Water Meters Inaccuracies Registrations: A First Approach of a Portuguese Case Study

M. Filomena Teodoro[1,2], Marina A. P. Andrade[3,4(✉)], Sérgio Fernandes[5], and Nelson Carriço[6]

[1] CINAV, Center of Naval Research, Naval Academy, Portuguese Navy, 2810-001 Almada, Portugal
[2] CEMAT, Center for Computational and Stochastic Mathematics, Instituto Superior Técnico, Lisbon University, 1048-001 Lisboa, Portugal
maria.alves.teodoro@marinha.pt
[3] ISCTE -IUL, Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal
[4] ISTAR -IUL, Instituto Universitário de Lisboa, 1649-026 Lisboa, Portugal
marina.andrade@iscte.pt
[5] ESTSetúbal, Escola Superior de Tecnologia de Setúbal, Instituto Politécnico de Setúbal, Campus do IPS Estefanilha, 2914-508 Setúbal, Portugal
[6] ESTBarreiro, Escola Superior de Tecnologia do Barreiro, Instituto Politécnico de Setúbal, Rua Américo da Silva Marinho, 2839-001 Lavradio, Portugal

Abstract. The work described in this article results from a problem proposed by a water utility company in the framework of ESGI 140*th*, during June 2018. The objective is to evaluate water meters performance using historical data, knowing that, being a mechanical device, the water meters suffer a deterioration with time and use, losing accuracy throughout its cycle of use. We intend to approach a problem capable of identifying anomalies on water consumption pattern. In present work, ARIMA modeling was considered to obtain a predictive model. The results show that in the time series traditional framework revealed significant and adequate in the different estimated models. The in-sample forecast is promising, conducting to adequate measures of performance.

Keywords: Water meter performance · Anomalies identification · ARIMA models

1 Introduction

The study of the behavior of time series data is considered one of the current challenges in data mining [1,2]. A wide number of methods for water demand Forecasting can be found in literature, for example, in [3] we have a good description of

© Springer Nature Switzerland AG 2020
O. Gervasi et al. (Eds.): ICCSA 2020, LNCS 12251, pp. 429–445, 2020.
https://doi.org/10.1007/978-3-030-58808-3_31

such methods which are labeled in five categories. Data Mining refers the extraction of knowledge by analyzing the data from different perspectives and accumulates them to form useful information which could help the decision makers to take appropriate decisions. Due to the unique behavior of time series data, we can find several references about time series data mining in various application domains, namely survey articles [4,5], PhD thesis where detailed techniques and case studies are analyzed [6], private communications [7], which takes into consideration how Artificial Neural Networks may assist in formulating a GLM, chapters in books [8] where the authors define the major tasks considered by the time series data mining community. Besides that, the existing research is still considered not enough. In [1] time series data is considered one of the 10 challenges in data mining. In particular the discovery of an interesting pattern, also called motif discovery, is a non-trivial task which has become one of the most important data mining tasks. In fact motif discovery can be applied to many domains [2].

Under the 140th European Study Group with Industry, Infraquinta submitted the mathematical challenge: they would like to have an algorithm for evaluating water meters performance by using historical data (hourly water consumption).

In Sect. 2 is described a brief review about the enormous quantity of available methods, analyzing their strengths and weaknesses. Such methods can be clsaasified in four categories: multi-agent models, fundamental models, reduced-form models, statistical models and computational intelligence models.

In present work, usual techniques in time series modeling are used. The ARMA, ARIMA, ARIMAX, SARIMAX are considered useful techniques to obtain a predictive model where its predictive power is discussed.

The outline of this article is developed in six sections. In Sect. 2 is presented some background about short term forecast techniques applied to this kind of problems. Section 3 makes a brief summary about the time series modeling approach. Section 4 provides more details about the challenge proposed by Infraquinta and on the provided data. In Sect. 5 is displayed a short summary about exploratory analysis of the data sets provided and continues with the study on the co-variables that may predict the time series breakpoint. The results of our ARIMA approach are presented. Finally in Sect. 6 some conclusions are drawn and suggestions for future work are pointed.

2 Preliminaries

Water utility companies need to control the supply network system so they can detect undesirable occurrences such as damages, leaks or others issues, so the water distribution network can be considered by the users as reliable and adequate. Another issue that contributes largely to this aim is that the water loss shall be reduced, by one hand due the trend of increasing water cost, by another, due the repair of damages that is usually very expensive. In [9,10] is evidenced that supervision of network water supply is mandatory to the improvement of the performance of water supply.

With the aim of a contribution to the improvement of the reliability of water distribution networks, the authors of [11] propose a fault measurement detection

considering a model-based approach. Considering a fuzzy concept, the diagnosis procedure proposed in [11] takes in consideration all available data and knowledge, revealing that it should be considered in water meters faults management.

In [12] is presented a good contribution to the validation and reconstruction of the flow meter data from a complex water system supply - the Barcelona water distribution network. It is processed a signal analysis to validate (detect) and reconstruct the missing and false data from a large dataset of flow meters in the tele-control system of a water distribution network. It considers two time scale in the proposed models: the daily data is considered in an ARIMA time series approach; the 10 min scale models uses a $10-$ min demand pattern using correlation analysis and a fuzzy logic classification.

The study presented in [13] concerns about inefficient water metering together with the practice of low tariffs, problem that contributes negatively to the financial sustainability of utilities. Many water utilities in developing countries face this problem. In this work, the performance of 3 m models is analyzed, also the influence of sub-metering is studied.

An interesting approach is done in [14], where the estimated value of not measured water volume in residential connections by loss of accuracy in hydrometers. The authors present the study developed in a Brazilian city from Bahia region, with the intention of estimating the sub-measurement index (percentage of the volume supplied to the consumer that is not measured) in residential consumers, due to the progressive loss of accuracy of the water meters installed due to the installation time. The work was developed based on direct measurements of the consumption profile of typical residential consumers and the evaluation of the real conditions of the meters installed in the city. The work was developed under the aim of the National Program to Combat Water Waste.

A study to find the important explanatory factors of residential consumption [15], spread in several levels categories, built a data base designed with detailed qualitative and quantitative data taking into consideration smart water metering technology, questionnaire surveys, diaries, and household water stock inventory audits. The water demand forecasting models were built using a wide number of several statistical techniques per each user category: cluster analysis, dummy coding, independent t-test, independent one-way ANOVA, bootstrapped regression models. The authors of [15] concluded that socio-demographic-economical and physical characteristics are most significant factors of water demand and that should be taken into consideration in water supply management.

The water demand management is giving more and more importance to smart metering. In last years, we can see that, for example in Australasian water utilities, where customer service, labor optimization, and operational efficiency are important keywords. In [16] are presented surveys and in-depth interviews that are accomplished of smart metering and intelligent water network, projects implemented is last years that got an important feedback. It is evidenced that digital metering combined with data analytics can be used to increase the utility efficiency and costumer service quality inducing an improvement of costumer satisfaction.

In [17], the authors propose the application of the Poisson generation models where the intensity and duration of residential water demand pulse generation

are considered. The models can use the readings from households using smart metering technologies, aimed to preserve the mean and cumulative trend of water demand. When several case studies are considered, the models estimated the water demand with adequate quality, even considering multiple time aggregation scales (from 1–15 min until 1 day).

The work presented in [18] follows the same idea: the authors used the data acquired from smart metering of a big number of householders and the detailed information per individual consumer. This work describes a mixed methods study with a certain detail. They have modeled the profile behavior of householder water consumers information together with the infrastructure report details, fact that allowed to improve the efficiency of water use and to promote water conservation. This improvement was evidenced by a home water updates detailed feedback provided by a group of selected households.

3 ARIMA Approach

The identification of an ARIMA model to model the data can be considered one of the most critical phases when using ARIMA approach. For a stationary time series[1] the selection of the model to be used is mainly based on the estimated auto-correlations and partial auto-correlations, which we will use to compare with the theoretical quantities and identify a possible model for the data. Some useful references about ARIMA approach can be found in [19–21].

The auto-regressive models with order p AR(p), the moving average with order q MA(q) and their combination, ARMA(p,q) models have their auto-correlation functions (ACF) with a certain specific feature, similarly to a finger print:

The ACF of an autoregressive process with order p is infinite in extent that decays according to a damped exponential/sinusoidal;
The ACF of a moving average with order q process is finite, i.e. presents a cut after the lag q;
The ACP of ARMA process (p,q) is like a mixture of the processes described in previous items, the ACF has infinite that decays according to exponents/damped sinusoidals after the lag $q - p$.

The idea is to identify a pattern that behaves with the same profile that some theoretical model. In particular, the it is useful to identify MA models but it is not so simple to identify other kind of models. As a possible solution, we can compute the partial auto-correlation function (PACF). This function corresponds to the correlation of X_t, X_{t-k+1} removing the effect of the observations $X_{t-1}, X_{t-2}, \ldots, X_{t-k-1}$ and is denoted by ϕ_{kk}. In the case a stationary time series we can use the Yule-Walker equations to compute the PACF. Again, the PACF have a specific profile for each process like a proper finger print:

The PACF of a MA(d) is infinite in extent that decays according to a damped exponential/sinusoidal (similarly to the behaviour of an ACP from a AR(d) process;

[1] A time series is classified as stationary when it is developed in time around a constant mean.

The PACF of AR(p) the process is finite, i.e. presents a cut after the lag q, like the behaviour of an ACP from a MA(p) process;;
The PACF of ARMA process (p,q) is similar to an ACF from a MA(q) process.

A general method for finding a.c. for a stationary process with f.a. is using the Yule-Walker equations.

This method seems to fail in the case of non-stationary time series (the irregular component is significant). To solve this issue we differentiate the non stationary series so many times (d times) as necessary to get a stationary series. After these differences of successive terms of the chain, applied d times, we can apply the same technique: identify which model(s) are identified from ACF and PACF. A model that represents a non-stationary series, differenciated d times, with an auto-regressive component with order p and a moving average component with order q is represents as an ARIMA(p,d,q).

To estimate the best models between several proposals, we usually apply the information criteria AIC, BIC: the best models have the lowest values of AIC and BIC. Also the log of likelihood function is a good statistic to evaluate the quality of the estimated models: the lowest value means a better model.

After selection, the models need to be validated. One of the rules is to analyze the residuals (the ith residual is the difference between the ith obervation and its estimate). Residuals are supposed to be Gaussian and non-correlated. To verify this can be use several statistical tests procedures and other techniques (Ljung-Box test, box-Pierce test, Kolmogorov-Smirnov test, Bera and Jarque test, some graphics, e.g. boxplots, qq plots).

The estimates precision evaluation is another step to include in all process. For that we can compute the usual measures: MAPE, MADE, etc.

In general the procedure of all process is composed by the following iterates:

1. Models formulation: use of ACF and PACF;
2. Models adjustment: estimation of model parameters, application of suitability measures of estimates;
3. Validation of models: selection of variables, diagnostics, residual analysis and interpretation.
4. Analysis of precision and updating the models.

4 Problem Description and Data Available

Infraquinta, E.M. is the water utility that manages the water and the waste water services of a well-known tourist place in Algarve (Portugal), known as Quinta do Lago. Tourism increase trend and climate change scenarios can be identified as two predominant drivers which will strongly influence Infraquinta, E.M. as a water utility. Due to a forecasted decrease in precipitation, especially in warm season, water supply at peak demand times could be reduced and water revenue limited. In this context Infraquinta, E.M. needs to reinforce mechanisms for predictive planning based on data analysis.

They would like to have an algorithm for evaluating water meters performance by using historical data (hourly water consumption). The main purpose of the algorithm should be to find out the meter performance breakpoint and hence where should water meters be replaced.

The main challenge of this task is segmenting the hourly water consumption data into different contributions: seasonality, trend and noise. The trend component is the one related to meter performance.

The available data is composed by:

1. Monthly data on water consumption, including water meter replacement (from 2006 to 2018), of one hotel;
2. Hourly data from 2 water meters, different models, installed in serie (from 2014 to 2018);
3. Hourly data from one water meter (from 2014 to 2018), replaced at 19/08/2014.

It should be noted that the information regarding the second water meter model, from (ii) dataset was not analyzed since it only regards information from 2017 until 2018, hence scarce information to be implemented a robust time series analyzes.

5 Empirical Application

5.1 Preliminary Approach

As it was referred in Sect. 4 data available in this problem have different registration options: hourly, daily, monthly. This is something that should not be ignored. Naturally, the hourly may be aggregated in days and, of course, days may be aggregated in months. But, one shall not forget that each time we group data, necessarily, some information is lost. Therefore, there is the need to, at least have a look on the 'pure' data, where 'pure' is used to mention the data as they were recorded.

5.2 A Brief Descriptive Analysis

Starting with a graphical inspection of the data representation, and considering the thinner representation presented in the data, the choice was to obtain the box-plot representation for the different hours of the day. From this we tried to observe similarities and dissimilarities, for the water meters (dataset (ii)).

Given one set of data has the hourly consumption of two water meters in a serie disposal, the same representation was considered for the other water meter.

As one can observe the graphical representations highlight the different behavior consumption, during day and during night. That is for the period after midnight until 8 a.m. the mean consumption is reduced but there are lots of anomaly observations with huge values for consumption. Even though the two water meters are in a serial disposal it looks like the consumptions are of different

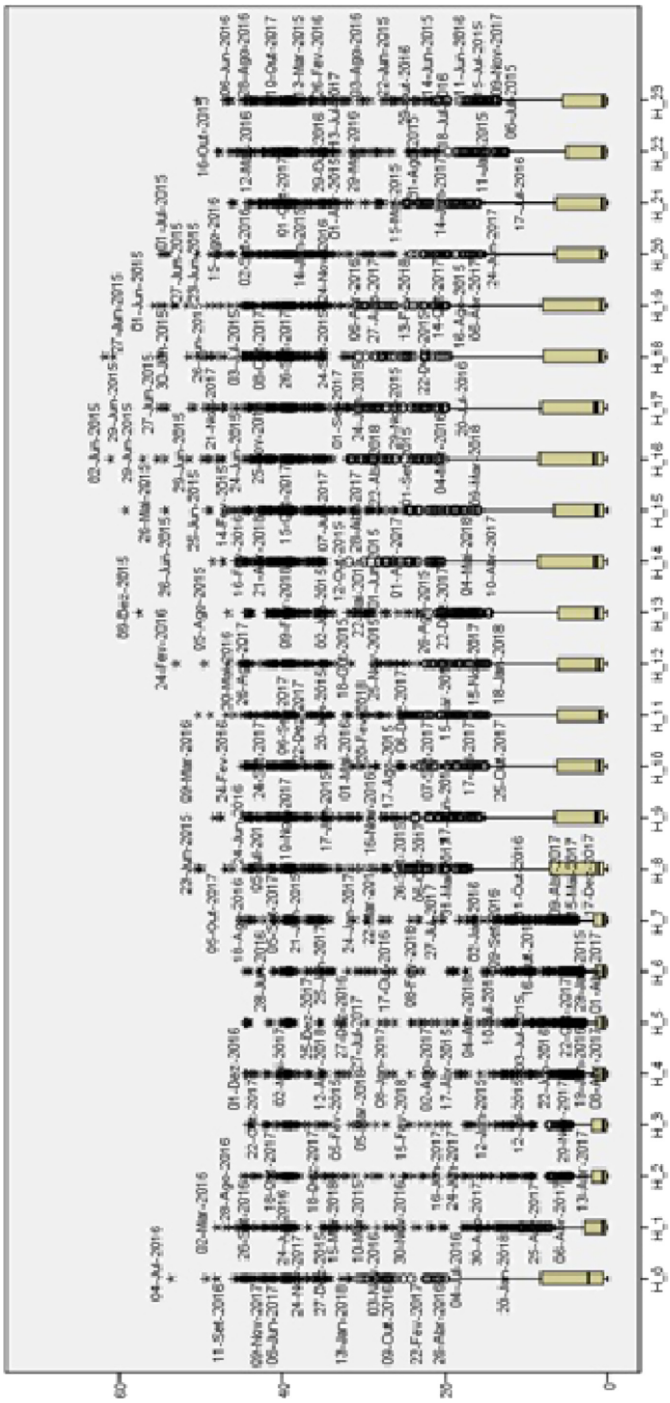

Fig. 1. Boxplot diagrams – water meter 1440.

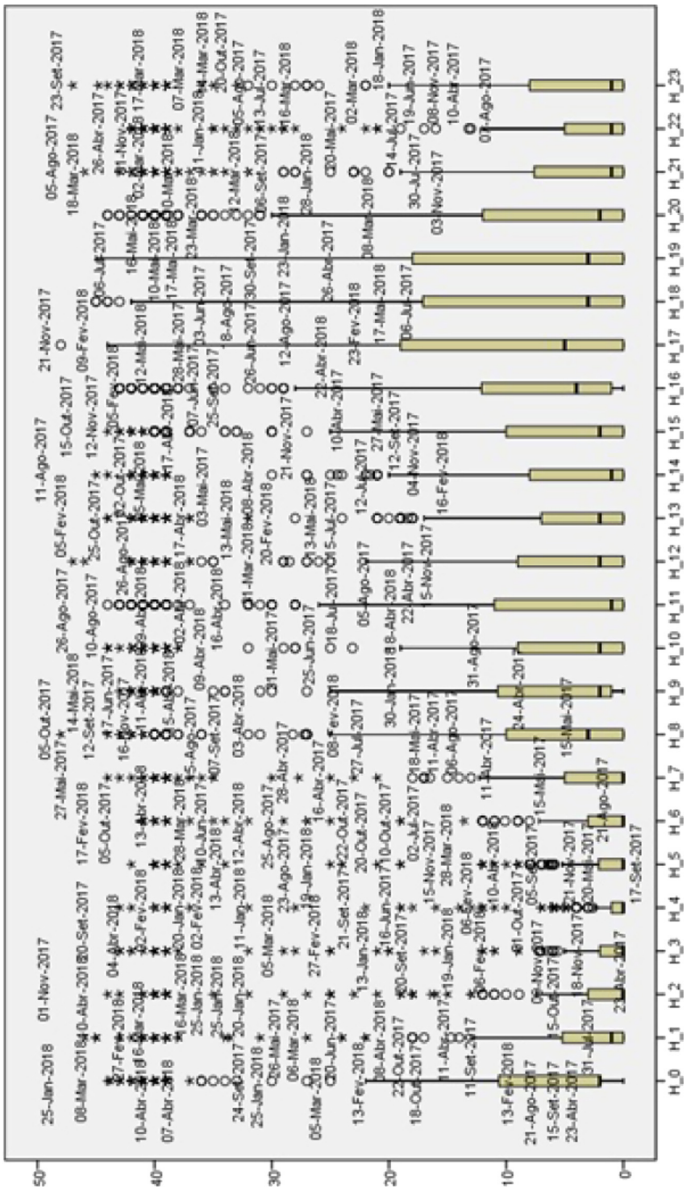

Fig. 2. Boxplot diagrams – water meter 2180.

populations. For water meter 1440 during night, about 75% of the consumption is very reduced but appears a large number of huge values, which should correspond to garden watering. Still the whole consumption seems more homogeneous in comparison to water meter 2180. In what concerns water meter 2180 there are

also anomalies but the water meter seems much more sensitive to the different hours, during a day consumption.

For the other set of data expressed in hours (dataset (iii)) has a huge number of zeros at several hours of the day (in many days), which is strange, and therefore no graphical representation was possible, that is, it is not possible to obtain the box-plot. The hourly box-plot representation was not possible to obtain for one data set since there were only monthly registrations of the consumption in one hotel (Figs. 1 and 2).

5.3 The Serial Disposal Water Meters 1440 and 2180 – Some Details

For the serial disposal two water meters, given the disposition of the equipment, what should be expected was, if things were working properly, the same measurements (or at least very close) for the consumption since they were measuring the same. Thus, after observing the data it was considered to obtain the daily aggregation and then to watch what was happening to the overlapping of both commonly days interval registrations, for which a representation is presented in Fig. 3.

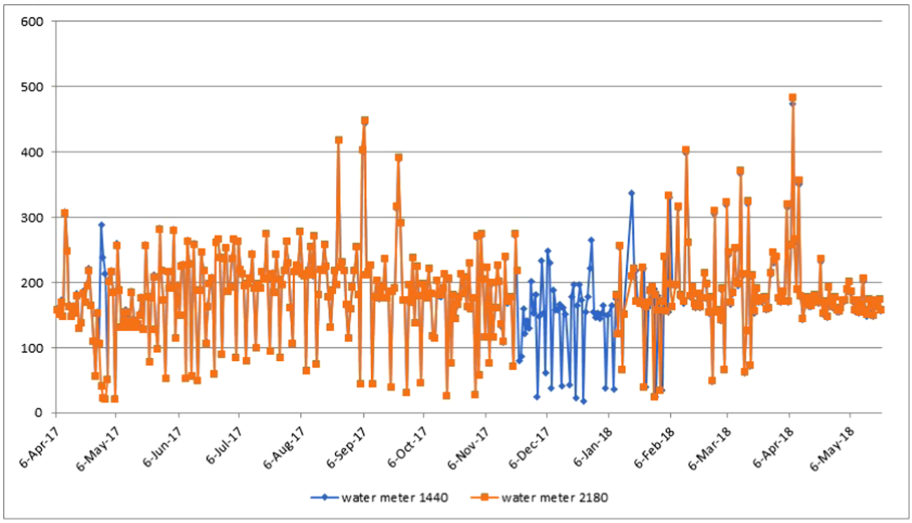

Fig. 3. Water meters series for both 1440 and 2180.

After that, and realizing that the measurement should be the same, it seemed natural to consider the difference between the measurements of each of the water meters on the same days of observation. And from that it was possible to check some irregularities.

When considered all the data registration of data consumption for the two serial disposal water meters was noticed that there is an interval time, from November 22 2017 until January 8 2018 for which water meter 2180 has no registered data. That is clearly identified in Fig. 4. Also, there are two peaks in the graph observed in distinct moments of time.

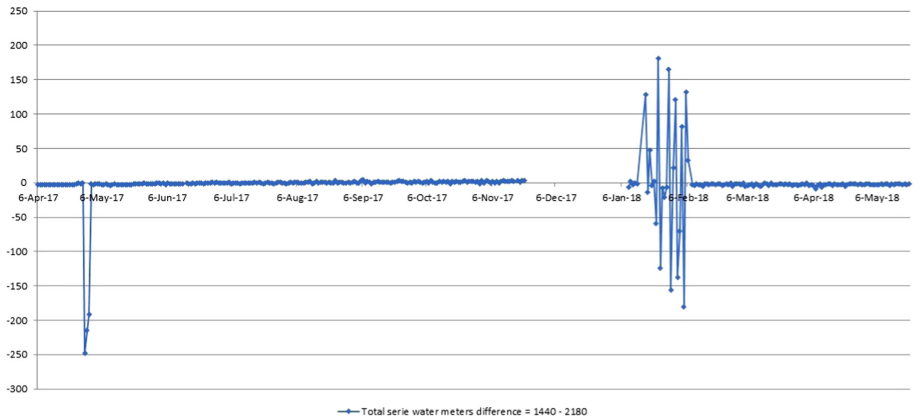

Fig. 4. Water meters difference series.

The first one, may even be an irregular occurrence lonely of a day. But the second peak of irregularities spreads for several days, see in more detail the zoom presented in Fig. 5. In fact, for a whole monthly the difference exhibits big discrepancies whether those are positive or negative.

When observed the whole series of differences, for a certain period it occurs that the difference is positive, that is, the water meter 1440 registers more consumption. It is only after a certain time (last measurements of 2018) that the difference between the measurements is negative, see again Fig. 5, that is, the water meter 2180 begins to measure more accurately, so it was expressed when that observation came up.

5.4 Water Meter 4773 – The Time Series (traditional) Approach

Considering the monthly data on water consumption, including a water meter replacement (dataset (i)), given these are monthly data and some external factors may influence the water consumption. For instance winter/summer seasons consumption or weather variables related with more or less rain, that may constraint the water consumption and it is not possible to have a detailed information on the daily consumption.

Therefore, the first differences, see Fig. 6, and the homologous differences were considered, see Fig. 7, which allows to remove partial auto–correlation issues an seasonality issues.

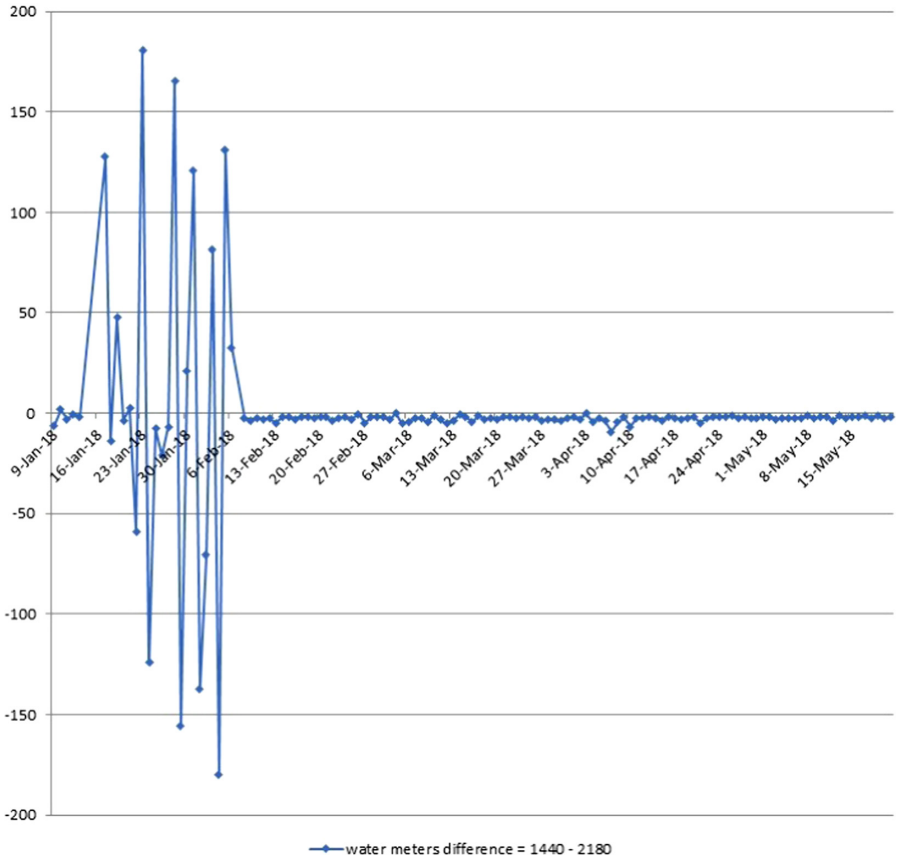

Fig. 5. Water meters difference zoom.

It is easily identified when the water meter was replaced (June 15, 2014) where a peak is clearly exhibited in the graph, in Fig. 6. In the same figure it is possible to identify as areas which may indicate the breakpoints – the slope is higher, in absolute terms, than the pattern or it is subsequently decreasing or subsequently increasing. To consider whether those are or not a breakpoints depends on the company definitions.

When the homologous differences are considered it is also possible to identify the water meter replacement, as already mentioned. In this, it is possible to observe that from January 2014 until June 2014 the consumption decreases.

Then the water meter is replaced (June 2014). From the replacement moment until June 2015 (a year) the water meter shows irregular consumption, which is probably the time needed to the water meter to stabilize, usual time necessary for this kind of equipment to work properly.

By last, were tried and obtained models using the traditional techniques, [25]. In Fig. 8 we have an additive model which decomposes the time series in:

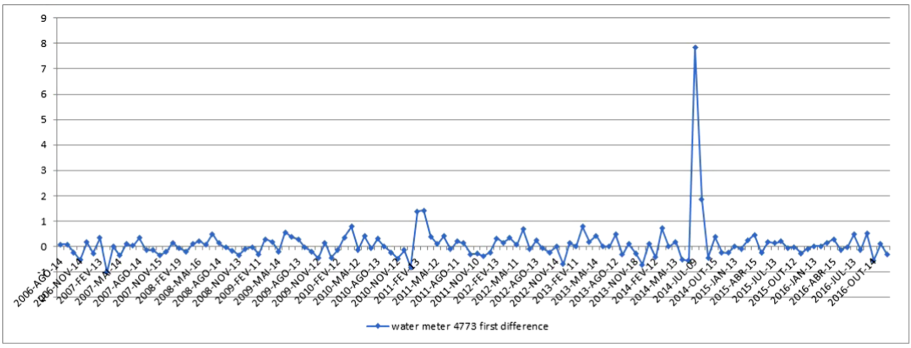

Fig. 6. First difference observation for 4773 water meter.

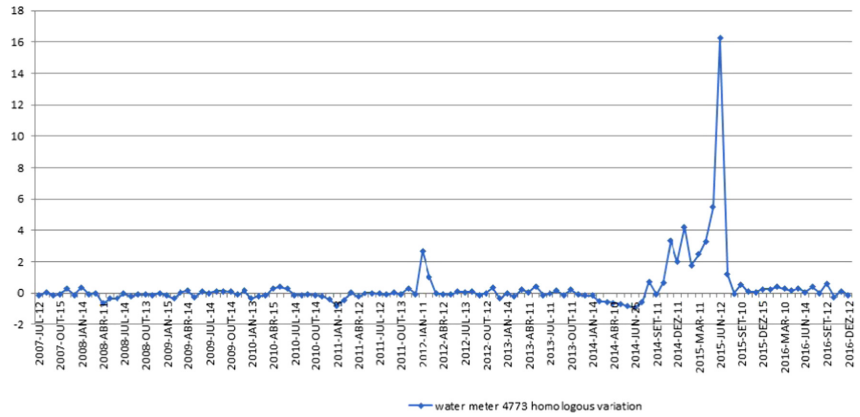

Fig. 7. Homologuos variation for 4773 water meter.

the consumption; the error for consumption from season; the seasonal adjusted series for consumption from season; the seasonal factors for consumption from season; the trend-cycle for consumption from season; with a length of seasonal period of 12 considering an additive effect.

The results obtained by this model are similar to the ones obtained by the non-parametric approach.

For the same data set, another model was obtained. In this case, some times series models were considered and tested. An ARIMA model was obtained presenting adequate goodness measures of fit and prediction.

Again, graphically, as it can be seen in Fig. 9, it is very similar to the non-parametric models.

These models may be considered as a preliminary approach of the non-parametric approach.

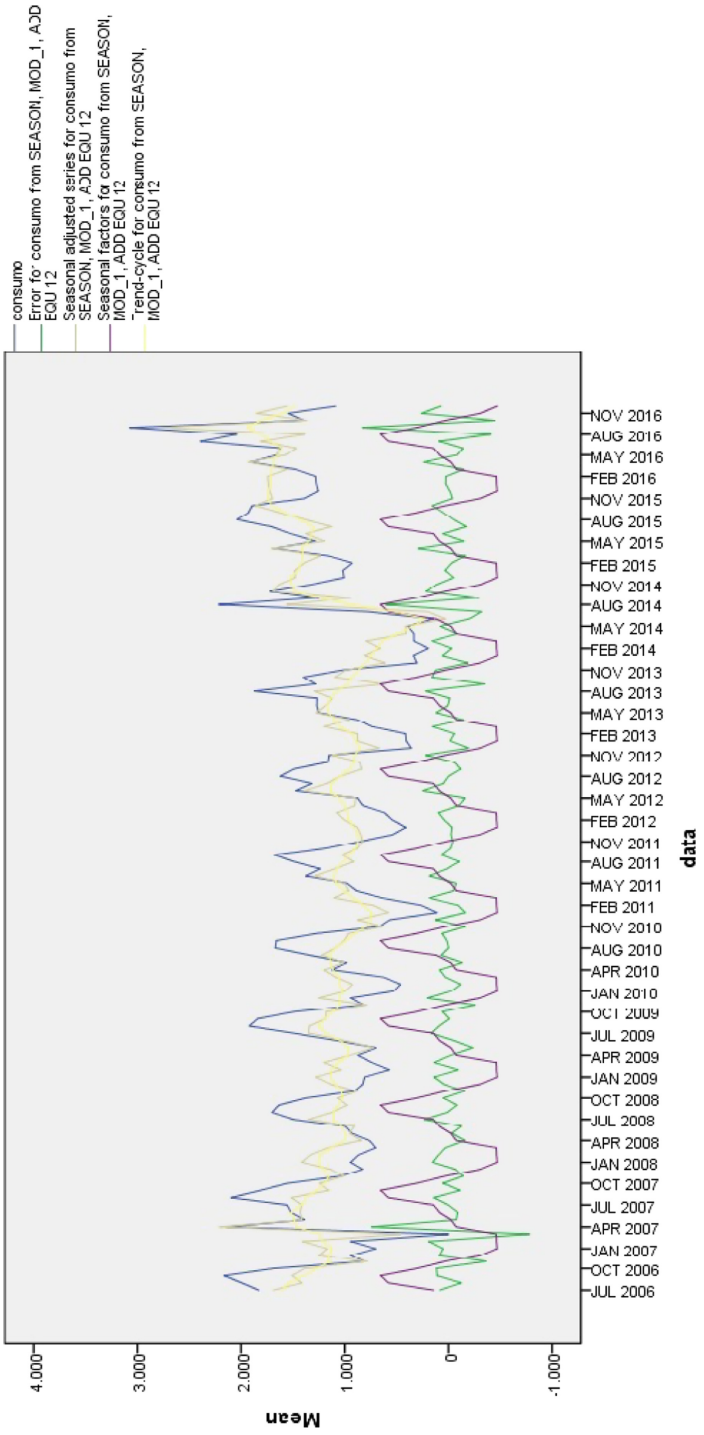

Fig. 8. Water meter 4773 adjusted series with additive effect.

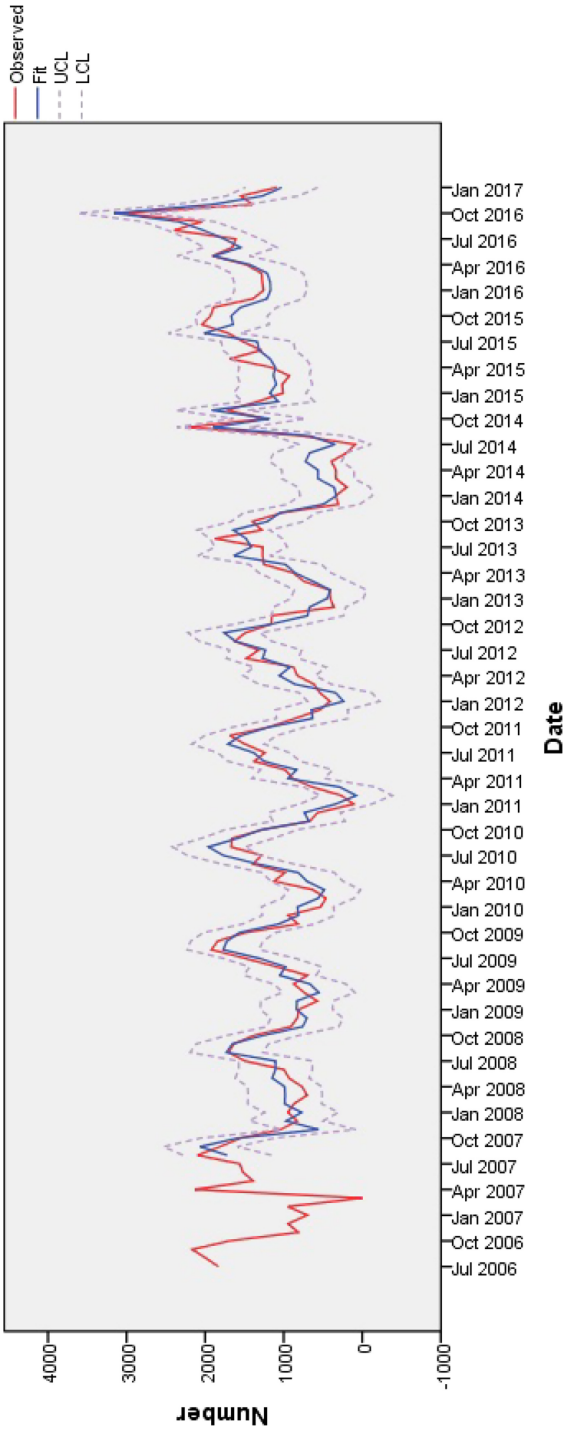

Fig. 9. Water meter 4773 ARIMA $(1, 1, 0)$ model adjustment.

6 Final Comments

In terms of the data analysis of the available data it is important to proceed with a preliminary descriptive analysis and also to consider the traditional time series analysis in a previous data analysis, in a way that the results may work as indicative measures of the non-parametric procedure following the work presented in [22–25] or a more recent approach [26,27] where a breaking point is identified using a recent Package Software [28] developped for time series structural changes. Both ways are complementary, and therefore, both can contribute to the problem solving. However, more work is required to optimize approaches to enable a significant contribution towards more sustainable urban water management. As future work, we can propose consider at least the parametric and non-parametric approach, using the multi-criteria algorithm proposed in [29].

Acknowledgements. This work was supported by Portuguese funds through the *Center of Naval Research* (CINAV), Portuguese Naval Academy, Portugal and *The Portuguese Foundation for Science and Technology* (FCT), through the *Center for Computational and Stochastic Mathematics* (CEMAT), University of Lisbon, Portugal, project UID/Multi/04621/2019.

References

1. Yang, Q., Wu, X.: 10 challenging problems in data mining research. Int. J. Inf. Technol. Decis. Making **5**(4), 597–604 (2006)
2. Fu, T.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
3. Weron, R.: Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int. J. Forecast. **30**(4), 1030–1081 (2014)
4. Vasimalla, K.: A survey on time series data mining. Int. J. Innovative Res. Comput. Commun. Eng. **2**(5), 170–179 (2014)
5. Pal, S.M., Palet, J.N.: Time series data mining: a review. Binary J. Data Min. Network. **5**, 01–04 (2015)
6. Zhu, Y.: High performance data mining in time series. Techniques and Case Studies. Ph.D. Thesis. Department of Computer Science. University of New York (2004)
7. Mulquiney, P.: Combining GLM and data- mining techniques for modelling accident compensation data. In: XIth Accident Compensation Seminar. Actuaries Institute, 2nd-4th April 2007, Australia. https://actuaries.asn.au/Library/3.a_ACS07_paper_Mulquiney_Combining
8. Ratanamahatana, C.A., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., Das, G.: Mining time series data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA (2009). https://doi.org/10.1007/978-0-387-09823-4_56
9. Watson, T.G., Christian, C.D., Mason, A.J., Smith, M.H., Meyer, R.: Bayesian-based pipe failure model. J. Hydroinformatics **6**, 259–264 (2004)
10. Kapelan, Z., Babayan, A.V., Savic, D.A., Walters, G.A., Khu, S.T.: Two new approaches for the stochastic least cost design of water distribution system. Water Supply **4**(5–6), 355–363 (2005)

11. Ragot, J., Maquin, D.: Fault measurement detection in an urban water supply network. J. Process Control **16**, 887–902 (2006)
12. Quevedo, J., et al.: Validation and reconstruction of flow meter data in the Barcelona water distribution network. Control Eng. Pract. **18**(6), 640–651 (2010). https://doi.org/10.1016/j.conengprac.2010.03.003
13. 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). https://doi.org/10.4314/wsa.v37i4.18
14. Sanchez, J. G., Motta, L. A., Alves, W. A.: Estimativa de volume de água não medido em ligações residenciais por perde de exactidão dos hidrômetros, na cidade de Juazeiro - BA. In: Congresso Interamericano de Engenharia Sanitária e Ambiental, 27ª, 2000, Porto Alegre. Anais eletrônicos. Porto Alegre-RS: ABES (Associação Brasileira de Engenharia Sanitária e Ambiental), II-051, pp 1–14 (2000)
15. Makki, A.A., Stewart, R.A., Beala, C.D., Panuwatwanich, K.: Novel bottom-up urban water demand forecasting model: revealing the determinants, drivers and predictors of residential indoor end-use consumption. Resour. Conserv. Recycl. **95**, 15–37 (2015)
16. Beal, C.D., Flynn, J.: Toward the digital water age: Survey and case studies of Australian water utility smart-metering programs. Utilities Policy **32**, 29–37 (2015)
17. Creaco, E., Kossieris, P., Vamvakeridou-Lyroudia, V., Makropoulos, C., Kapelan, Z., Savic, D.: Parameterizing residential water demand pulse models through smart meter readings. Environ. Model. Softw. **80**, 33–40 (2016)
18. Liu, A., Giurco, D., Mukheibir, P.: Urban water conservation through customised water and end-use information. J. Cleaner Prod. **112**(44), 3164–3175 (2016). https://doi.org/10.1016/j.jclepro.2015.10.002
19. Box, G.E., Jenkins, G.M., Reinsel, G.C.: Time Series: Forecasting and Control, 3rd edn. Prentice Hall, New Jersey (1994)
20. Chatfield, C.: The Analysis of Time Series: An Introduction, 6th edn. Chapman and Hall/CRC, Boca Raton (2004)
21. Hamilton, J.D.: Time Series Analysis. Princeton University Press, New Jersey (1994)
22. Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.: STL: a seasonal-trend decomposition procedure based on loess. J. Official Stat. **6**(1), 3–73 (1990)
23. Hyndman, R.J., Khandakar, Y.: Automatic time series forecasting: the forecast package for R. J. Stat. Softw. **26**(3), 1–22 (2008). http://www.jstatsoft.org/article/view/v027i03
24. Zeileis, A., Kleiber, C., Kramer, W., Hornik, K.: Testing and dating of structural changes in practice. Computat. Stat. Data Anal. **44**(109–123), 6 (2013). https://doi.org/10.1016/S0167-9473(03)00030-
25. Gonçalves, E., Lopes, N.M.: Séries temporais - Modelações lineares e não lineares, 2nd edn. Sociedade Portuguesa de Estatística, Lisboa, Portugal (2008)
26. Cristina, S., Cordeiro, C., Lavender, S., Costa Goela, P., Icely, J., Newton, A.: MERIS phytoplankton time series products from the SW Iberian Peninsula (Sagres) using seasonal-trend decomposition based on loess. Remote Sens. **8**(6), 449 (2016)
27. Goela, P.C., Cordeiro, C., Danchenko, S., Icely, J., Cristina, S., Newton, A.: Time series analysis of data for sea surface temperature and upwelling components from the southwest coast of Portugal. J. Mar. Syst. **163**, 12–22 (2016)

28. Zeileis, A., Leisch, F., Hornik, K., Kleiber, C.: Strucchange: an R package for testing for structural change in linear regression models. R Package version 1.5 (2015)
29. Fernandes, S., Clímaco, J. and Captivo, M. E.: A cooperative multicriteria group decision aiding tool - a guided tour of the desktop application. In: Proceedings of Group Decision and Negotiation - GDN 2017, Stuttgart-Hohenheim, Germany, pp. 361–368 (2017)