SCHOOL OF
MANAGEMENT AND
TECHNOLOGY
POLYTECHNIC OF
PORTO

MASTER
METHODS FOR BUSINESS DECISION MAKING

Anomaly Detection of Water Consumption
Marta Cristiana Barbosa Dos Santos

Marta Cristiana Barbosa Dos Santos. Anomaly Detection of Water Consumption

06/2021

SCHOOL OF
MANAGEMENT AND
TECHNOLOGY
POLYTECHNIC OF
PORTO

MASTER
METHODS FOR BUSINESS DECISION MAKING

Anomaly Detection of Water Consumption

Marta Cristiana Barbosa Dos Santos

06/2021

SCHOOL OF
MANAGEMENT AND
TECHNOLOGY
POLYTECHNIC OF
PORTO

M

MASTER
METHODS FOR BUSINESS DECISION MAKING

Anomaly Detection of Water Consumption

Marta Cristiana Barbosa Dos Santos

Supervisors

PhD Ana Isabel Coelho Borges

PhD Davide Rua Carneiro

PhD Flora José Rocha Ferreira

"A persistência é o caminho do êxito."
Charles Chaplin

Agradecimentos

Mais uma etapa que termina e chega a hora de agradecer a todos os que estiveram lá e de uma maneira ou de outra contribuíram para conseguir chegar aqui.
Um obrigada à Águas do Norte pelo desafio proposto e me darem a possibilidade de trabalhar com um problema real.
Sem eles nada disto seria possível, um enorme agradecimento aos meus orientadores. Professora Ana Borges obrigada por acreditar em mim, por acompanhar todo o meu percurso, por toda a dedicação prestada, partilha de conhecimento e por todas as palavras de incentivo e motivação ao longo destes anos. Ao professor Davide Carneiro, o meu enorme obrigada por ter aceite o convite em fazer parte deste desafio, pela dedicação, disponibilidade e partilha de conhecimento ao longo deste tempo. À professora Flora Ferreira, o meu profundo agradecimento por ter confiado em mim, pela paciência, disponibilidade e por toda a partilha ao longo desta caminhada. Sou grata por ter o privilégio de trabalhar com pessoas que me ensinaram tanto como vocês!
A todo o corpo docente do mestrado, um obrigada por partilharem a vossa experiência e sabedoria e nos fazerem evoluir dia após dia.
À minha família que foi incondicional e sempre me apoiou em todas as minhas decisões, o meu muito obrigada por me motivarem a alcançar sempre mais e estarem sempre lá para mim.
À Patrícia, ao Emanuel e ao Rafael o meu profundo obrigada por estarem sempre presentes, pelas horas de estudo que passamos juntos e por terem sempre uma palavra de motivação nas horas menos boas.
À Fátima e à Rosa, obrigada por me acompanharem nesta caminhada e estarem sempre prontas a ajudar.
Aos meus amigos, obrigada por estarem presentes, pela paciências e pelas palavras de incentivo.
Por fim, um obrigada à Escola Superior de Tecnologia e Gestão, por tudo o que tive oportunidade de viver e aprender ao longo desta minha caminhada que termina aqui.

O meu sincero obrigada!

Resumo

Com os aumentos substanciais de perdas de água tem existido uma forte preocupação no desenvolvimento de soluções sustentáveis de combate às perdas. As empresas de fornecimento de água enfrentam novos desafios e procuram soluções para reduzir as perdas, minimizando os custos inerentes, aumentando a qualidade dos serviços prestados e tornando-se mais sustentáveis e competitivas.
Este projeto foi desenvolvido em colaboração com uma empresa de fornecimento de água, que pretende minimizar as suas perdas com a deteção de alterações significativas nos consumos de água dos seus clientes. Para auxiliar a empresa na deteção de anomalias nos consumos de água e a tomar decisões oportunas de forma a minimizar essas perdas, foi construído um *dashboard* no software R que permite o *upload* de um arquivo CSV com o histórico de consumo, e devolve ao utilizador mudanças bruscas detetadas nos padrões de consumos. Os métodos implementados na *dashboard* foram inicialmente analisados e testados. Este estudo preliminar está reportado em 2 artigos publicados em *proceedings* indexados no Scopus e que foram apresentados em conferências internacionais.
O primeiro artigo incide sobre a aplicação e avaliação de diferentes métodos para a detecção de pontos de mudança em consumos de água. Para tal foram analisados dois datasets reais fornecidos pela empresa, que correspondem aos registos do consumo de água num período de um ano de um hotel e um hospital situados a norte de Portugal.
No segundo artigo, de forma a validar os métodos utilizados no artigo anterior e como não existiam garantias que os pontos de mudança detetados correspondiam a anomalias ou quebras nos consumos de água, recorreu-se à simulação de conjuntos de dados com base em distribuições probabilísticas sugeridas na literatura. Nesses conjuntos de dados foram inseridas quebras nos consumos de 5%, 10%, 15% e 20% na magnitude dos parâmetros, resultando um total de 1200 conjunto de dados simulados. Com estes novos dados foi elaborada uma análise de performance utilizando o RMSE (Raiz do Erro Quadrático Médio) e MSE (Erro Quadrático Médio) para determinar a eficácia dos métodos de deteção de pontos de mudança na deteção de quebras nos consumos de água.

Palavras chave: Perdas de Água, Perdas Aparentes, Análise de séries temporais, Deteção de Anomalias, Apoio à tomada de decisão

Abstract

With the substantial increase in water losses, there has been a strong concern towards the development of sustainable solutions to combat losses. Water supply companies face new challenges and seek solutions to reduce losses, minimizing inherent costs, increasing the quality of services provided, and becoming more sustainable and competitive.
This project portrays the case of a water distribution company, which intends to minimize its losses by detecting significant changes in its customers' water consumption. To assist the company in detecting anomalies in water consumption and taking timely decisions in order to minimize these losses, a dashboard in R software was created that allows the upload of a CSV file with the consumption history and returns sudden changes to the user detected in consumption patterns. The methods implemented in the dashboard were initially analyzed and tested. This preliminary study is reported in 2 articles published in proceedings indexed in Scopus and presented at international conferences.
The first article focuses on the application and evaluation of different methods for detecting points of change in water consumption, having analyzed two real datasets provided by the company, which correspond to the records of water consumption in a period of one year of a hotel and a hospital located in the north of Portugal.
In the second article, in order to validate the methods used in the previous article and as there were no guarantees that the detected points of change corresponded to actual anomalies or breaks in water consumption, synthetic datasets generated based on probabilistic distributions suggested by the literature were used. In these datasets, drops in consumption of 5%, 10%, 15% and 20% were inserted in the magnitude of the parameters, resulting in a total of 1200 simulated datasets. With these synthetic datasets, a performance analysis was performed using the RMSE (Root Mean Squared Error) and MSE (Mean squared error) to determine the effectiveness of the change point detection methods in detecting drops in water consumption.

Keywords: Water losses, Apparent losses, Time series analysis, Anomaly detection, Support decision making

Contents

List of Figures

Acronyms

AIC - Akaike Information Criterion
BIC - Bayesian information criterion
BS - Binary Segmentation
IWA - International Water Association
LSTM - Long Short -Term Memory
MSE - Mean Square Error
RMSE - Root Mean Square Error
NRW - Non Revenue Water
PELT - Pruned Exact Linear Time
STL - Seasonal Decomposition of Time Series by Loess method

Chapter 1

Introduction

Water is the most precious asset in human life. The droughts and the rapid urban population development and the negative impact that is being observed due to climate change, have led to an abrupt decrease in water resources available on the planet [1]. There is growing concern about sustainability and equal access to safe drinking water, on the 2030 agenda of the United Nations [2]. With a strong concern for sustainability and the significant increase in water losses, companies seek to implement systems for controlling and assessing losses [3], as these can significantly affect the availability of water resources in the future, as well as the quantity/quality of water to be used and the quality of services provided to customers [4]. Water losses in supply systems can be around 50% [5] worldwide and according to the International Water Association's water balance [6], there are three types of losses in water systems, the unbilled authorized consumption, the apparent losses and the real losses (see Figure 1.1).

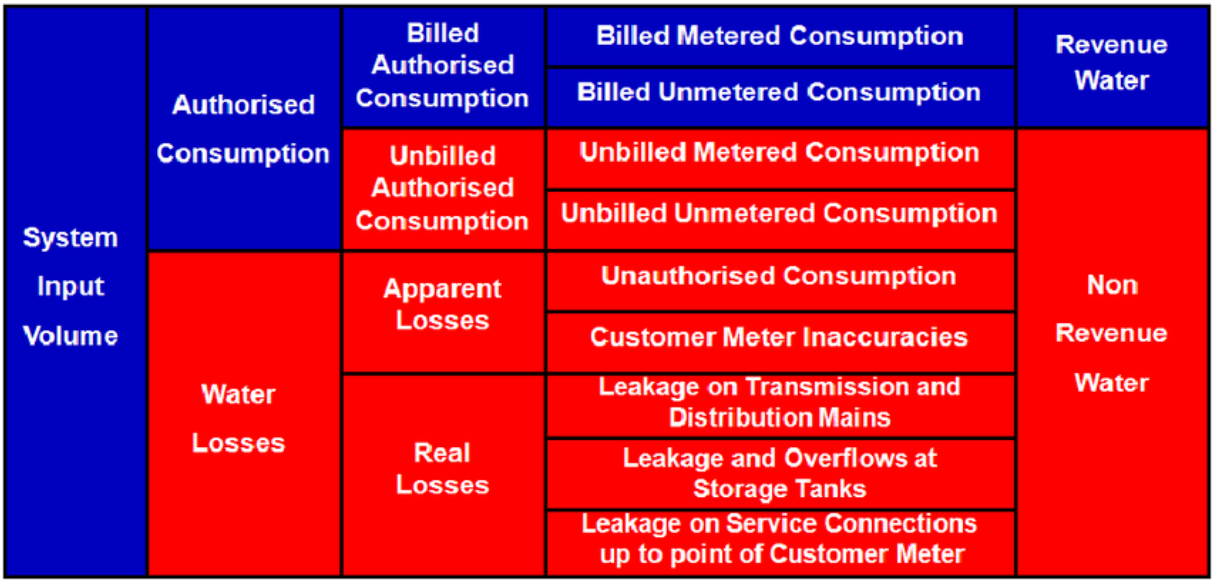

Figure 1.1: IWA Water Balance

The unbilled authorized consumption of water is that which is accounted for but not invoiced by the company and is assigned for social purposes.

Real losses are losses that occur along the distribution chain, as in valves, pipes and other components [7]. These are the most difficult losses to be accounted for.
Financial losses, also known as apparent losses, are one of the most important components of non-revenue water (NRW) and are divided into Unauthorised Consumption and Customer Meter Inaccuracies.
Unauthorized Consumption is where consumers are able to illegally reduce their billing consumption by tampering with water meters. These adulterations can range from direct connections to the removal and/or contouring of the water meter, with the greatest probability of adulteration occurring with older water meters, as they are composed of a piston or oscillating disk through the placement of a strong magnet, which will cause the speed of the meter to decrease and consequently the consumption measured is reduced [8]. It is a difficult technique to be detected by the concessionaires, as there is no change in the meter.
Measurement errors can be avoided if companies adopt strategies for the replacement of water meters [9] thus avoiding wear and tear, lack of maintenance and calibration that cause them to lose their efficiency throughout time [10]. These errors can also be associated with an incorrect installation or choice of the type, class or size of the water meter [11].
The volumes of apparent losses tend to be smaller than the volumes of real losses, but in terms of value for the distribution companies, they are identical. It is estimated that $1m^3$ of water lost in the distribution network has a cost for the company equal to the cost of production, while $1m^3$ that is provided to the customer but is not registered by the accountant, has a cost to the company equal to the last m^3 charged to that customer [4].
Due to increased concerns about the scarcity of water resources, water losses have been the subject of an increasing number of publications and studies to minimize them.

1.1 The Company and Case Study

Águas do Norte [12] is the management entity responsible for the multi-municipal water supply and sanitation system in northern Portugal.
The main objective of the company is to minimize its losses and ensure the best quality in the provision of its services to its customers. For this, the company is interested in a system that allows the detection of abrupt anomalies in water consumption, which may be associated to several factors, such as fraud, leakage, losses in the distribution networks, lack of calibration or deterioration of the meters.
For this study, two databases provided by the company were used, which correspond to the consumption of a hotel and a hospital in the north of Portugal in a time span of 365 days between December 2018 and November 2019.

1.2 Contributions of this Project

There has been a great concern about the scarcity and sustainability of water resources, which leads to a growing number of publications, demonstrating a great importance for the topic under

study.
With the decrease or timely detection of possible anomalies in the consumptions of its users, Águas do Norte will be able to provide a better quality service in supply and minimize its losses in revenues.
With the ultimate goal of facilitating the anomaly detection process, the main contributions of this project are as follows:

- Analysis of consumption patterns of two company clients.
- Implementation and evaluation of different change points methods in detecting anomalies in water consumption based on real data provided by the company.
- Creation of a user-friendly dashboard to assist in the company's water consumption analysis process, thus facilitating the detection of water consumption losses and assisting in decision-making.
- Using the dashboard, any anomaly can be detected more easily, thus reducing the time for water losses, maximizing the company's revenues and the quality of its services.
- Synthetic datasets generated based on real data using probabilistic distributions suggested by the literature and application of consumption breaks. These synthetic datasets were used to analyze the methods of detection of change points related to a break in the consumer's water consumption patterns.
- The datasets generated in this study can be useful to analyze water consumption patterns.
- Provide methodologies to generate new dataset's in several areas based on real data.

1.3 Structure

This dissertation is structured with two articles that were developed to respond to the problem presented by the company Águas do Norte, and another chapter with the description of the developed dashboard, is organized into 5 chapters as described in Figure 1.2.

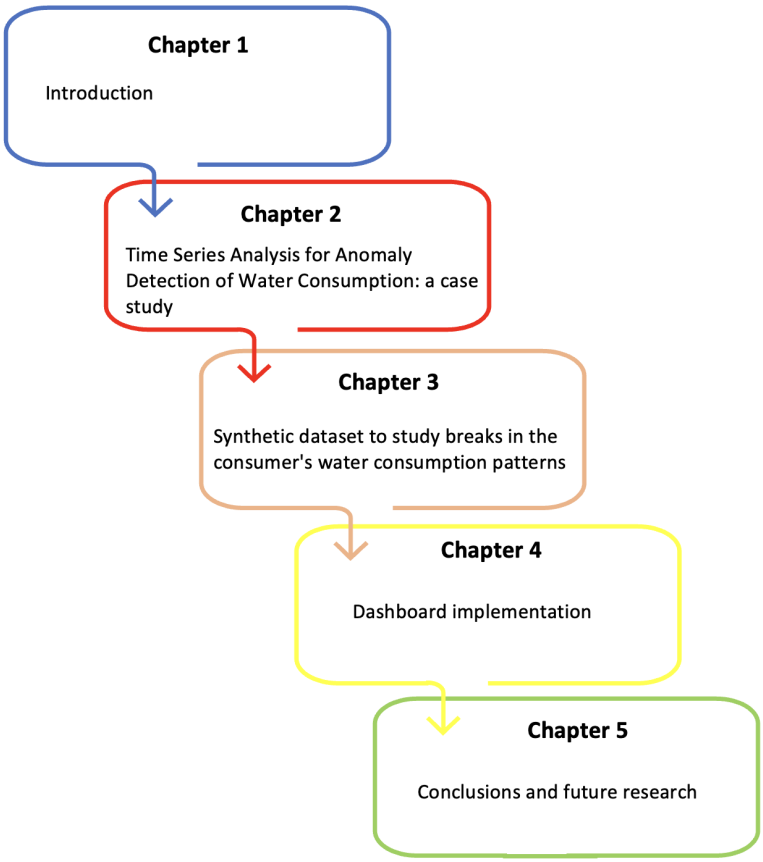

Figure 1.2: Structure of the project

Chapter 2 consists of an article entitled "Time Series Analysis for Anomaly Detection of Water Consumption: a case study" [13] presented at the ICIE - International Conference - INNOVATION IN ENGINEERING (June 28-30, 2021). The article presents a literature review on types of losses that exist in water consumption and what can lead to faults in water meters. In this study, two daily datasets provided by the company were analyzed, where consumption patterns were explored, and methods for detecting points of change in the time series were applied.

Chapter 3 includes the article "Syntehetic dataset to study breaks in the consumer's water consumption patterns" presented at the 3^{rd} International Applied Mathematics, Modeling and Simulation Conference (AMMS 2021, June 24-26, 2021). In this study, dataset simulation was applied using probabilistic distributions suggested by the literature that best fit water consumption. This study was designed to validate the methods applied in the previous article, since there was no guarantee that the detections made by the methods would actually detect breaks in the consumption of the water meters. Initially, the datasets provided by the company were fitted to five probabilistic distributions: Weibull, gamma, log-logistics, normal and exponential. For those that had a better fit, datasets were simulated with decreasing mean in intervals of 5%, 10%, 25% and 50% of water consumption. A total of 1200 datasets were simulated, where methods of change point detection were applied and the performance of those methods was determined through the calculation of RMSE and MAE.

Chapter 4 presents the dashboard, its structure and implementation. Although still in demo version it already allows the detection of anomalies in consumption and the user is provided with a set of information such as the slope of the consumption decrease, as a proxy for the magnitude of water loss, that can help in the decision making.
Finally, in the fifth chapter the main study conclusions and future suggestions for research and improvement of this work, are presented.

Chapter 2

Time Series Analysis for Anomaly Detection of Water Consumption: a case study

Time Series Analysis for Anomaly Detection of Water Consumption: A Case Study

Marta Santos[1], Ana Borges[1], Davide Carneiro[1], and Flora Ferreira[2( )]

[1] CIICESI, ESTG, Politécnico do Porto, Rua do Curral, Casa do Curral, Margaride, 4610-156 Felgueiras, Portugal
{8150180,aib,dcarneiro}@estg.ipp.pt
[2] Center of Mathematics, University of Minho, Campus de Azurém, 4800-058 Guimarães, Portugal
fjferreira@math.uminho.pt

Abstract. Water loss is one of the factors that most affect a concessionaire's financial sustainability. Early detection of any anomaly in water consumption is very valuable. This article aims to carry out a preliminary study to detect change points in consumption associated with water meter malfunction. The dataset is composed of water consumption measurements of two different companies (a hotel and a hospital) located in the north of Portugal, obtained during a complete year. Different methods were implemented in order to study its effectiveness in the detection of change points in the time series related to a sharp decrease in water consumption. Results suggest that the Seasonal Decomposition of Time Series by Loess method (STL) and the combination of several breakpoint detection methods is a suitable approach to be implemented in a software system, in order to help the company in anomaly detection and in the decision-making process of substituting the water meters.

Keywords: Time-Series Decomposition · Breakpoint detections · Water losses · Water meters

1 Introduction

Water is an essential good for human life, as water resources are increasingly scarce, so it is necessary to use measures to bridge current and future gaps in water supply and demand [1].

Water loss could have multiple causes including metering errors, leakage, public usage such as firefighting, and theft [2]. Defective water meters do impair water supply management, but they also affect the financial sustainability of utilities, justifying the need for a detailed analysis of water consumption data by meter [3]. The apparent loss has a significant role in water loss because these represent 30–40% of total leakage [4].

J. Machado et al. (Eds.): icieng 2021, LNME, pp. 234–245, 2022.
https://doi.org/10.1007/978-3-030-78170-5_21

In the literature, we can find various techniques for loss detection in water meters [5], suggesting the creation of indicators based on age, pressure and volume water estimating the meter replacement period [6].

This is a preliminary study of water consumption time-series, in order to implement/construct an algorithm that will be able to determine structural changes associated with water meter malfunction. The long-term purpose is to implement the algorithm in order to help a company, Águas do Norte, in anomaly detection and in the decision-making process of substituting the water meters.

To test de adequacy of the methods proposed in this study we apply them in two case studies: a hotel and a hospital in the time-horizon between 01/12/18 and 30/11/19.

This article is divided into five parts: the first part summarizes the existing literature on losses in water meters and factors that may lead to measurement failures. The second part describes the case study and the dataset, followed by the methodology chapter description. In the fourth, the part we perform an analysis of the results, ending with a brief discussion and conclusion of the results obtained, where remarks on the future investigation are also described.

2 Literature Review

Due to the consumption of uninvoiced water, concessionaires may experience breaks in their financial sustainability. Recently, regarding a European study, Liemberger and Wyatt [7] reported losses of approximately 26.8 million m^3 per day, equivalent to 9.8 billion m^3 in losses in a year. Another result of this same study is that Europe occupies the 6^{th} position in 13^{th} positions with the biggest breaks on the sustainability of concessionaires.

According to the water balance of the International Water Association (IWA), are able to obtain the amount of water without a prescription (Non Revenue Water (NRW)) through the difference between the amount of water that is introduced into the system and the amount of water that is billed by companies to their customers [8]. Water without revenue not only corresponds to actual and apparent losses but also to authorized and uninvoiced consumption. All systems have water losses [9], and it is neither possible nor economically viable to reduce them in their entirety. The reduction, however small, has a positive impact on companies revenues and is also one of the reasons why they seek new methods to avoid them.

Water meters are devices used by utilities to measure and bill the consumption of water that has been made available and consumed by their customers. No water meter is capable of registering 100% of the volume that passes, but there are more accurate meters according to the type of flow that passes through it [10].

Real losses or physical losses correspond to the volume of water that is lost in the network and in the infrastructure of the management entities [9]. These losses are usually caused by leaks in the pipes, valves, and other components of the network [11]. When there are high quantities of water in the network, and they remain for a long period of time there is a greater likelihood of leaks and

broken pipes or plumbings [8]. Real losses exist in greater quantities and are more difficult to detect by concessionaires.

Unbilled authorized consumption is the use of water but the concessionaire has no monetary return from it [8]. These are consumptions that, although sometimes unmeasured, are actually consumed by the suppliers themselves, or by established social commitments such as when used in fires by firefighters [9].

The apparent losses are not physical losses but financial losses [12]. It is extremely difficult to quantify them and despite being in a smaller volume it generates a greater economic impact for companies [13]. The International Water Association (IWA), according to the water balance describes the apparent losses as unauthorised consumption and meter inaccuracies. These are the type of losses that this empirical work relies on.

Measurement errors can be avoided if the management company performs adequate maintenance of water meters and stipulates periods of exchange for them [14], since as the meters get older the water volume that is not registered tends to increase [15]. Criminisi et al. [12] affirm that the wear and the lack of maintenance and calibration of the meters, make the meters obsolete, thus losing their efficiency. The authors also point out that the incorrect choice of the class and application of it can also lead to these inaccuracies. Meter errors not only affect the control of water systems but also the financial management of entities. According to [3] the water losses of the meters caused by inaccuracies can be minimized by analyzing their performance and investigating the causes of these same inaccuracies.

Several approaches may be found in the literature to deal with the problem of water demand prediction (see Benítez et al. [16], for an exhaustive description of the latest works on that thematic). However, only a few dedicate their interest in explicitly exploring and finding changes in water consumption. In that context, we highlight the study of Hester and Larson [17] where they successfully employ breakpoint and decomposition analyses to explore changes in water use for three North Carolina municipalities between 1990 and 2014.

3 Case Study and Dataset Description

The company Águas do Norte, in the year 2018, had losses of around 2,700,000 m^3 of non-invoiced water, corresponding to about 3.67% of the value of the entire system [18]. With such a high volume of losses, the company seeks strategies to decide the best moment to replace the water meters. There is a need to seek and adopt strategies to make decisions to replace meters not only based on the maximum estimated volume, or the useful life, because it does not define which mechanism is working out of the ordinary.

In order to respond to this need, this study aims to develop an algorithm that considers different ways of detecting discrepancies in consumption, integrating later with a Business Intelligence platform where alerts will be given when anomalies in consumption are verified. This tool will give important support to the company in managing and the decision-making of replacing or repairing the water meters.

To demonstrate the adequacy of the methods proposed in this study, we apply them to two company datasets: a hospital water consumption time-series and a hotel water consumption time-series with a time horizon of 01/12/18 until 30/11/19. Both the hotel and the hospital are located in the north of Portugal. The collected data is broken down into periods of 15 in 15 min, hour to hour or daily observation. In the present study, the daily data corresponding a 2 water meters will be analysed.

4 Methodology

The methodology adopted focuses on the application of a set of statistical instruments that aim to validate for the subsequent implementation of an effective algorithm for change point detection on water consumption time-series.

Detecting the structural breaks in the trend of the series, related to a sharp decrease in water consumption, allows us to detect a possible failure in the water meter and alert the company. Note that, a malfunction in the water meter will be reflected in a lower than expected registration of the water consumption. Our analysis goes through three main phases:

Step 1 – Exploratory analysis, in order to detect consumption patterns, variations, and seasonality in the series under study;

Step 2 – Time-series decomposition, to extract the trend component associated with the water meter performance;

Step 3 – Time-series structural change detection, to detect critical moments on the water meter performance.

These steps will allow, in the end, the historical interpretation of data in order to decide if the water meter should be replaced.

4.1 Step 1 - Exploratory Analysis

During the last decades, the L-moments, introduced by Hosking [19], have gained great popularity and have been widely used in many hydrological applications, such as in the analysis and statistical characterization of residential water demand data [20]. In the present study, we examine three statistics based on the first three L-moments λ_i: mean value (=L-mean = λ_1), L-variation, $\tau_2 = \lambda_2/\lambda_1$, and L-skewness, $\tau_3 = \lambda_3/\lambda_2$. The mean value expresses the central tendency of the dataset. L-variation is analogous to the conventional coefficient of variation (ratio of the standard deviation to the mean value) and quantify the variability of the data values. For positive variables such as water flow measures, L-variation takes values in the range $[0, 1]$. L-skewness is a dimensionless measure of the asymmetry of the random variable analogous to the conventional skewness coefficient. L-skewness takes values in the range $[-1, 1]$ where L-skewness equal to zero means that the distribution is symmetric while the distribution is right or left for positive and negative values, respectively. The L-variation and L-skewness coefficients have the advantage to be more robust, to outliers in the data, than

conventional measures, coefficient of variation, and skewness, which means that they suffer less from the effects of sampling variability [19].

4.2 Step 2 - Time Series Decomposition

When analyzing the time series of daily water consumption (in m^3) we need to be able to decompose the series eliminating seasonality and the non-explained variability (also known as remainder). Thus, eliminating measurement error due to noise (that varies depending on the flow rate at which the meter is working) and seasonality effects.

For that purpose, we consider in the proposed algorithm three different ways of decomposing a time series, depending on its nature (additive or multiplicative). The general representation of a time-series decomposition approach is given as:

$$Y_t = f(s_t, h_t, \epsilon_t)$$

where Y_t is the time series value (actual data) at period t, s_t is the seasonal component (or index) at period t, h_t is the trend-cycle component at period t, and ϵ_t is the irregular component at period t. In this particular analysis, water meter performance is directly related to the trend component, h_t. The additive decomposition is, then, given by: $Y_t = s_t + h_t + \epsilon_t$, and the multiplicative decomposition by: $Y_t = s_t \times h_t \times \epsilon_t$.

Seasonal trend decomposition based on Loess (Locally Estimated Scatterplot Smoothing), also known as STL decomposition, has been the preferred procedure for additive seasonal decomposition, in water related time-series decomposition (e.g. Hester and Larson [17] and Benítez et al. [16]). The smoothed Loess based in adjustment the polynomial regression, weighted for an observation time when the weights decrease with a distance nearest neighbor [21].

The package *stats* includes the function *decompose* that implements mechanisms to remove seasonality, trend, and error using moving averages, where they are calculated through the series sequential samples. This function allows the two types of decomposition, additive and multiplicative. To simplify, the function determines the trend through moving average, the seasonality is calculated through the mean in all periods for each time-space, centralizing it. The error is determined by removing the trend and seasonal values from the original series.

As we want a flexible algorithm to accommodate any kind of time-series we opt to implement the three decompositions described above and iterate each subsequent step (2–4) to the three decompositions and compare the final results.

4.3 Step 3 - Breakpoint Detection

As Hester and Larson [17] explains, structural change methods may be used to estimate a long-term trend, identifying periods of statistically significant change. They are, indeed, effective because these methods are specifically intended for exploratory cases where a regression parameter may have changed, but at an unknown point.

To clarify, in an ordered sequence of data $y_{1:n} = (y_1, \ldots, y_n)$ we say that a change point has occurred if there is a moment, $\tau \in (1, \ldots, n-1)$ such that the statistical properties of $(y_1, \ldots, y_\tau)$ and $(y_{\tau+1}, \ldots, y_n)$ are somehow different.

The detection of the change point can be put as a hypotheses test, with the null hypothesis being that there is no change point the alternative hypothesis is that there is a point of change.

There are several packages in R, with functions implemented able to detect breakpoints in a time-series, whether it is structural changes in the mean, variance, or the general distribution.

The *strucchange* package tests structural changes in linear regression models. In this package, we will use the *breakpoint* method that will allow us to calculate the breakpoints in the time series, given the number of breaks. The function will return the number of ideal points [22].

The *changepoint* package provides an algorithm that allows multiple detections of the change points, the package is composed of three methods: the *neighborhood segment*, the *binary segmentation*, and the *PELT (Pruned Exact Linear Time)*. In the study we will test the *binary segmentation* and the *PELT* procedure, testing the changes on the mean and variance of the time-series [23], not use a neighborhood segment because it's necessary attributes a maximum number of change points, limited the size of segmentation [24].

The *Binary Segmentation (BS)* is the most used method, regarding the changepoint analysis. Scott and Knott [25] applied the initial algorithm in search of binary segmentation. As Eckley et al. [26] simplifies, this method initiates by applying a single changepoint test statistic to the entire data if a changepoint is identified the data is split into two at the changepoint location. The procedure is then repeated on the two new data sets, before and after the change. If changepoints are identified in either of the new data sets, they are split further. The process continues until no changepoints are detected in any partition of the data.

The method *PELT* implements an algorithm proposed in [29], that minimizes the expression given by

$$\sum_{i=1}^{m+1} [C(y_{(\tau_{i-1}+1):\tau_i}] + \beta f(m),$$

where C is a cost function for a segment and $\beta f(m)$ is a penalty to guard against overfitting. The minimization is done by using a dynamic programming technique to obtain the optimal segmentation for $m+1$ changepoints reusing the information that was calculated for m changepoints.

The package *Cpm* provides several functions to allow parametric and nonparametric distribution-free change detection in the mean, variance, or general distribution of a time-series, which constitutes an advantage since it relaxes parametric assumptions on the data under analysis.

In analysis will use only the nonparametric methods because the algorithm tests the time series without needing to know its nature. In particular, we made

use of the function *processStream* used to detect multiple change points in a sequence of observations. As the Ross et al. [27] explain, the observations are processed in order, starting with the first, and a decision is made after each observation if changes occur at the point. In this empirical analysis, we address different tests statistics such as: *Mann-Whitney* test statistic, to detect location shifts in a stream with a (possibly unknown) non-Gaussian distribution; *Mood* test statistic, to detect scale shifts in a stream with a (possibly unknown) non-Gaussian distribution; *Lepage* test statistics to detect the location and/or shifts in a stream with a (possibly unknown) non-Gaussian distribution; *Kolmogorov-Smirnov* test statistic, to detect arbitrary changes in a stream with a (possibly unknown) non-Gaussian distribution; and *Cramer-von-Mises* test statistic, to detect arbitrary changes in a stream with a (possibly unknown) non-Gaussian distribution [27].

5 Results

5.1 Exploratory Analysis

Figure 1 displays the mean value, L-variation and L-skewness coefficients by month-to-month, day-to-day, and hour-to-hour for the hotel and hospital. The mean value in the hospital exhibits small monthly changes with slightly higher values between August and October. In contrast, large variation among months is observed in the hotel with a significant increase in the mean value (Fig. 1(a)) during the summer months. Regarding L-variation and L-skewness coefficients (Fig. 1(b), (c)), a small fluctuation in the L-variation coefficient is noted from month-to-month in both companies, while higher fluctuations are noted in the L-skewness coefficient mainly for the hospital during the summer months. Despite the high fluctuation, all L-skewness values are positive (i.e., right-skewed data).

From day-to-day, both companies show different mean values on the weekend days comparing to the mean values on the weekdays (Fig. 1(d)). However, while the hotel shows the higher mean values in the weekend days, the mean values in the weekend days in the hospital are the smallest ones. Regarding the two shape statistics (Fig. 1(e), (f)) a small fluctuation is noted in both. L-variation values remain almost invariant at around 0.4 in the hotel and at around 0.1 in the hospital. Most L-skewness values are concentrated in a narrow interval which ranges from 0.1 to 0.4 for both companies.

From hour-to-hour, while the hotel shows a typical daily consumption pattern, i.e., low flows during night hours, a sharp morning and evening/night peaks and a moderate flow during the day [28], (Fig. 1(g)). Observing the pattern of L-variation and L-skewness coefficients (Fig. 1(h), (i)) the main differentiation is noted between the daily ad night hours. For both shape statistics, a sharp night peak is observed in the hotel. In the hospital, the L-variation values remain almost invariant and a day peak is observed for the L-skewness.

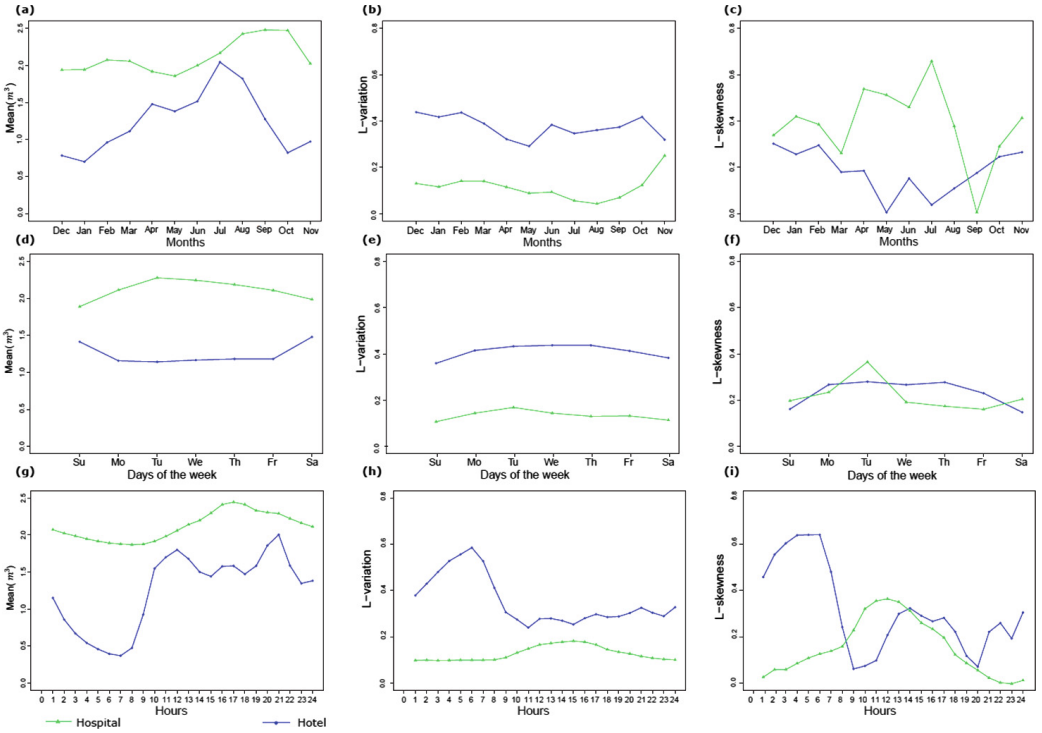

Fig. 1. Variation per month (Figures a, b and c), per day (Figures d, e and f), and per hour (Figures g, h and i) of the Mean, L-variation, and L-skewness values.

5.2 Time Series Decomposition

Additive decomposition is used when the variation in seasonality is constant throughout the series, whereas multiplicative decomposition is used when seasonality increases throughout the series. We note that the additive and multiplicative decomposition of both the hotel and the hospital, the trend, seasonality, and observations are identical.

5.3 Breakpoint Detection

Results of hotel water consumption time-series (Table 1) indicate that the package *Strucchange* has detected 3 change points. The *PELT* algorithm applied to changes on the series mean was the one that detected a higher number of change points (215). However, the same algorithm applied to the detection of changes in variance detected only 4 change points. The *Binary Segmentation* for detection on variance disturbances detected 3 change points, while detecting 5 change points on mean changes of the series. Using the *Cpm* package, the *Lepage* statistic detected 29 change points, the *Kolmogorov-Sminorv* detecting 26, the *Cramer-von-Mises* and *Mann-Whitney* detected 28, and, finally, the *Mood* process detected 17 change points.

For the hospital case (Table 1), the *Strucchange* package detected 4 change points, the *PELT* algorithm, for mean disturbances, detected 235 change points and only 2 change points, for variation changes in the time-series. The *Binary Segmentation* for mean disturbances detected 5 change points while the same method applied for variance disturbances detected only 2 change points. Using the *Cpm* package, the *Lepage* statistic detected 29 change points, the *Kolmogorov-Sminorv* detected 24 change points, the *Cramer-von-Mises* and *Mann-Whitney* detected 26 change points and finally, the *Mood* process detected 16 change points.

Table 1. Number of breakpoints detected and its locations, for hotel Water Consumption Time-Series Trend (decomposed by STL)

	Water meter 1			Water meter 2		
Method	Breakpoints detected	Location of breakpoint	Corresponding dates	Breakpoints detected	Location of breakpoint	Corresponding dates
Strucchange	3	135;236; 299	14/04/19; 24/07/19; 25/09/19	4	107;161; 225;279;	17/03/19; 10/05/19; 13/07/19; 05/09/19
PELT (change in mean)	215	———	———	235	———	———
BinSeg (Change in mean)	5	135;236; 297;318; 338;	14/04/19; 24/07/19; 23/09/19; 14/10/19; 03/11/19;	5	4;108; 162;231; 269;	04/12/18; 18/03/19; 11/05/19; 19/07/19; 26/08/19
PELT (change in variance)	4	133;236; 298;339	12/04/19; 24/07/19; 24/09/19; 04/11/19	2	252; 268	09/08/19; 25/08/19
BinSeg (Change in variance)	3	133;237; 297	12/04/19; 25/07/19; 23/09/19	2	228; 276	16/07/19; 02/09/19
processStream (Lepage statistic)	19	———	———	29	———	———
processStream (Kolmogorov-Smirnov)	26	———	———	24	———	———
processStream (Cramer-von-Mises)	28	———	———	26	———	———
processStream (Mann-Whitney)	28	———	———	26	———	———
processStream (Mood)	17	———	———	16	———	———

Analyzing Fig. 2, we can see that some methods detect the same change points, for both time-series under study, the method implemented by *struccha-nges*, the *binary segmentation* for mean changes, and the *PELT* algorithm for variation changes, detect several proximal structural breaks between them.

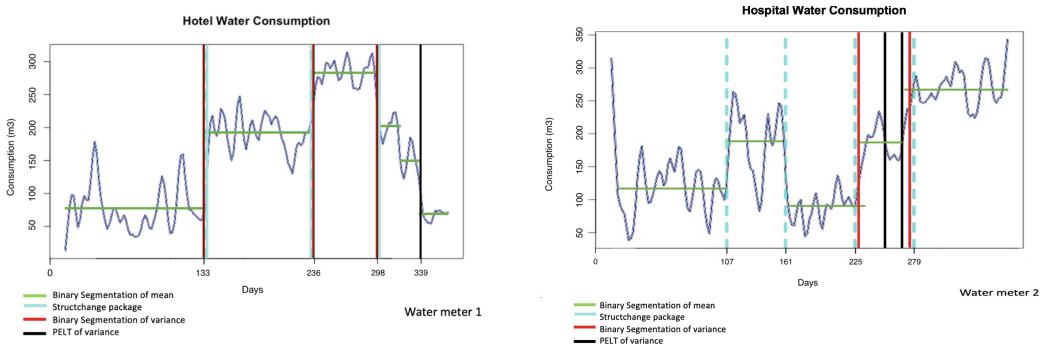

Fig. 2. Comparison of the different methods applied

6 Discussion and Conclusions

This empirical study consists in the application of methods to detect change points, i.e., breakpoints, in the water consumption time-series of a water company located in the north of Portugal. For that, we considered a case study concerning a hotel and a hospital data.

The analysis of seasonality shows a different water consumption pattern in the two understudy companies. While the hotel shows a pattern similar to residential consumption [20,28,30]; with higher consumptions during summer months (Fig. 1(a)), weekend days (Fig. 1(d)), and daily hours (Fig. 1(g)), the hospital shows a different pattern with higher consumptions in the last months of the summer and in the autumn months (Fig. 1(a)), weekdays (Fig. 1(d)) and small variation between daily and night hours consumptions (Fig. 1(g)).

The time-series analysis shows that there is a weekly seasonality and that, implementing the additive and multiplicative decomposition, we encounter no differences in the time-series trend between both procedures.

With the breakpoint detection methods used, we can understand that the *PELT* algorithm for detecting changes on the mean is not very effective since it detects more structural breaks, compromising the algorithm efficiency, augmenting the probability of false positives detection.

For a better prevision power of the algorithm, it will be used a combination of several methods for detection incorporated in software R. The alarm will only be given when the structural break detected has a negative slope and is above an error that the company considers critical. Also, the methods implemented in the *cpm* package may not be effective for our purpose since there was the detection of several points that could generate a great number of false positives and take the concessionaire making bad decisions, what we want to avoid.

Although the results obtained in the study are promising and allowed us to detect change points, critical indicators for the substitution of the water meters, it is necessary to continue to deepen and test new methods in order to obtain an improved algorithm to accurately detect the change points, not indicating false positives and lead the concessionaire into taking a bad decision, which we need to prevent.

Further research aims to extend the analysis of time series to hourly periods and, also, to every 15 min. Also, to perform an analysis of the influence of external variables (e.g. temperature, precipitation) on water consumption exploring different machine learning techniques, for example LSTM (Long short-term memory).

Acknowledgement. This work has received funding from FEDER Funds through P2020 program and from National Funds through FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) under the projects UID/GES/04728/2017, UIDB/00013/2020 and UIDP/00013/2020.

References

1. Ncube, M., Taigbenu, A.: Assessment of apparent losses due to meter inaccuracy-a comparative approach. Water SA **45**(2), 174–182 (2019)
2. El-Abbasy, M.S., Mosleh, F., Senouci, A., Zayed, T., Al-Derham, H.: Locating leaks in water mains using noise loggers. J. Infrastruct. Syst. **22**(3), 04016012 (2016)
3. Mutikanga, H.E., Sharma, S.K., Vairavamoorthy, K.: Investigating water meter performance in developing countries: a case study of Kampala. Uganda. Water SA **37**(4), 567–574 (2011)
4. Szilveszter, S., Beltran, R., Fuentes, A.: Performance analysis of the domestic water meter park in water supply network of Ibarra. Ecuador. Urban Water J. **14**(1), 85–96 (2017)
5. Fontanazza, C.M., Freni, G., La. Loggia, G., Notaro, V., Puleo, V.: A composite indicator for water meter replacement in an urban distribution network. Urban Water J. **9**(6), 419–428 (2012)
6. Puleo, V., Fontanazza, C., Notaro, V., De Marchis, M., La Loggia, G., Freni, G.: Definition of water meter substitution plans based on a composite indicator. Procedia Eng. **70**, 1369–1377 (2014)
7. Liemberger, R., Wyatt, A.: Quantifying the global non-revenue water problem. Water Supply **19**(3), 831–837 (2019)
8. AL-Washali, T., Sharma, S., Al-Nozaily, F., Haidera, M., Kennedy, M.: Monitoring nonrevenue water performance in intermittent supply. Water **11**(6), 1220 (2019)
9. Sardinha, J., Serranito, F., Donnelly, A., Marmelo, V., Saraiva, P., Dias, N., Rocha, V.: Controlo ativo de perdas de água. EPAL-Empresa Portuguesa das Águas Livres, Lisboa (2015)
10. Richards, G.L., Johnson, M.C., Barfuss, S.L.: Apparent losses caused by water meter inaccuracies at ultralow flows. J. Am. Water Works Ass. **102**(5), 123–132 (2010)
11. Arregui, F.J., Gavara, F.J., Soriano, J., Pastor-Jabaloyes, L.: Performance analysis of ageing single-jet water meters for measuring residential water consumption. Water **10**(5), 612 (2018)
12. Criminisi, A., Fontanazza, C.M., Freni, G., Loggia, G.L.: Evaluation of the apparent losses caused by water meter under-registration in intermittent water supply. Water Sci. Technol. **60**(9), 2373–2382 (2009)
13. Arregui, F., Soriano, J., Cabrera, E., Jr., Cobacho, R.: Nine steps towards a better water meter management. Water Sci. Technol. **65**(7), 1273–1280 (2012)
14. Arregui, F., Cabrera, E., Cobacho, R., Garcia-Serra, J.: Reducing apparent losses caused by meters inaccuracies. Water Pract. Technol. **1**(4) (2006)

15. Arregui, F., Cobacho, R., Cabrera Jr., E., Espert, V.: Graphical method to calculate the optimum replacement period for water meters. J. Water Resour. Plan. Manag. **137**(1), 143–146 (2011)
16. Benítez, R., Ortiz-Caraballo, C., Preciado, J.C., Conejero, J.M., Sánchez Figueroa, F., Rubio-Largo, A.: A short-term data based water consumption prediction approach. Energies **12**(12), 2359 (2019)
17. Hester, C.M., Larson, K.L.: Time-series analysis of water demands in three North Carolina cities. J. Water Resour. Plan. Manag. **142**(8), 05016005 (2016)
18. Silva, D.V., Sampaio, M.J., Milagres, C., Alves, V., Ferreira, F.: Flow4Link - the flow in the hand. In: 18th International Flow Measurement Conference (2019)
19. Hosking, J.R.: L-moments: analysis and estimation of distributions using linear combinations of order statistics. J. Roy. Stat. Soc. Ser. B (Methodol.) **52**(1), 105–124 (1990)
20. Kossieris, P., Makropoulos, C.: Exploring the statistical and distributional properties of residential water demand at fine time scales. Water **10**(10), 1481 (2018)
21. Dagum, E.B., Luati, A.: Global and local statistical properties of fixed-length nonparametric smoothers. Stat. Methods Appl. **11**(3), 313–333 (2002)
22. Zeileis, A., Leisch, F., Hornik, K., Kleiber, C., Hansen, B., Merkle, E.C., Zeileis, M.A.: Package 'strucchange'. R package version, 1–5 (2015)
23. Killick, R., Eckley, I.: Changepoint: an R package for changepoint analysis. J. Stat. Softw. **58**(3), 1–19 (2014)
24. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. **107**(500), 1590–1598 (2012)
25. Scott, A.J., Knott, M.: A cluster analysis method for grouping means in the analysis of variance. Biometrics **30**, 507–512 (1974)
26. Eckley, I.A., Fearnhead, P., Killick, R.: Analysis of changepoint models. Bayesian Time Series Models **205–224** (2011)
27. Ross, G.J.: Parametric and nonparametric sequential change detection in R: the CPM package. J. Stat. Softw. **66**(3), 1–20 (2015)
28. Walski, T.M., Chase, D.V., Savic, D.A., Grayman, W., Beckwith, S., Koelle, E.: Advanced water distribution modeling and management (2003)
29. Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., Tsai, T.T.: An algorithm for optimal partitioning of data on an interval. IEEE Signal Process. Lett. **12**(2), 105–108 (2005)
30. Memon, F.A., Butler, D.: Water consumption trends and demand forecasting techniques. Water Demand Manag. **2006**, 1–26 (2006)

Chapter 3

Synthetic dataset to study breaks in the consumer's water consumption patterns

Synthetic dataset to study breaks in the consumer's water consumption patterns

MARTA C SANTOS[1]
CIICESI, ESTG, Politécnico do Porto, Rua do Curral, Casa do Curral, Margaride,4610-156 Felgueiras, Portugal

ANA I BORGES[2]
CIICESI, ESTG, Politécnico do Porto, Rua do Curral, Casa do Curral, Margaride,4610-156 Felgueiras, Portugal

DAVIDE R CARNEIRO[3]
CIICESI, ESTG, Politécnico do Porto, Rua do Curral, Casa do Curral, Margaride,4610-156 Felgueiras, Portugal

FLORA J FERREIRA[4]
Center of Mathematics, University of Minho, Campus de Azurém,4800-058 Guimarães, Portugal

Breaks in water consumption records can represent apparent losses which are generally associated with the volumes of water that are consumed but not billed. The detection of these losses at the appropriate time can have a significant economic impact on the water company's revenues. However, the real datasets available to test and evaluate the current methods on the detection of breaks are not always large enough or do not present abnormal water consumption patterns. This study proposes an approach to generate synthetic data of water consumption with structural breaks which follows the statistical proprieties of real datasets from a hotel and a hospital. The parameters of the best-fit probability distributions (gamma, Weibull, log-Normal, log-logistic, and exponential) to real water consumption data are used to generate the new datasets. Two decreasing breaks on the mean were inserted in each new dataset associated with one selected probability distribution for each study case with a time horizon of 914 days. Three different change point detection methods provided by the R packages *strucchange* and *changepoint* were evaluated making use of these new datasets. Based on Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) performance indices, a higher performance has been observed for the breakpoint method provided by the package *strucchange*.

CCS CONCEPTS • Computing methodologies→Modeling and simulation.

Additional Keywords and Phrases: Water management, Anomaly detection, Structural breaks, Probability Distribution, Water Consumption Simulation

[1] 8150180@estg.ipp.pt

[2] aib@estg.ipp.pt

[3] dcarneiro@estg.ipp.pt

[4] fjferreira@math.uminho.pt

ACM Reference Format:

Marta Santos, Ana Borges, Davide Carneiro, and Flora Ferreira. 2021. Synthetic dataset to study breaks in the consumer's water consumption patterns. 1, 1 (Jun 2021), 11 pages.

1 INTRODUCTION

Water loss is one of the factors that strongly affect a concessionaire's financial sustainability. The early detection of any anomaly in water consumption is very valuable for both economic and environmental sustainability [1]. Particularly, breaks in water consumption records can represent apparent losses, which correspond to theft and illicit consumption. These losses may be associated with water theft by its consumers, billing anomalies, data handling errors, or metering errors [2].

To minimize this type of loss, new solutions to assist the companies in detecting anomalies in water consumption are sorely needed. Given that the water consumption measurement can be seen as a time series, change point methods are suitable for point anomaly detection [3]. In the water consumption context, few studies adopted this methodology such as [4] and [5]. However, studies that assess the performance of such methods in the detection of point anomalies in consumer's water consumption are scarce [6]. The real datasets from water consumption are not always large enough, or abnormal water consumption patterns are not identified. In this context, the production of data that mimics the proprieties of a real dataset (synthetic data [7]) is essential to evaluate, validate and/or test the existent or new methods on the detection of anomalies. Furthermore, using synthetic data for evaluation and testing gives the advantage to tailor the data properties to meet various conditions that are not available in the real datasets [7].

In particular, when applying an anomaly detection method, information about the anomalies is needed to evaluate the rate of success. While in a real dataset the anomalies are not easily identified, in a synthetic dataset anomalies can be inserted. Recently, Kofinas et al. [8] proposed a synthetic data generation technique for household water consumption data, in which the flowrate values were obtained with the use of the estimated probability distributions. The suggested methodology was revealed to successfully reproduce data that mimic the source household's consumption patterns. However, the generated datasets do not present abnormal consumption patterns.

In this article, we present an approach to generate synthetic data with representative points of abnormal consumption from real water consumption data using as study cases two different consumers, a hotel and a hospital. The consumption values are obtained with the use of a probability distribution. Before the generation of consumption data, gamma, Weibull, log-Normal, log-logistic, and exponential distributions are examined in order to identify the one that best describes the real consumer's water consumption. These five probability distributions were selected based on previous studies [8, 9, 10, 11, 12] in which the probabilistic models were used to model hydrological variables with similar characteristics.

A change point corresponds to an abrupt variation in the behavior of the data over time [13], to simulate abrupt changes in the water meters, two decreasing structural breaks were inserted in each time series of water consumption generated. These synthetic datasets are intended to evaluate and test different change point detection methods namely: (i) the breakpoint method incorporated in the R software [14] package *strucchange* [15]; (ii) the binary segmentation, and (iii) the PELT (Pruned Exact Linear Time) method implemented on the *changepoint* package [16].

The change point detection methods have already been implemented in several studies in the hydrological context. For example, Hester and Larson [5] made use of the breakpoint method implemented in the *strucchange* package to explore changes in water use in three cities in North Carolina. Recently, Shao et al. [17] analyzed changes in the monthly flood-related variables via change point detection methods implemented in the *changepoint* package. The inclusion of the PELT method is due to the fact that it presented good results detecting change point with great precision [18]. In a study previously performed by us [19], these three methods were also the ones that detected an acceptable number of change points on water consumption. Thus, in this study, we intend to validate these methods and determine which performs better on detecting the simulated breaks. The rest of this paper is organized as follows: in Section 2 the two real datasets are firstly described. Next, the methodology is described in a step-by-step format. In Section 3 the results of this study are presented and, finally, in Section 4 the main conclusions, limitations, and future work to be developed are presented.

2 MATERIALS AND METHODS

2.1 Data description

The data used in this study was collected from real water consumptions registered during a period of 12 months - from December 1st, 2018 to November 30th, 2019, from a hotel and a hospital located in the north of Portugal. These two datasets are a daily time series of water consumption demand, i.e., each time series value is the total amount of water consumed in a day (24 hours). A statistical study including a seasonal variation analysis of these two datasets can be found in [19].

2.2 Method description

The methodology proposed in this work was developed over three different phases: the investigation of the water consumption values' distribution, the generation of the synthetic data, and the evaluation of the change point methods. It is described throughout the following steps of the three phases:
Steps 1 to 3 correspond to the distribution fitting (phase 1):

1st Step: Estimate the parameters' values of the probability distributions (Table 1). The *fitdist* function from the *fitdistrplus* package in R [20] was used to estimate the parameters of each probability distribution. This function uses the maximum likelihood estimation method by default, where the parameters θ, maximizing the probability definition to find the most probable values.

Table 1: Probabilistic distributions

Distribution	Probability distribution
Weibull distribution	$f_w = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{(k-1)} exp\left(-\left(\frac{x}{\lambda}\right)^{k}\right)$ Where λ is a scale parameter and k is a parameter of shape.
Gamma distribution	$f_G(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\left(x^{\alpha-1}e^{-\beta x}\right)$ Where α is a parameter of shape and β is a parameter of rate.
Log-logistic distribution	$f_{LL}(x) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{(\beta-1)}}{(1+\left(\frac{x}{\alpha}\right)^{\beta})^2}$ Where α is a parameter of scale and β is a parameter of shape.

Distribution	Probability distribution
Log-normal distribution	$f_{LN}(x) = \frac{1}{x\sigma\sqrt{2\pi}} exp\left(-\frac{(lnx-\mu)^2}{2\sigma^2}\right)$
	Where μ is a parameter of mean log and σ is a parameter of standard deviation log.
Exponential distribution	$f_E = \lambda e^{-\lambda x}$
	Where λ is a parameter of rate.

2nd Step: Analyze the probability plots (Histogram, Q-Q plot, and P-P plot) to visually select the probability distributions that best fit a given data.

3rd Step: Analyze the quality of the fit using Kolmogorov-Smirnov, Cramer-von Mises, and Anderson-Darling statistics, and two fit criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The fit statistics are used (choosing the lowest value) to determine the goodness of fit [21]. Anderson Darling's statistic is the most used since it places the same importance on the entire distribution (either in the tails or in the center), where as the Cramer-Von-Moses and Kolmogorov-Sminorv statistics do not take into account the complexity of the model (number of different parameters). In this case, we are comparing distributions with the same number of parameters, so any of the fit statistics can be used to choose the best distribution. Additional, AIC and BIC measures were also calculated to assess model fit.

The following two steps concern the generation of synthetic data (phase 2):

4th Step: Generate time series with random values for each selected probability distribution.

5th Step: Insert decreasing structural breaks in the time series obtained in the previous step. Breaks on the average of the distribution with magnitudes of 5%, 10%, 25%, and 50% have been introduced to simulate anomaly points of water consumption. For each variation of % decrease in the average of the distributions, intervals were simulated every 10 months during 914 days.

The final two steps concern the application and evaluation of change-point methods (phase 3):

6th Step: Application of break detection methods. After a preliminary analysis where several methods of detecting change points were tested, not detailed in this study due to space constraints, we selected the three most consistent methods to evaluate their performance. The methods selected are the breakpoint methods incorporated in the package *strucchange* [15], the binary segmentation, and the PELT (Pruned Exact Linear Time) method of the *changepoint* package [16] of the software R [14]. The breakpoint method allows calculating interruptions in time series, where through the number of breaks it is possible to obtain the number of ideal points [15]. The binary segmentation method starts by applying a single point of change to the dataset, if any point is detected the dataset is divided in two, this process is repeated until no point of change is detected in the division of the datasets [22] and [23]. The PELT method detects points of change by minimizing a cost function over possible locations of change points [24].

7th Step: Calculation of performance indices to evaluate the performance of each method described in the previous step to detect the breaks inserted on the time series (5th Step). To understand which method detects the exact number of change point simulated, we propose a performance index that relates the number of detected breakpoints (n_d) with the number of simulated breakpoints (n_s), as follows:

$$\frac{n_s}{n_d} \times 100. \quad (1)$$

Values of this indicator above 100% mean that the method detected fewer change points than the simulated ones. On the other hand, values under 100% indicate that the method detected more points than the simulated change points (i.e., false alarms).

For the methods with a performance of 100%, i.e., where the correct number of change points simulated were detected, we calculated the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). These are criteria widely used in studies to evaluate models [25]. In this study, these indicators will also be applied to measure the performance of the methods in detecting change points, to quantify the distance between the estimated change point and the simulated break:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}, \quad (2)$$

and

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right| \quad (3)$$

where $n\ (= n_d = n_s)$ is the number of points detected, y_j is a value of breakpoint detected and $\hat{y}_j$ is a value of breakpoint estimated. These indicators are used to understand if the points detected by the methods are the same, or close to the simulated points.

3 RESULTS

3.1 Probabilistic Models (phase 1)

The parameters determined for each distribution can be observed in Table 2.

Table 2: Parameters estimated with a package *fitdistrplus*

Statistic	Weibull		Gamma		Log-logistic		Log-normal		Exponential
	k	λ	α	β	β	α	μ	σ	μ
Hotel	2,296	33,665	4,663	0,157	3,563	27,246	3,282	0,485	0,034
Hospital	3,548	54,923	26,034	0,513	9,108	49,814	3,907	0,193	0,197

Graphically (Figures 1a-c), we can observe that although all selected distributions fit the real data of the hotel well, the Weibull and gamma distributions show a higher agreement of the center quantiles. Furthermore, the gamma distribution seems to be the one that adjusts better to the data in the tails (Figure 1b).

Regarding the hospital data, the selected distributions seem to not adjust so well as observed in the hotel data (Figures 1d-f). The gamma and log-logistic distributions are the distributions that show better adjustment of the theoretical quantiles of the distributions (Figure 1f) and better fit of the tails (Figure 1e).

Analyzing the goodness-of-fit statistics (Table 3) and AIC and BIC values (Table 4) corresponding to hotel data, the gamma distribution is the one that shows lower statistics values. Regarding the hospital data, while the gamma distribution is the one that shows lower fit statistics values (Table 3), the log-logistic distribution is the distribution with lower AIC and BIC values (Table 4).

In conclusion, the gamma distribution showed the best fit for the data standard of the hotel, while for hospital data there is no unanimity between the choice of gamma and log-logistic distributions. Thus, we selected the

gamma distribution for generating the synthetic datasets for both cases and also the log-logistic distribution in the hospital case.

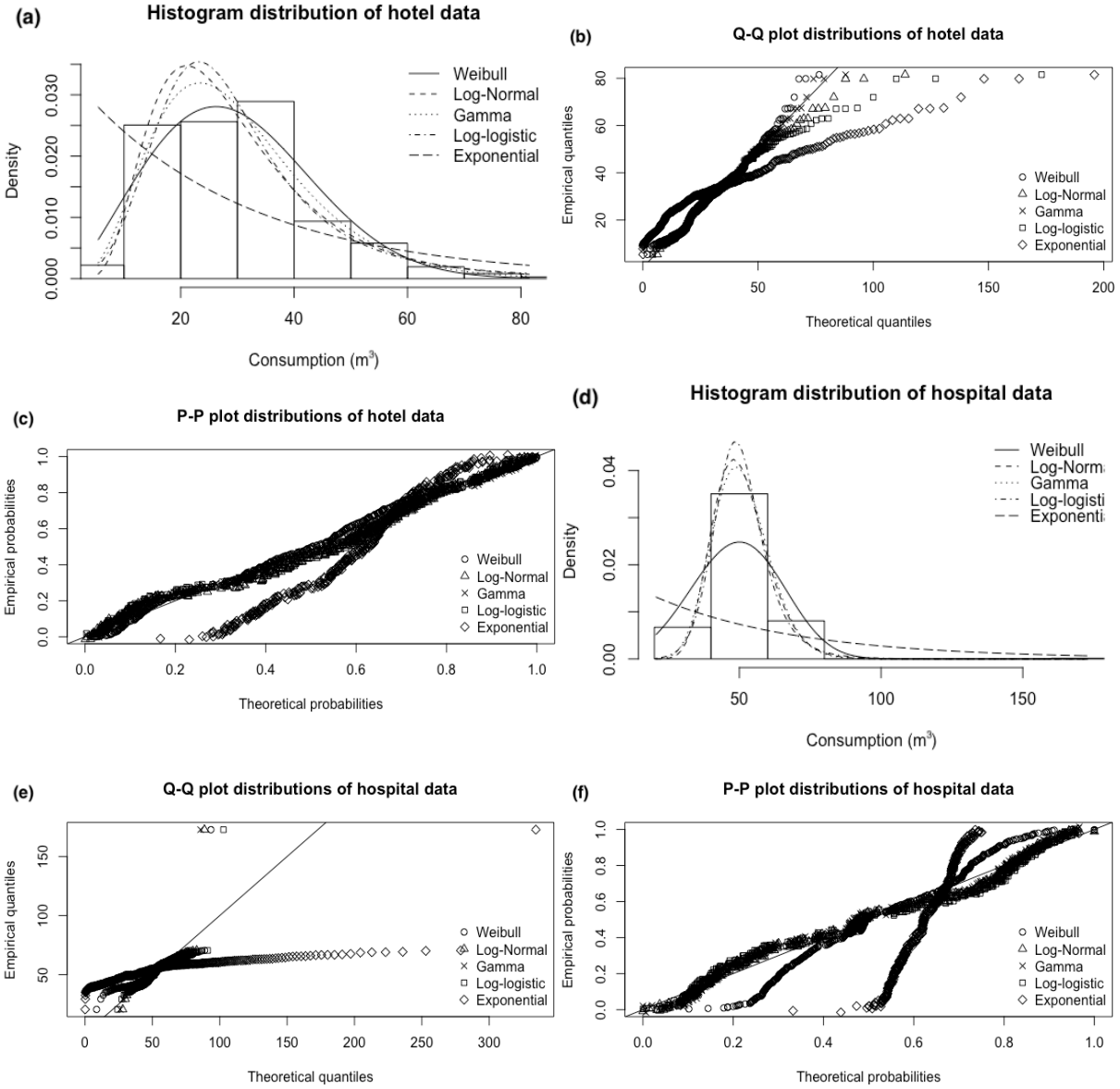

Figure 1: **(a)** Histogram distribution of hotel dataset. **(b)** Q-Q Plot distributions of hotel dataset. **(c)** P-P Plot distributions of hotel dataset. **(d)** Histogram distribution of hospital dataset. **(e)** Q-Q Plot distributions of hospital dataset. **(f)** P-P Plot distributions of hospital dataset.

Table 3: Goodness-of-fit statists

Statistic	Weibull	Gamma	Log-logistic	Log-normal	Exponential
Kolmogorov-Smirnorv statistic(hotel)	0.057	0.057	0.073	0.075	0.276
Cramer-von Mises statistic (hotel)	0.211	0.210	0.389	0.516	9.521
Anderson-Darling statistic (hotel)	1.735	1.357	2.813	2.791	49.698
Kolmogorov-Smirnov statistic (hospital)	0.204	0.061	0.087	0.073	0.494
Cramer-von Mises statistic (hospital)	3.087	0.419	0.704	0.460	23.903
Anderson-Darling statistic (hospital)	Inf	3.349	4.692	3.374	112.479

Table 4: Goodness-of-fit criteria

Statistic	Weibull	Gamma	Log-logistic	Log-normal	Exponential
AIC (hotel data)	2900.329	2883.640	2904.566	2891.688	3191.039
BIC (hotel data)	2908.110	2891.428	2912.354	2899.476	3194.934
AIC (hospital data)	2917.303	2707.229	2685.109	2690.134	3598.564
BIC (hospital data)	2925.103	2715.029	2692.909	2697.934	3602.464

3.2 Generation of synthetic data (phase 2)

A decrease of percentage (%) in the magnitude of the mean of the probability distribution was applied every 10 months in each time series to simulate a break of consumption. In this study, $p = 5, p = 10, p = 25$, and $p = 50$ (other p values can be assumed) were considered to validate which of the methods used in a previously prepared study presents greater sensitivity/performance and for what percentage of breaks. Each dataset has a time horizon of 914 days and 100 datasets were simulated for each type of break, consumption, and distribution and a total of 600 datasets were generated. The values used for each type of decrease were obtained from the decrease in the percentage type of value obtained previously described in Table 2. Examples of generated datasets are available online on [26].

The graphs in Figure 2 are an example of an output generated in the analysis of the detection of points. In this case, it corresponds to one of the datasets generated (from the hotel data) by the gamma distribution with breaks of 50%.

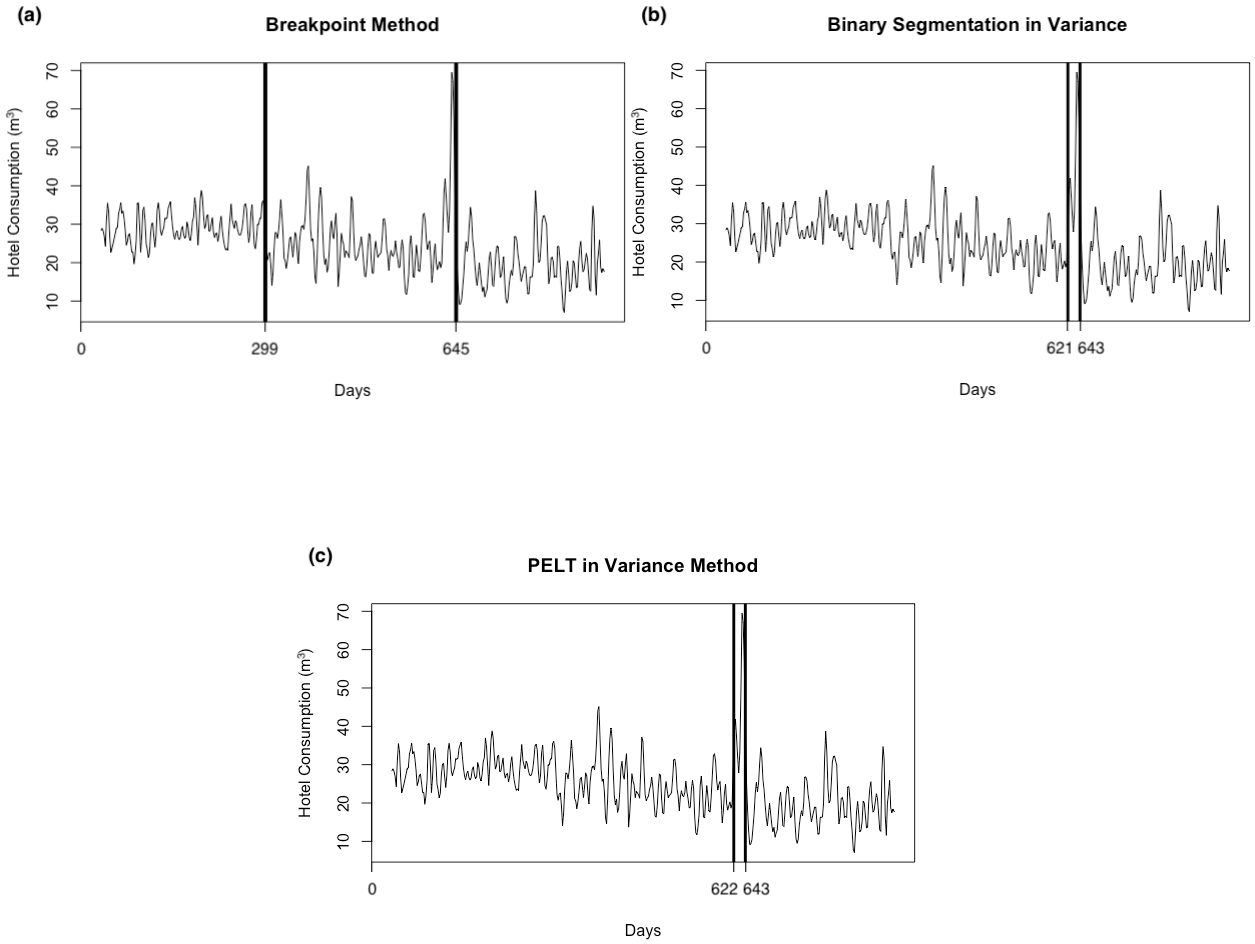

Figure 2: **(a)** Breakpoint Method the *strucchange* package. **(b)** Binary Segmentation in Variance Method the *changepoint* package. **(c)** PELT in Variance Method the *changepoint* package. The vertical lines correspond to the change points found by each method.

In the simulated datasets, breaks were introduced after 304 and 609 days. As can be seen in Figure 2 (a) the breakpoint method of the package *strucchange* detected the 299 and 645 days (represented by the vertical line), the binary segmentation in the variance also detected only two points of change, corresponding to the 621 days and 643, the PELT method also detected two drops in consumption that refer to days 622 and 643. As can be seen, in this example no method accurately detected the simulated points, but the breakpoint method was the one that detected the first closest point of the simulated.

3.3 Evaluation of the change-point methods (phase 3)

For the methods that detected the same number of anomalies as those that were simulated (that is, 2 breaks), the RMSE and MAE were used to assess the performance of the methods as can be seen in Table 5.

Table 5: RMSE and MAE results for distributions

Distribution	Consumption	breaks	RMSE breakpoint	MAE breakpoint	RMSE BSVar[5]	MAE BSVar	RMSE PELTVar[6]	RMSE PELTVar
Gamma	Hospital	5%	201,238	175,065	308,100	294,750	253,629	222,583
		10%	160,759	140,870	266,984	233,375	303,573	275,238
		25%	157,220	141,891	243,555	221,217	269,519	243,818
		50%	174,864	162,344	228,246	200,870	229,875	208,071
	Hotel	5%	158,601	137,829	277,003	256,176	258,649	229,868
		10%	172,345	152,341	267,008	247,020	294,175	269,341
		25%	111,450	92,250	240,031	214,281	248,021	224,750
		50%	108,596	97,857	231,908	205,703	184,717	148,444
Log-Logistic	Hospital	5%	92,244	78,833	270,313	250,441	214,175	187,964
		10%	42,497	32,796	77,228	63,875	81,925	64,094
		25%	17,541	12,796	90,483	64,900	88,321	63,304
		50%	29,461	21,054	26,743	19,378	31,121	22,246

For the two distributions that obtained the best fit to the data, it can be verified with the analysis of Table 5 that the breakpoint method of the package *strucchange* obtained a better performance (analyzing the RMSE and the lowest MAE), either for the hotel data using gamma distribution or for hospital data using the gamma and log-logistic distribution. The method also demonstrated better performance for almost all types of breaks that were inserted to simulate the decreases in water consumption.

4 CONCLUSIONS AND FUTURE RESEARCH

In this study, a method for generating water consumption data with punctual anomalies (breaks) is introduced. The consumption values are modeled with the use of the estimated probabilistic distributions and two breaks were inserted in the generated time series. Different synthetic datasets were generated with different magnitude of breaks. Three different change point methods (breakpoint, segmentation binary in variance and PELT in variance) were applied and evaluated, where the breakpoint method of the package *strucchange* [15] shows a better performance in the detection of the introduced punctual anomalies.

The synthetic datasets provided can be used to test and evaluate other or new anomaly detection methods to detect breaks in the consumer's water consumption patterns. Furthermore, the proposed approach can be used to generate synthetic datasets in other areas to test detection methods of breaks.

The suggested approach constitutes a potential basis for building up a tool, which will support the water company to detect anomalies in water consumption. The synthetic data can be assumed as the expected

[5] Binary Segmentation in Variance

[6] Pruned Exact Linear Time in Variance

consumer's water consumption. However, some improvements are needed. Information about seasonality (e.g. consumption pattern throughout the week) should be incorporated in the generation of the expected water consumption. On the other hand, the algorithms here evaluate show higher accuracy in the detection of breaks with higher slopes, and more research is need when the slope of breaks is smaller. Other anomaly detection methods should thus be tested and evaluated. Moreover, it is also necessary a deeper study on what the level of the consumption break for each type of consumer can be considered as an anomaly.

ACKNOWLEDGMENTS

This work was supported by FCT - Fundação para a Ciência e a Tecnologia, through projects UIDB/04728/2020 and UIDB/00013/2020. The authors thank the Portuguese water company Águas do Norte for providing the real data used in this study.

REFERENCES

[1] Martin Oberascher, Michael Möderl, and Robert Sitzenfrei. 2020. Water Loss Management in Small Municipalities: The Situation in Tyrol. Water 12, 12 (2020), 3446.
[2] FJ Arregui, J Soriano, E Cabrera Jr, and R Cobacho. 2012. Nine steps towards a better water meter management. Water Science and Technology 65, 7 (2012), 1273–1280. https://doi.org/10.2166/wst.2012.009
[3] Samaneh Aminikhanghahi and Diane J Cook. 2017. A survey of methods for time series change point detection. Knowledge and information systems 51, 2 (2017), 339–367.
[4] Kimberly J Quesnel and Newsha K Ajami. 2017. Changes in water consumption linked to heavy news media coverage of extreme climatic events.Science advances 3, 10 (2017), e1700784.
[5] Cyrus M Hester and Kelli L Larson. 2016. Time-series analysis of water demands in three North Carolina cities. Journal of Water Resources Planning and Management 142, 8 (2016), 05016005. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000659
[6] José Carlos Carrasco-Jiménez, Filippo Baldaro, and Fernando Cucchietti. 2020. Detection of Anomalous Patterns in Water Consumption: An Overview of Approaches. In Proceedings of SAI Intelligent Systems Conference. Springer, 19–33.
[7] Emilie Lundin Barse, Hakan Kvarnstrom, and Erland Jonsson. 2003. Synthesizing test data for fraud detection systems. In 19th Annual Computer Security Applications Conference, 2003. Proceedings. IEEE, 384–394.
[8] Dimitris T Kofinas, Alexandra Spyropoulou, and Chrysi S Laspidou. 2018. A methodology for synthetic household water consumption data generation. Environmental modelling & software 100 (2018), 48–66.
[9] Panagiotis Kossieris and Christos Makropoulos. 2018. Exploring the Statistical and Distributional Properties of Residential Water Demand at Fine Time Scales. Water 10, 10 (2018). https://doi.org/10.3390/w10101481
[10] Seevali Surendran and Kiran Tota-Maharaj. 2018. Effectiveness of log-logistic distribution to model water-consumption data. Journal of Water Supply: Research and Technology—AQUA 67, 4 (may 2018), 375–383. https://doi.org/10.2166/aqua.2018.175
[11] Rudy Gargano, Carla Tricarico, Francesco Granata, Simone Santopietro, and Giovanni De Marinis. 2017. Probabilistic models for the peak residential water demand. Water 9, 6 (2017), 417.
[12] Kiran Tota-Maharaj and Seevali Surendran. 2020. 3-Parameter Log-Logistic Distribution Modelling and Scenario Development to Evaluate the United Kingdom's Water Demand. Institute of Water Journal 4 (2020).
[13] Shilpy Sharma, David A Swayne, and Charlie Obimbo. 2016. Trend analysis and change point techniques: a survey. Energy, Ecology and Environment 1, 3 (2016), 123–130.
[14] R Core Team et al. 2013. R: A language and environment for statistical computing. (2013).
[15] Achim Zeileis, Friedrich Leisch, Kurt Hornik, Christian Kleiber, Bruce Hansen, Edgar C Merkle, and Maintainer Achim Zeileis. 2015. Package 'strucchange'. R package version (2015), 1–5.
[16] Rebecca Killick and Idris Eckley. 2014. changepoint: An R package for changepoint analysis. Journal of statistical software 58, 3 (2014), 1–19.
[17] Manqing Shao, Gang Zhao, Shih-Chieh Kao, Lan Cuo, Cheryl Rankin, and Huilin Gao. 2020. Quantifying the effects of urbanization on floods in a changing environment to promote water security—A case study of two adjacent basins in Texas. Journal of Hydrology 589 (2020), 125154.
[18] Rebecca Killick, Paul Fearnhead, and Idris A Eckley. 2012. Optimal detection of changepoints with a linear computational cost. J. Amer. Statist. Assoc. 107, 500 (2012), 1590–1598.
[19] Marta Santos, Ana Borges, Davide Carneiro, and Flora Ferreira. 2021. Time Series Analysis for Anomaly Detection of Water Consumption: a case study. In In 6th International Conference on Industrial Engineering. Lecture Notes in Mechanical Engineering. Springer, Cham. [accepted].
[20] Marie Laure Delignette-Muller, Christophe Dutang, Regis Pouillot, Jean-Baptiste Denis, and Maintainer Marie Laure Delignette-Muller. 2020. Package 'fitdistrplus'.
[21] Marie Laure Delignette-Muller and Christophe Dutang. 2012. Fitting parametric univariate distributions to non censored or censored data using the R fitdistrplus package.
[22] Idris A Eckley, Paul Fearnhead, and Rebecca Killick. 2011. Analysis of changepoint models. Bayesian Time Series Models(2011), 205–224. https://doi.org/10.1017/CBO9780511984679.011

[23] Christian Rohrbeck. 2013. Detection of changes in variance using binary segmentation and optimal partitioning.

[24] G Dorcas Wambui, Gichuhi Anthony Waititu, and Anthony Wanjoya. 2015. The power of the pruned exact linear time (PELT) test in multiple changepoint detection. Am. J. Theor. Appl. Stat 4, 6 (2015), 581–586.

[25] T. Chai and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Development 7, 3 (2014), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[26] Marta Santos, Ana Borges, Davide Carneiro, and Flora Ferreira. 2021. Synthetic Datasets of Water Consumptions. Mendeley Data, V2. http://dx.doi.org/10.17632/v4ynw83j6k.2.

Chapter 4

Dashboard Implementation

The dashboard is intended to be a tool that aggregates all information and presents it to the user in a simple and easy to interpret visual way, so that it is easy for the user to interpret and extract information. For the development of this dashboard we used the shiny R software package that allows the creation of web applications in a simple way. A shiny application is divided into the UI where it has all the front-end logic that is made available to the user and the server part where all the back-end logic is implemented for the application to work ([14]).

Figure 4.1: Structure implemented in the dashboard

The structure implemented in the dashboard Figure 4.1 has its main focus 3 phases:

- 1^{st} - Obtaining consumption data: by uploading the CSV file by the user;
- 2^{nd} - Data treatment: through the application of change point analysis methods;
- 3^{rd} - Results visualization: where the results are made available through the dashboard to assist in the decision-making process.

4.1 Magnitude of change

A function was created to determine the magnitude of the change point, by calculating the slope before and after the change point. This value is displayed on the dashboard and the company should define a threshold regarding this value above which the decision of the water

meter substitution should be made.

4.1.1 Slope before change point

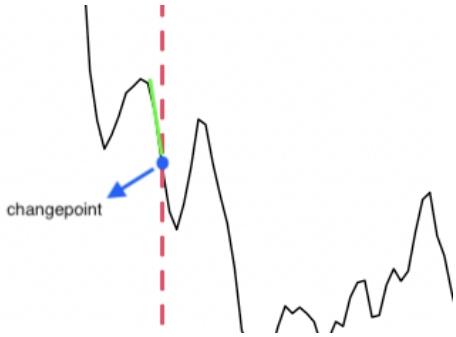

Figure 4.2: Slope before change point

When the change point is found (represented in blue in Figure 4.2), the slope determined is between the consumption of previous days and the consumption at the change point (as represented in green in Figure 4.2). Through the analysis of data sets provided by the company, it was detected that they had weekly seasonality. When the change point is detected, it will be compared with the previous 7^{th} day due to the detected seasonality. As consumption varies over time, nothing indicates that consumption on the 7^{th} day before is higher than consumption at the change point, and it is not possible to obtain the slope of the loss. If the consumption on the seventh day prior to the detected one is less than the consumption of the detected point, it is incremented one more day and this comparison is made again (the code developed is available in appendix A.1).
When the consumption of the n previous days (that day is shown on the dashboard) is higher than the consumption at the detected change point, it is applied as follows:

$$Slope_{BeforeChangepoint} = \frac{ConsumptionBreakpoint - ConsumptionDaysBefore}{valueBreakpoint - (valueBreakpoint - daysMin)} \tag{4.1}$$

where the ConsumptionBreakpoint is the consumption at the detected change point, ConsumptionDaysBefore is the consumption determined days before, as explained above, valueBreakpoint is the value of the breakpoint and (valueBreakpoint - daysMin) is the value obtained by the value of the breakpoint minus the days used in the detection of ConsumptionDaysBefore.

4.1.2 Slope after change point

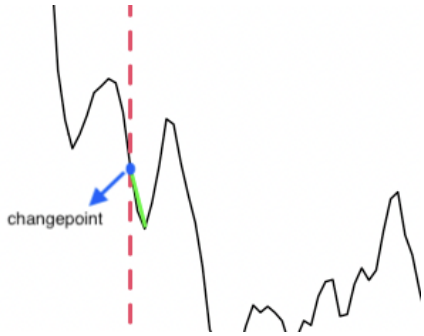

Figure 4.3: Slope after change point

The slope after the change point (Figure 4.3 follows the same routine, the number of base days used is 7 (due to weekly seasonality), but in this case if the consumption 7 days past the consumption of the detected point is higher, it will increase in days until that consumption is smaller, because if we are facing a negative slope (see Figure 4.3). The second point consumption will have to be necessarily smaller than the first consumption (the code developed is available in appendix A.2), in order to be able to determine the slope of the loss after the change point.
The slope expression determined after the change point is demonstrated as:

$$Slope_{AfterChangepoint} = \frac{ConsumptionDaysAfter - ConsumptionBreakpoint}{(valueBreakpoint + daysMax) - valueBreakpoint} \tag{4.2}$$

where the ConsumptionBreakpoint is the consumption at the detected change point, ConsumptionDaysAfter is the consumption determined days after, as explained above, valueBreakpoint is the value of the breakpoint and (valueBreakpoint + daysMax) is the value obtained by the value of the breakpoint plus the days used in the detection of ConsumptionDaysAfter.

4.2 Dashboard Visualization

The demo of the dashboard implemented for the detection of anomalies in the Águas do Norte company can be seen in the Figure 4.4.
It gives the user back an analysis of the consumptions and the main points of change detected. On the left, the user uploads the file, for now the dashboard only accepts csv files, but in the future can be connected to the company's Database, thus avoiding the export of files and facilitating the process. After the upload, a graphic is returned to the user, wherein the consumptions, the breaks detected by the breakpoint method of the structchange package are indicated (this being the method that obtained the best performance in detecting failures as noted in Chapter 2), it also returns the information on the slope value before and after the change point, as mentioned above. It should be noted that the significance of the value of this slope must be analyzed

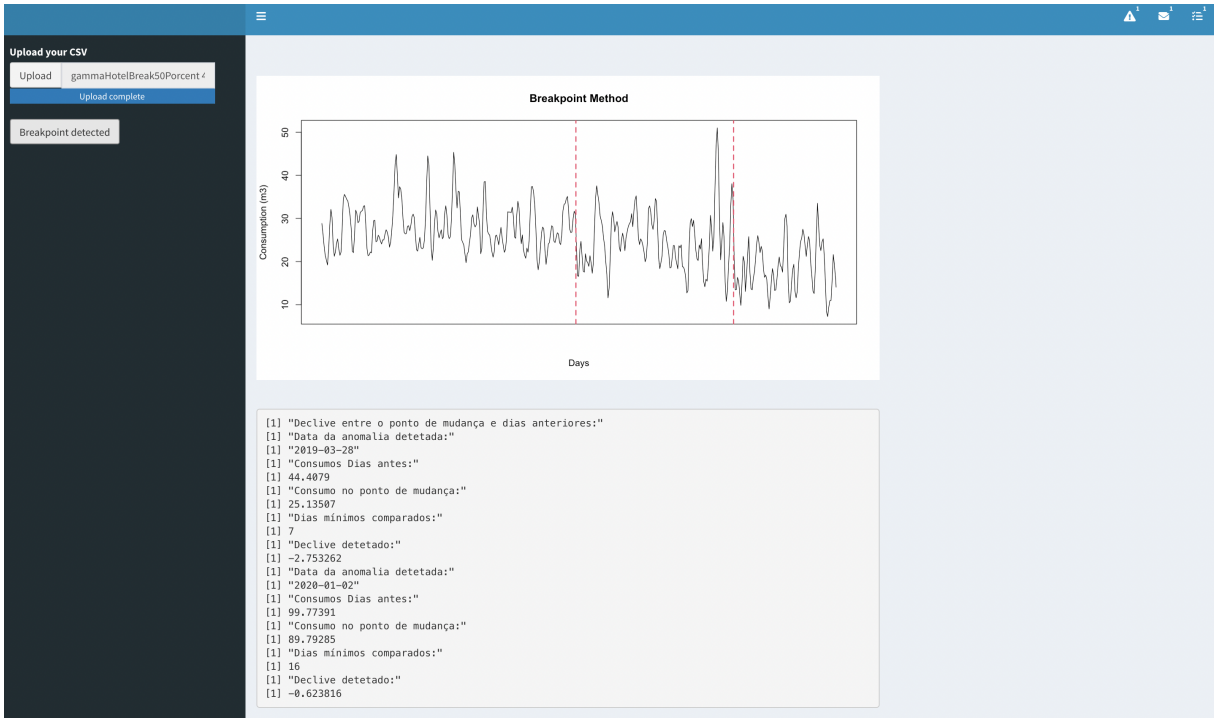

Figure 4.4: The demo of the dashboard implemented

by the company in order to be able to make decisions in favor of its objectives.
The dashboard focuses on water consumption, but it can be adapted to any context of analysis of anomalies that happen over a time series.

Chapter 5

Conclusion

Water losses are one of the most relevant problems affecting the efficiency of water distribution worldwide and it is estimated that in water supply systems they can reach around 50% on a global scale [5]. Distribution companies face new challenges for the sustainable management of water resources [15].
In order to help Águas do Norte to manage its resources in a sustainable way, this project arose, which makes an important contribution to improving the process of detecting anomalies in the water consumption of Águas do Norte. Based on two datasets, methods were used to detect change points in consumption. In order to validate these same methods, 1200 datasets were simulated where breaks with magnitudes of 5%, 10%, 25% and 50% were introduced and it was concluded that the method had a better performance for detecting anomalies in consumption independently of the simulated break was the breakpoint method of the *structchange* package.
To assist the company in the process of detecting and reducing losses, a dashboard was created, still in a demo version. The company will only have to upload the csv file, where the consumptions for the period of time are described and will be made available on the dashboard, the detected change points, the dates they occurred, and a calculating the magnitude of the loss reduction, which has to be evaluated by the company to understand the level of importance the same. With the availability/visualization of the results, it is intended to clarify and simplify the decision process for the company.

5.1 Limitations and future research

With the unusual times we live in due to COVID-19, it has not yet been possible to make the demo version available for the company to use. Once available, they can add new features to the dashboard, and implemented improvements in the feature. With strong concerns about the scarcity of water resources, an add-on for the dashboard may be created in the future, to detect significant increases in consumption in order to combat unnecessary water losses, since water is essential for human survival.
It will be intended to improve the dashboard through comparisons between current consumption and forecasts, through the use of forecast models such as the SARIMA model (among

others). It is also intended to include machine learning models for the detection of anomalies in consumption, which is already being developed.

Bibliography

[1] Roland Liemberger and Alan Wyatt. “Quantifying the global non-revenue water problem”. In: *Water Supply* 19.3 (2019), pp. 831–837.

[2] *Organização das Nações Unidas*. Organização das Nações Unidas, 2021. URL: https://unric.org/pt/ (visited on 2021-05).

[3] Ewa Ociepa, Maciej Mrowiec, and Iwona Deska. “Analysis of water losses and assessment of initiatives aimed at their reduction in selected water supply systems”. In: *Water* 11.5 (2019), p. 1037.

[4] Francisco J Arregui et al. “Performance analysis of ageing single-jet water meters for measuring residential water consumption”. In: *Water* 10.5 (2018), p. 612.

[5] Nelson Pimenta and Paulo Chaves. “Study and design of a retrofitted smart water meter solution with energy harvesting integration”. In: *Discover Internet of Things* 1.1 (2021), pp. 1–15.

[6] *International Water Association*. International Water Association.

[7] Allan O Lambert. “International report: water losses management and techniques”. In: *Water Science and Technology: Water Supply* 2.4 (2002), pp. 1–20.

[8] Iñigo Monedero et al. “An approach to detection of tampering in water meters”. In: *Procedia Computer Science* 60 (2015), pp. 413–421.

[9] FJ Arregui et al. “Graphical method to calculate the optimum replacement period for water meters”. In: *Journal of Water Resources Planning and Management* 137.1 (2010), pp. 143–146.

[10] A Criminisi et al. “Evaluation of the apparent losses caused by water meter under-registration in intermittent water supply”. In: *Water Science and Technology* 60.9 (2009), pp. 2373–2382.

[11] Gregory L Richards, Michael C Johnson, and Steven L Barfuss. “Apparent losses caused by water meter inaccuracies at ultralow flows”. In: *Journal-American Water Works Association* 102.5 (2010), pp. 123–132.

[12] *Águas de Portugal*. Águas do Norte, 2015. URL: https://www.adp.pt/pt//?id=78&emp=6 (visited on 2020-12).

[13] Marta Santos et al. “Time Series Analysis for Anomaly Detection of Water Consumption: A Case Study”. In: *Innovations in Industrial Engineering*. Ed. by José Machado et al. Cham: Springer International Publishing, 2022, pp. 234–245.

[14] Hadley Wickham. *Mastering shiny*. " O'Reilly Media, Inc.", 2021.

[15] Helena M Ramos et al. "Smart water management towards future water sustainable networks". In: *Water* 12.1 (2020), p. 58.

Appendix

Appendix A

A.1 Slope before change point

```
daysMin<-7
for(i in 1:length(detection$breakpoints)){
    ConsumptionDaysBefore <-df$Consumption[detection$breakpoins[i]
    - daysMin]
    consumoBreakpoint<-df$Consumption[detection$breakpoints[i]]
        while(ConsumptionDaysBefore <= consumoBreakpoint){
            daysMin<-daysMin+1
            ConsumptionDaysBefore<-df$Consumption[detection$breakpoints[i]
            - daysMin]
        }
    slope<-(consumoBreakpoint-ConsumptionDaysBefore)/
    (detection$breakpoints[i]-(detection$breakpoints[i]-daysMax))
}
```

A.2 Slope after change point

```
daysMin1<-7
for(i in 1:length(detection$breakpoints)){
    ConsumptionDaysAfter<-df$Consumption[detection$breakpoints[i]
    +daysMin]
    consumoBreakpoint<-df$Consumption[detection$breakpoints[i]]
        while(ConsumptionDaysAfter >= consumoBreakpoint){
            daysMin<<-daysMin+1
            ConsumptionDaysAfter<-df$Consumption[detection$breakpoints[i]
            + daysMin]
        }
    slope<-(ConsumptionDaysAfter-consumoBreakpoint)/
    ((detection$breakpoints[i]+daysMin) - detection$breakpoints[i])
}
```