *Sérgio Teixeira Coelho* [a], *Diogo Vitorino* [a]

[a] Baseform, R. Borges Carneiro 34 RC – 1200 Lisboa, Portugal
[b] Laboratório Nacional de Engenharia Civil, Av. Brasil 101 – 1700 Lisboa, Portugal

ABSTRACT

Several methodologies can be applied to predict the 24-hour demand pattern in district metering areas (DMA) of water distribution systems. However, in most cases, it is essential to know in advance the characteristics of each flow time-series, to choose the most suitable combination of parameters to use as the input of the method. In particular, the demand pattern derived is dynamic and may be influenced by the period of time considered. An adaptive predictive model of the pattern based on some selection criteria, which automatically choose the most appropriate parameterizations is described, in an attempt to overcome these drawbacks. This new approach is built on the concept of weighted percentiles, allowing the prediction to quickly adjust to changes in consumption habits, seasonality and other trends. The proposed procedure has undergone extensive testing and validated through the application to a large number of DMA, as preparation for use in a commercially deployed software.

Keywords – Water distribution systems, flow metering, demand pattern, adaptative models, changepoint detection.

RESUMO

Diversas metodologias podem ser aplicadas para prever o padrão de procura a 24 horas em zonas de medição e controlo (ZMC) de sistemas de distribuição de água. Contudo, na maior parte dos casos, é essencial conhecer de antemão as características de cada série temporal de caudal, para escolher a combinação de parâmetros mais adequada para o método. Em particular, o padrão de procura derivado é dinâmico e pode ser influenciado pelo período de tempo considerado. Numa tentativa de superar estes inconvenientes, é descrito um modelo adaptativo de previsão do padrão, baseado numa seleção de critérios, que escolhe automaticamente as parametrizações mais apropriadas. Esta nova abordagem é baseada no conceito de percentis pesados, permitindo às previsões ajustarem-se rapidamente a mudanças nos hábitos de consumo, sazonalidades e outras tendências. O procedimento proposto foi sujeito a testes exaustivos e validado através da aplicação a um grande número de ZMC, no âmbito da sua implementação para utilização num *software* comercial.

Palavras Chave – Abastecimento de água, medição de caudal, padrão de consumo, modelos adaptativos, deteção de *changepoint*s.

* *Autor para correspondência. Corresponding author.*
E-mail: margarida.azeitona@baseform.com

1 INTRODUCTION

The availability of network flow data in water utilities is steadily increasing as sensors and telemetry become more affordable. The modern practice of partitioning the systems into district metering areas (DMA) for greater detail contributes to a deeper understanding of the different components of urban consumption – domestic consumption, non-domestic consumption and water losses (Farley & Trow, 2003; Thornton, 2002). DMA are equipped with permanent or temporary network meters for continuous flow monitoring. In recent years, advances in remote metering technology and its wider availability at lower cost has made it effective to continuously monitor the most important consumers in the DMA (e.g., large consumers).

A number of techniques are applied to extract relevant flow information for use in leakage assessment, on-line burst event detection and operational control of the systems (Puust *et al.* 2010). These includes heuristic predictors as well as statistical tools that generate daily demand averages, minima, accumulated volume and other estimators.

Although the most common sense approach to representing a demand pattern is the 24-hour cycle of historically observed values (e.g., for network model use), the richness of information contained in the daily cycle behaviour is seldom used in this context. Failing to retain the value of the daily cycle as the key repeatable pattern of water consumption leads to effectively failing to extract useful information from the detailed available data.

There is also a widespread tendency to overlook the daily pattern significant variability with geographical location, consumer types and heterogeneity, presence of large consumers, as well as with the day of the week and the time of the year at each location.

2 THE CHALLENGE: THE 24-HOUR DEMAND PATTERN

The 24-hour demand pattern is expressed as a sequence of observed distributions of demand values at each instant of the day, sampled from a given range of dates, usually in the immediatelly contiguous past (Coelho, 1988; Alegre & Coelho, 1993) — as opposed to the more common pattern representation as a 24-hour single-line time series of some representative value, usually a median or a mean. The advantages of using the actual observed distributions include an accurate representation of the high-granularity historical data increasingly available at water utilities, and the efficiency in compressing that history, retaining rich detail.

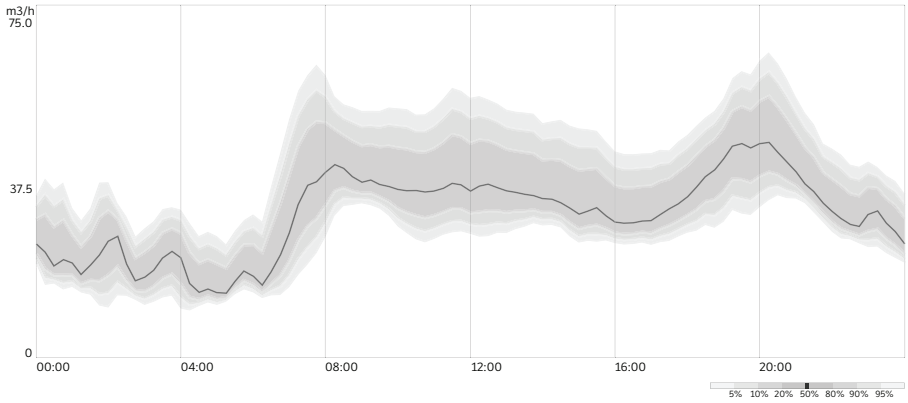

Figure 1. The 24-hour demand pattern: a sequence of observed population of demand values at each instant of the day (15 min. intervals, grey scale from percentiles 0%/100% – white, to 50% – black)

This model effectively encloses a representation, and quantification, of normality within a metered zone of the network. Different lines may be extracted from it, depending on the purpose of the analysis: the median is still a good representation of the most likely behaviour; the width of the 5%-95% percentile band can be used to assess variability, for example in testing the resilience of a particular network design; the 5% percentile is useful in water-loss auditing; etc.

The main value of this 24-hour demand pattern lies in its quantified depiction of normal behavior for that given set of consumers, for the network between them and the meter, and for the period of time during which the measurements were taken. The word normal is used here is its root meaning: *what is to be expected*.

The variability of measured consumption is influenced by a range of factors that should be taken into consideration when setting up an analysis:

i) noise in the data, usually generated by data acquisition problems, e.g. inadequately sized meters, transmission interference, data logger problems, etc.;

ii) consumer heterogeneity; a DMA is effective if its consumers are homogeneous and behave similarly. This also leads to measuring and removing demands from large consumers or those with local storage, such as hotels or hospitals;

iii) the size of the metered zone: the law of large numbers leads to an expectation that behaviour becomes more predictable as the number of individual consumers increases;

iv) the period(s) under analysis: consumers will behave differently depending on the day of the week and the time of the year; certain areas display significant climate- or human-related seasonality; periods of regular work may be interspersed with holidays or vacation periods; and network topology and operation (e.g., boundary limits, operating pressure, background leakage levels) may also change over time.

The proposed expression of the 24-hour demand pattern is designed to make no assumptions about the distribution of the observations of the time series. Avoiding normal, lognormal or other parametric distributions and using robust statistics helps prevent the unduly effects of outlier readings; however, limiting the variability and enhancing the predictability of the behaviour favors better analysis. The above factors should therefore be taken into account to seek the best expression of the behaviour of a metered zone.

Additionally, a better expression of normality can be obtained by filtering out abnormal events from the data used to build the 24-hour pattern. The identification of these abnormal events depends of the purpose of the analysis. If, for example, the 24-hour demand pattern is to be applied to identifying pipe bursts — as events that deviate from normality — in turn these readings should not feed the 24-hour demand pattern data; the same rationale can be used as basis to filter out atypical consumption or isolated network operation events.

The importance of obtaining adequate 24-hour demand patterns relates to the need to establish, with a certain amount of confidence, whether the behaviour of the consumption recorded at a given DMA, during a specific period, fulfils the requirements to be considered normal or, on the contrary, may reveal abnormal events, such as a leak, a pipe break, an unusual water use or other types of network problems.

In this context, 'normal' is often associated with a recurrency in time as a result of repeated human behaviour. But an accurate definition of normality depends on having an adequate predictive model based on previously recorded consumption, which provides an indication of what should be expected in the following consumptions. Then, if a new observation deviates from the confidence bands, there is evidence to believe that it is an abnormality.

Figure 2 illustrates the differences in daily flow patterns for Sundays and Mondays in winter and summer periods, for a DMA located in a system in the Greater Lisbon area. In particular,

the days represented correspond to February 8-9 and July 19-20 in 2015. This DMA presents a demarked seasonality and deriving a demand pattern for this area is greatly influenced by the period of time considered, since abrupt changes in the consumption habits may occur along time. This example shows that, even within the same DMA, there is no single static 24-hour demand pattern that adequately describes consumption (Vitorino *et al.* 2014).

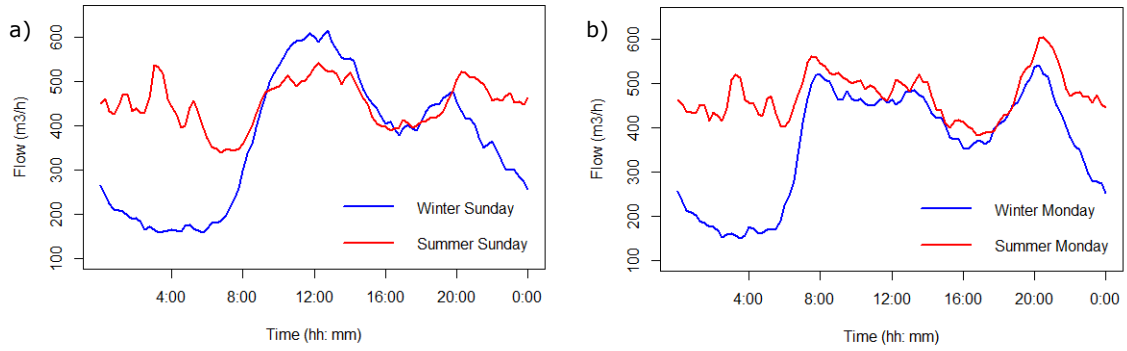

Figure 2. Daily flow patterns for winter and summer periods for a DMA: a) Sunday b) Monday.

3 THE PROPOSED APPROACH

3.1 Overview

The methods for detection of outlier flow events in DMA proposed by Loureiro *et al.* (2015) were adjusted for the estimation of 24-hour demand patterns. According to previous studies (Loureiro 2010, Tricarico *et al.* 2010), most DMA follow a periodic behaviour during the day and for that reason daily flow data were disaggregated into 96 15-min intervals. Moreover, the demand pattern predicted for each one of these instants was obtained from a set of representative previous observations, namely, data regarding the same instant in previous days of a specific type, within a fixed-width window.

As for the type of day, 3 grouping possibilities were considered: (A) all days were equally important; (S) only the same days were used; (W) days were grouped into workdays and weekend days. It is important to mention that holidays were considered as Sundays and thus, included in the weekend day's group. It was found that, in general, strategy (W) is the most adequate regardless of season of the year (Loureiro 2010, Tricarico *et al.* 2010). Regarding the time window, the following options were considered: use the data from the previous week, 2 weeks, 3 weks, 1 month, 2 months, 3 months, 6 months, year or 2 years.

For each estimation method, and having chosen a combination of day grouping strategy and time window, it was possible to obtain a different expression of the 24-hour demand pattern. However, in the absence of a method to select the most appropriate setting for each DMA, this wide range of possibilities may hinder the effectiveness of such demand patterns in commercial applications, such as the Meters app from Baseform software (Baseform, 2015). Thus, the development of an adaptive predictive model of the pattern based on some selection criteria, which automatically chooses the most appropriate options concerning the time window and grouping of days, has become imperative.

To accomplish this goal, a weighted percentile-based approach was considered, weighting the time distance and the effect of type of day for each observation.

One of the first steps of this process was the computation of a normality index, based on the same principles of normal Q-Q plots, with the aim of measuring the agreement between normal distribution and observed data. The calculation of this index for several DMA during

different time periods allowed concluding that a 3-month window was the most appropriate choice and thus, a reasonable starting point to consider as the maximum time range.

Furthermore, for several DMA an optimization of the weights assigned to each observation was performed to measure the gain that could be achieved with this new strategy. In this optimization setting, the objective function to minimize was the sum of the absolute differences between weighted medians and the corresponding observed values. In addition, the optimization of weights regarding the type of day (problem 1) and the history (problem 2) were considered separately and jointly. Problem 1 consisted in the optimization over 3 variables, namely, the weights assigned to all days (x_A), to days equal to the day to predict (x_S) and to days belonging to the same type (x_W). On the other hand, problem 2 was an optimization over 4 variables, the weights designated to days within a specific period (1 week, 2 weeks, 1 month, 3 months). In both cases, these weights were optimized for each day and for each month to evaluate the need to obtain specific weights for each day. The approaches used in addressing those two problems are described next.

3.2 Weighting the type of day

Water demand is dictated by human behaviour, which follows daily and weekly cycles, and for that reason water demand also presents the same type of patterns. As stated in the beginning of this section, 3 types of grouping the days were considered, however the challenge was to define a procedure to select the most suitable option for each DMA. In this context, it appeared to be a logical step to define a set of variables to characterize each day. The first variables considered were some basic descriptive statistics, like the minimum, the maximum and the median of the daily flow. Nevertheless, this kind of variables had the disadvantage of being scale dependent and for instance, the maximum was extremely instable. Additionally, the time of occurrence of these variables (t_{min} and t_{max}) were also computed. As for the median, the variable t_{med} was defined as the first time after the t_{min} in which the median value is registered. This last variable proved to be very useful to distinguish between types of days, since it appears to be related with the wake up time of the consumers of a given DMA. Figure 3 depicts these time variables for the same winter daily flow patterns considered in Figure 2. For the DMA analysed, the variable t_{med} (green line) presented similar values for the workdays, far enough apart (2h gap) from the ones on the weekend days. Thus, strategy (W) is the most adequate grouping of days.

In general, if the t_{med} of the days equal to the day to predict are far enough apart from the value obtained for the other days, then the weights are assigned according with strategy (S). Else, if the t_{med} of the workdays is significatively different from the t_{med} of the weekend days, then strategy (W) should be applied. Otherwise, if the t_{med} is approximately the same for all days, there is no need to separate the days.

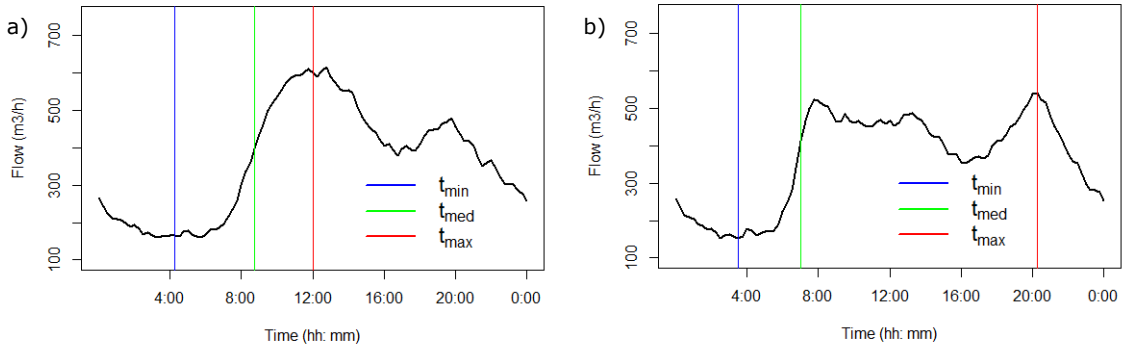

Figure 3. Time variables derived to characterize daily flow patterns in winter: a) Sunday, b) Monday.

3.3 Weighting recent history

After deciding the most suitable type of day grouping to apply, it is necessary to identify relevant periods in the recent history, to which should be given more weight. In particular, here the goal was to identify the location of multiple change points within the water demand time series. In this context, a changepoint is assumed to be a point at which the statistical properties (e.g. mean, variance, mean and variance) change. Over the years, several algorithms for changepoint detection have been proposed, namely binary segmentation, segment neighborhoods and the pruned exact linear time (PELT). However, in this article we only focus on the binary segmentation, one of the most popular changepoint detection methods. This algorithm consists in appling a single changepoint test statistic to the entire data, if a changepoint is found the data is divided into two at that point and the procedure is repeated on the two resulting data sets, until no changepoints are identified or the maximum number of changepoints to search for is achieved (Killick and Eckley, 2014).

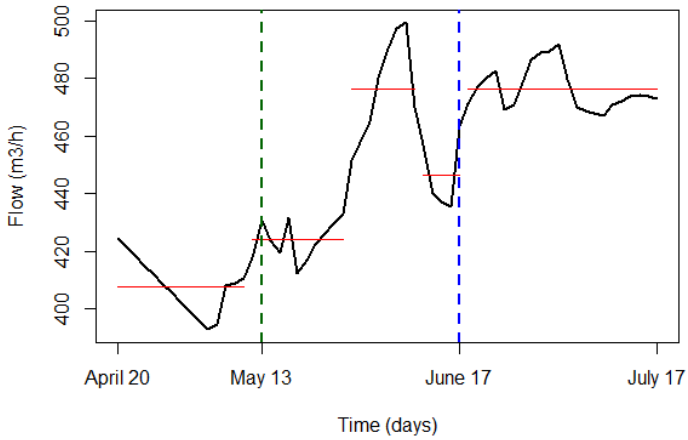

Figure 4. Change point detection for the workdays within the period April 19 – July 19.

The application of this changepoint detection algorithm to the 3 months history, with type of day grouping, allows finding inflexion points in the recent history. Then, starting from the more recent changepoint, the obtained points are used to define new segments, with increasing duration. Finally, weights are allocated to the days within these final segments so that each weight is inversely proportional to the duration of the period. It is important to mention that the detection algorithm is performed on the history after events' filtration, following a similar process to generate a filtered 24-hour demand pattern, as the one described in Vitorino *et al.* 2014.

The exemplification of the methods introduced above for July 20th, 2015, is described below. This date fell on a Monday and according to the previous subsection, the best grouping day's option was to consider only days of the same type, i.e. workdays. Thus, the changepoint analysis represented on Figure 4 was based on workdays only, after excluding days with events, resulting in 61 of the original 90 days. Initially, 4 changepoints were found however, according to the weighting procedure described, in this case only two changepoints will be used, which one defining periods lasting about 1 month. In particular, the period from June 17 to the previous day (July 19) includes the days to which more weight was assigned.

3.4 Computing the median and the pattern

Combining the weights selected in the previous steps yields a vector of weights, for each day in the previous 3 months, used as input in the estimation of the daily demand pattern based on weighted sample percentiles. It is worth mention that the same vector of weights is applied in the prediction of all the 96 15-min intervals within a day.

Consider n elements $x = (x_1, x_2, \ldots, x_n)$, with positive weights $w = (w_1, w_2, \ldots, w_n)$, such that the total sum of the weights is S. Then, the weighted median is defined as the element x_k, satisfying Eq. 1 and Eq. 2 (Cormen *et al.* 1989).

$$\sum_{x_i < x_k} w_i < \frac{S}{2} \qquad \text{(eq. 1)}$$

$$\sum_{x_i > x_k} w_i < \frac{S}{2} \qquad \text{(eq. 2)}$$

In general, the p^{th} weighted percentile is based on the same interpolated order statistic method as the corresponding unweighted sample percentile. The weighted percentiles algorithm considered here was based on the CRAN – Package Hmisc (2015) available in the statistical software R. This implementation combines the weights of equal observations before computing the estimates.

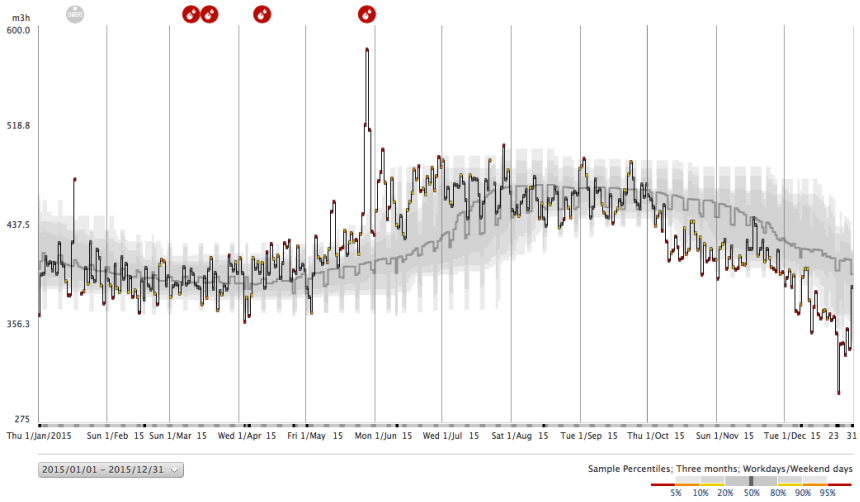

Figure 5. Charting the annual demand pattern estimated for the year 2015 with the sample percentiles-based approach (grey scale from percentiles 0%/100% – white, to 50% – black).

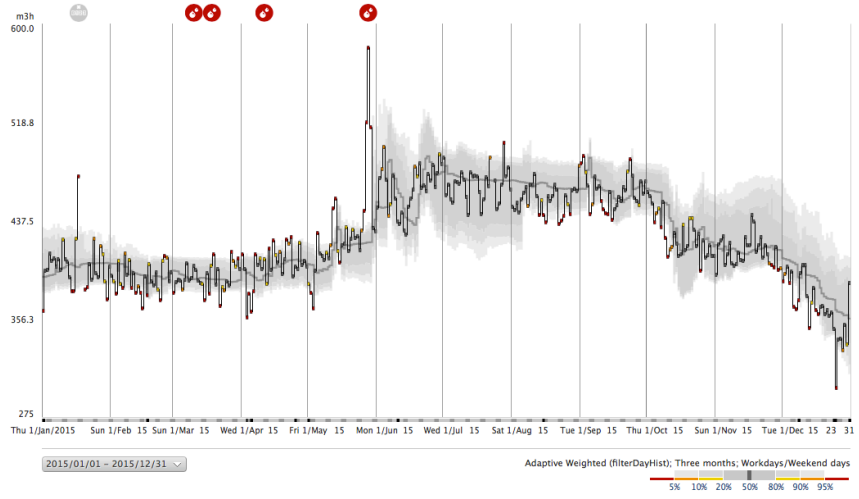

Figure 6. Charting the annual demand pattern estimated for the year 2015 with the weighted percentiles-based approach (grey scale from percentiles 0%/100% – white, to 50% – black).

Furthermore, the proposed approach is flexible and can be applied to different flow time series (e.g., instantaneous, hourly flow, daily flow, minimum night flow). In Figures 5 and 6 is represented the annual demand pattern estimated for the year 2015, in the previously

analysed DMA, with the sample percentiles-based approach and the new approach, respectively. It is clear the improvement achieved by the new approach based on the estimation of weighted percentiles. This kind of adaptive methods revealed to be particularly usefull in the modeling of DMA that, like this one, present a demarked seasonal behaviour due to climatic variations or human related seasonality (e.g. extra sprinkle water demand during the night).

4 DISCUSSION AND CONCLUSIONS

This paper presents a new approach for modeling water demand patterns in district metering areas (DMA) of water distribution systems. This new approach is based on the concept of weighted sample percentiles, allowing the prediction to quickly adjust to changes in the characteristics of each flow time-series. Moreover, this methodology does not require any previous knowledge on the time series' behavior out of the 3 months window and for that reason provides an adaptive framework to automatically choose the most suitable day grouping and time window size (number of past observations) to consider.

It is believed that the procedure developed will be a useful and effective addition to the range of tools currently available in commercial software, such as Baseform (2015), in use by various water utilities. This approach is currently undergoing extensive testing using data from 1500 flow meters, which together comprise more than 1000 years of data, so that the necessary adjustments can be performed and the method can be included in the software in the near future.

REFERENCES

Alegre, H., Coelho, S.T. (1993). *A methodology for the characterisation of water consumption*, In Integrated computer applications in water supply, Research Studies Press Ltd., pp. 369-384.

Baseform (2015). BF Software, Lda. http://baseform.com, accessed on May 2015.

Coelho, S.T. (1988). *A System for Demand Analysis and Forecasting in Water Supply Systems*. MSc dissertation, University of Newcastle upon Tyne.

Cormen T.H., Leiserson C.E., Rivest R.L. (1989). *Introduction to Algorithms*. The MIT Press. Massachusetts Institute of Technology.

CRAN – Package Hmisc (2015). *Hmisc: Harrell Miscellaneous. R package version* 3.17-1. Harrell Jr., F. https://CRAN.R-project.org/package=Hmisc, accessed on May 2015.

M. Farley and S. Trow, Losses in water distribution networks. A practitioner's guide to assessment, monitoring and control. Londres: IWA Publishing, 2003.

Killick R., Eckley I.A. (2014). *Changepoint: An R package for changepoint analysis*. Journal of Statistical Software, 58(3):1–19.

Loureiro D. (2010). *Consumption analysis methodologies for the efficient management of water distribution systems*. PhD Thesis PhD Thesis, Universidade Técnica de Lisboa, Lisbon, Portugal.

Loureiro D., Amado C., Martins A., Vitorino D., Mamade A., Coelho S. T. (2015). *Water distribution systems flow monitoring and event detection: a practical approach*. Urban Water Journal.

Puust, R., Kapelan, Z., Savic, D., Koppel, T. (2010). *A review of methods for leakage management in pipe networks*, Urban Water Journal, 7 (2010) 25-45.

Thornton, J. (2002). *Water Loss Control*. New York: McGraw-Hill.

Tricarico C., Marinis, G.d., Gargano, R. (2010). *Residential Water Demand-Daily Trends*. Water Distribution Systems Analysis, 1314–1323.

Vitorino D., Loureiro D., Alegre H., Coelho S., Mamade A. (2014). *In defense of the demand profile: a software approach*. Procedia Engineering (Science Direct), Vol 89, 2014, p.982–989, Elsevier