Data reconstruction of flow time series in water distribution systems – a new method that accommodates multiple seasonality

Rui Barrela, Conceição Amado, Dália Loureiro and Aisha Mamade

ABSTRACT

The purpose of this paper is to present a simple yet highly effective method to reconstruct missing data in flow time series. The presence of missing values in network flow data severely restricts their use for an adequate management of billing systems and for network operation. Despite significant technology improvements, missing values are frequent due to metering, data acquisition and storage issues. The proposed method is based on a weighted function for forecast and backcast obtained from existing time series models that accommodate multiple seasonality. A comprehensive set of tests were run to demonstrate the effectiveness of this new method and results indicated that a model for flow data reconstruction should incorporate daily and seasonal components for more accurate predictions, the window size used for forecast and backcast should range between 1 and 4 weeks, and the use of two disjoint training sets to generate flow predictions is more robust to detect anomalous events than other existing methods. Results obtained for flow data reconstruction provide evidence of the effectiveness of the proposed approach.

Key words | data reconstruction, flow data, forecasting models, multiple seasonality, TBATS model, water distribution systems

Rui Barrela (corresponding author)
Dália Loureiro
Aisha Mamade
Urban Water Division,
National Laboratory for Civil Engineering,
Avenida Brasil 101,
Lisbon 1700-066,
Portugal
E-mail: *r.guerra.barrela@tecnico.ulisboa.pt*

Conceição Amado
Department of Mathematics and CEMAT,
Instituto Superior Técnico, Universidade de Lisboa,
Avenida Rovisco Pais 1,
Lisbon 1049-001,
Portugal

INTRODUCTION

Urban water systems need to ensure adequate customer service and to improve efficiency through water loss control. Water losses represent a significant economic and environmental inefficiency in most water utilities. Flow monitoring through SCADA (Supervisory Control And Data Acquisition) or telemetry systems that remotely collect data from flow meters is one of the most important tools to improve network operation and management and to ensure the efficient use of water. The number of meters installed in networks has increased remarkably as a result of technological advances. However, extracting useful information can be a difficult task due to the need to combine data from multiple meters and the fact that flow data are often faulty (e.g., missing, duplicate or out of range values).

Missing data might be due to problems with flow meters, sensors, data loggers and central database communication, or with data processing. Having complete and accurate flow data is essential for a reliable billing system and a high quality customer relation and network management. Filling gaps from flow series is also relevant for online anomalous event detection (Loureiro *et al.* 2016a) and for a proper water loss assessment. Examples of this assessment include the calculation of annual water balance (Lambert & Hirner 2000; Alegre *et al.* 2006), or more detailed and complementary methods, such as night flow analysis (Farley & Trow 2003).

Since network flow data are usually characterized by daily and weekly cycles (de Marinis *et al.* 2008; Mamade 2013), the focus will be on models that can accommodate multiple seasonality.

Multiple seasonality models are used, for instance, in electricity load demand forecasting. Mohamed *et al.* (2010) investigated the use of a double seasonal ARIMA (Auto

doi: 10.2166/hydro.2016.192

Regressive Integrated Moving Average) model for electricity load demand forecasting in the context of electric power planning. Hassan *et al.* (2012) compared double seasonal ARIMA and double seasonal ARFIMA (Auto Regressive Fractionally Integrated Moving Average) models for forecasting half-hourly electricity load demand, with the ARFIMA model producing slightly better results.

In the context of water demand forecasting, various models and techniques have already been explored but the generality of work has been dedicated chiefly to forecasting and did not consider the problem of missing data. Alvisi *et al.* (2007) developed a short-term forecasting procedure of hourly water demand based on two modules: a daily module, which included annual and weekly seasonality, and an hourly module, which incorporated the intra-day patterns. Li & Huicheng (2010) employed multiple linear regression and fuzzy neural network in order to model the trend and cyclical components of yearly urban water demand, respectively. Caiado (2009) examined the forecasting performance of several univariate time series models, including Holt-Winters model, ARIMA model and GARCH (Generalized Auto Regressive Conditional Heteroskedasticity) model to predict daily water demand. Firat *et al.* (2010) compared several artificial neural network (ANN) techniques to forecast monthly water demand. Herrera *et al.* (2010) compared several techniques, such as ANN, projection pursuit regression, multivariate adaptive regression splines, random forests and support vector regression, to forecast hourly water demand. Quevedo *et al.* (2010) employed two forecasting models in order to validate and reconstruct missing and false flow meter data in water distribution systems: a daily model based on ARIMA time series model and a 10-min intra-day model based on a 10-min demand pattern, determined through correlation analysis and an unsupervised fuzzy logic classification (LAMDA algorithm).

To forecast intricate time series, De Livera *et al.* (2011) introduced a model incorporating Box–Cox transformations, Fourier representations with time varying coefficients and ARMA (Auto Regressive Moving Average) error correction. This enables forecasting complex seasonal time series, such as those with multiple seasonal periods, high-frequency seasonality, non-integer seasonality, and dual-calendar effects, and as such, can cover a broad range of applications. They applied this model to forecast electricity demand, gasoline and call centre data.

Despite the work put forth by previous studies, namely Quevedo *et al.* (2010), the need remains for a simple and robust methodology that can accommodate multiple seasonality to reconstruct online or offline flow data in water distribution networks. The novelty of this work resides in: (i) the proposed weighted function of the forecast and backcast values obtained from multiple seasonality time series models; (ii) its application in the reconstruction of water flow data; (iii) an extensive study of the performance of the proposed method in common situations for this type of data analysis, namely the existence of outliers and multiple seasonality.

The paper is organized as follows. In the Methodology section, the proposed approach, the tests carried out are presented and the general testing procedure is described. First, the forecasting model is defined and incorporated into three different reconstruction methods: a Forecast Method, a Backcast Method and a Combined Method. Furthermore, several tests are carried out to justify the model selection, to assess the window size for training the model, to study model robustness in terms of anomalous events' location in the training window, and in terms of reconstruction method, and finally to compare prediction accuracy in different reconstruction methods. The outcome of these tests is presented in the Results and discussion section. In the Conclusions section, we present the main conclusions drawn from the study.

METHODOLOGY

The proposed approach is based on a weighted function of forecasts and backcasts for estimating gaps in flow time series with multiple seasonality. Data were collected from existing flow meters that monitored the inflow in three different urban sectors, whose boundaries were clearly defined and watertight to ensure an adequate network operation and management. Electromagnetic flow meters with 150 mm of nominal diameter registered readings with 10- or 15-minute time steps throughout a SCADA system.

The first steps of data processing are known as data validation and normalization (Mamade 2013; Loureiro *et al.*

2016a). Data validation consists of detecting and correcting anomalous data caused by the measuring equipment, such as duplicate or out of range values. These values were removed as they might bias the subsequent analysis. Data normalization, in turn, consists of setting a regular time step for the measurement variable. A 15-minute time step was defined and water flow units were set to m^3/h. Once these first steps have been completed, the next step is data reconstruction.

Only daily and weekly seasonality were considered in the current study. Annual seasonality was not included, since only 1 year of historical data was used. The availability of accurate flow data from multiple years is less frequent among the water utilities. Moreover, to test the method it was important to analyse periods where network operation was stable (i.e., clearly defined and watertight network boundaries and constant operating configurations), in order to ensure data consistency throughout the time period.

The TBATS model

There are several approaches to forecasting time series with multiple seasonality, particularly in intra-day time series. However, for demand forecasting, the De Livera *et al.* (2011) forecasting model was adopted, since it is recent and adequate to model dynamic seasonality. This model is known as TBATS, an acronym for its key features: Box–Cox transformation, ARMA errors, trend, and seasonal components (the initial T standing for trigonometric, as in trigonometric representation of seasonal components). It can be formalized as follows. Consider a realization of a stochastic process with N positive observations, i.e., the sequence of positive observed data $\{y_t\}_{t=1}^{N}$, where y_t is the observation at time t. Applying a Box–Cox transformation, defined as:

$$y_t^{(\omega)} = \begin{cases} \dfrac{y_t^{\omega} - 1}{\omega}, & \text{if } \omega \neq 0 \\ \log y_t, & \text{if } \omega = 0 \end{cases}, \tag{1}$$

with parameter ω, we then have:

$$y_t^{(\omega)} = l_{t-1} + \phi b_{t-1} + \sum_{i=1}^{T} s_{t-m_i}^{(i)} + d_t, \tag{2}$$

where

$$l_t = l_{t-1} + \phi b_{t-1} + \alpha d_t \tag{3}$$

is the local level in period t,

$$b_t = (1-\phi)b + \phi b_{t-1} + \beta d_t \tag{4}$$

is the short-run trend in period t with b as the long-run trend, and

$$d_t = \sum_{i=1}^{p} \varphi_i d_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{5}$$

denotes an ARMA (p, q) process, with ε_t as a Gaussian white-noise process with zero mean and constant variance σ^2. The parameters of the ARMA model are given by φ_i $(i = 1, \ldots, p)$ and θ_i $(i = 1, \ldots, q)$.

The smoothing parameters are given by α and β, ϕ represents the damping parameter, and $m_1, \ldots, m_T$ denote the seasonal periods.

Furthermore,

$$s_t^{(i)} = \sum_{j=1}^{k_i} s_{j,t}^{(i)} \tag{6}$$

represents the i-th seasonal component at time t with the following trigonometric formulation based on Fourier series:

$$s_{j,t}^{(i)} = s_{j,t-1}^{(i)} \cos\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \sin\lambda_j^{(i)} + \gamma_1^{(i)} d_t, \tag{7}$$

$$s_{j,t}^{*(i)} = -s_{j,t-1}^{(i)} \sin\lambda_j^{(i)} + s_{j,t-1}^{*(i)} \cos\lambda_j^{(i)} + \gamma_2^{(i)} d_t, \tag{8}$$

where $s_{j,t}^{(i)}$ is the stochastic level of the i-th seasonal component, and $s_{j,t}^{*(i)}$ is the stochastic growth in the level of the i-th seasonal component that is needed to describe the change in the seasonal component over time. The smoothing parameters are given by $\gamma_1^{(i)}$, $\gamma_2^{(i)}$, and $\lambda_j^{(i)} = 2\pi j/m_i$, while k_i denotes the number of harmonics required for the i-th seasonal component, $i = 1, \ldots, T$.

To fit this model it is necessary to estimate not only the smoothing parameters and the damping parameter, but also the Box–Cox transformation parameter ω, as well as the

ARMA coefficients p and q. The implemented R function (Team 2014) of the model automatically handles these estimates, as well as optimal model selection through the Akaike information criterion (AIC) (Akaike 1973). The AIC is used in order to determine the best fit, and therefore allows a decision as to whether to use the Box–Cox transformation, whether to include a trend, whether to include a damping parameter in the trend and whether to include ARMA errors.

The Forecast Method

The initial idea for data reconstruction is to iteratively fit a forecasting model to the data preceding each sequence of consecutive missing values, and then generate forecasts in order to fill each sequence with reasonable values. In this study, this procedure is referred to as the Forecast Method.

Since flow data are provided with a regular 15-minute time step, there are four time steps in 1 hour, and $4 \times 24 = 96$ time steps in 1 day. Therefore, daily and weekly seasonalities are accounted for TBATS model with $T = 2$ seasonal components containing 96 and $96 \times 7 = 672$ time steps, respectively. The initial approach to flow data reconstruction is to apply the Forecast Method with this double seasonal TBATS model. A second TBATS model that only accommodates daily seasonality (with $T = 1$ seasonal component containing 96 time steps) is also considered. To compare TBATS models with a classical approach, the Forecast Method was also applied with a seasonal ARIMA model.

Furthermore, depending on the location of the sequence of missing values in the time series, the Forecast Method may not be applicable. For instance, if the very first values of flow data are missing, the Forecast Method lacks the flow data needed for fitting the model. In this case, there is a need for a reconstruction method that incorporates the flow data succeeding the sequence of missing values, in order to generate predictions for the missing past values.

The Backcast Method

Having high-resolution data (15-min time steps) allowed fitting the model in different sections of complete data, which enabled computing not only forecasts, but also backcasts. The term backcasting is introduced as a means to back-forecast the unknown past values (Wei 2006). This concept was applied in the context of flow data reconstruction: if we consider a given sequence of missing values, the flow data succeeding the sequence may be used to fit a model, thus generating predictions for the preceding missing values.

In essence, the Backcast Method allows us to predict missing values if the Forecast Method is not applicable due to lack of data. In instances where both methods are applicable, two sets of predictions are generated, which can then be combined into a third reconstruction method – the Combined Method.

The Combined Method

We consider that the uncertainty of the predictions generated by a forecast model should increase as we get further away from the left bound of the prediction window. Conversely, the predictions generated with a backcasting model are progressively more reliable as we approach the right bound of the prediction window. Therefore, when considering a sequence of missing values, a combination of predictions generated by the Forecast Method and the Backcast Method should assign progressively less weight to the forecast predictions, and progressively more weight to the backcast predictions.

The proposed Combined Method consists of a simple weighted combination of the forecast and backcast for a given sequence of missing values, and is constructed as follows:

$$c_i = \delta_i \times \text{forecast}_i + (1 - \delta_i) \times \text{backcast}_i, \; i = 1, \ldots, l \quad (9)$$

with

$$\delta_i = \begin{cases} 1/2, l = 1 \\ \dfrac{l - i}{l - 1}, l > 1 \end{cases} \quad (10)$$

where l is the length of the sequence of missing values, forecast_i and backcast_i are the i-th component of the forecast and backcast prediction sequences, respectively, and c_i is the prediction for i-th component of the sequence of missing values, as generated by the Combined Method.

When the Forecast Method (resp. the Backcast Method) is unable to generate predictions due to lack of data, the

Combined Method consists only in the application of the Backcast Method (resp. the Forecast Method).

Nevertheless, the Combined Method usually takes advantage of two disjoint sets of flow data in order to compute the predictions.

General testing procedure

Evaluation is the key to assess the actual performance of the prediction methods, and splitting data into training and testing sets is a central part of this evaluation (Witten & Frank 2005). The use of a set of independent data (test set), but with the same distribution of the training set, avoids overfitting and allows obtaining the performance characteristics of the models. In the scope of this study, the test set corresponds to a section of data that is removed from the time series data, previous to the application of a data reconstruction method. The training set is composed of a window of adjacent data preceding the test set (in the Forecast Method), succeeding the test set (in the Backcast Method), or both (in the Combined Method).

Prediction accuracy is determined by the following performance measures: the root-mean-square error (RMSE), a normalized root-mean-square error (NRMSE) and the mean absolute scaled error (MASE). The RMSE is a simple, useful measure that allows the figure to have the same dimensionality as the quantity being produced (Witten & Frank 2005). As scale invariant measures, the NRMSE and the MASE are used when the test requires comparison between predictions on flow data from several district metering areas (DMAs). These three measures generate a fair assessment of the prediction accuracy, and are defined as follows. Let y_t be the observed value at time t, and $\hat{y}_t$ the corresponding prediction value, with $t = 1, \ldots, n$. Then

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \tag{11}$$

and

$$NRMSE = \sqrt{\frac{\sum_{t=1}^{n} ((\hat{y}_t - \hat{\mu}/\hat{\sigma}) - (y_t - \mu/\sigma))^2}{n}} \tag{12}$$

with μ and σ as the mean value and standard deviation of the set of observed values, and $\hat{\mu}$ and $\hat{\sigma}$ as the mean value and standard deviation of the set of prediction values. Additionally,

$$MASE = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{(n/n-1) \sum_{t=2}^{n} |y_t - y_{t-1}|} \tag{13}$$

where the denominator corresponds to the average forecast error of a one-step naïve forecast method, in which the forecast is the previous observed value.

The key aspects of the reconstruction approach have been analysed through a series of five tests:

- In Test 1, we analysed the prediction accuracy of the Forecast Method with three different models: a daily seasonal ARIMA model, a daily seasonal TBATS model, and a daily and weekly seasonal TBATS model.
- In Test 2, we addressed the issue of the window size for fitting the TBATS model.
- In Test 3, we studied the robustness of the TBATS model by creating artificial anomalous events at various instants in the training data and then analysing their impact on the forecasts.
- In Test 4, we analysed the impact that the Combined Method has on robustness.
- In Test 5, we compared the prediction accuracy of the Forecast Method, the Backcast Method and the Combined Method.

RESULTS AND DISCUSSION

Collected data consist of three flow time series belonging to different DMAs in 2013. A full year view of the data is represented by the aggregate daily medians for DMA 1, DMA 2 and DMA 3 (see Figure 1). Note that we have chosen to represent the aggregated time series using the sample median as it is robust to noise and outliers. In order to illustrate daily and weekly cycles, the decomposition of components generated by a double seasonal TBATS model fitted on a window of 4 consecutive weeks of flow data for DMA 3 (see Figure 2) is presented. In Figure 2, the first plot represents the original observed flow data

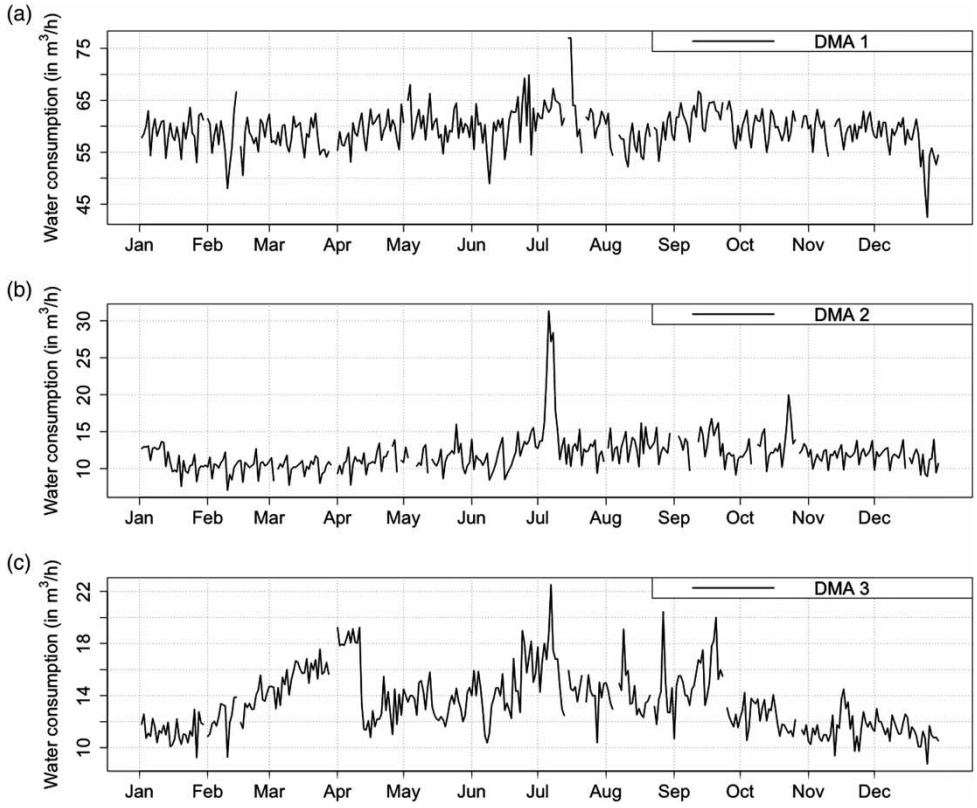

Figure 1 | Daily medians of the year 2013 for DMA 1 (a), DMA 2 (b), and DMA 3 (c).

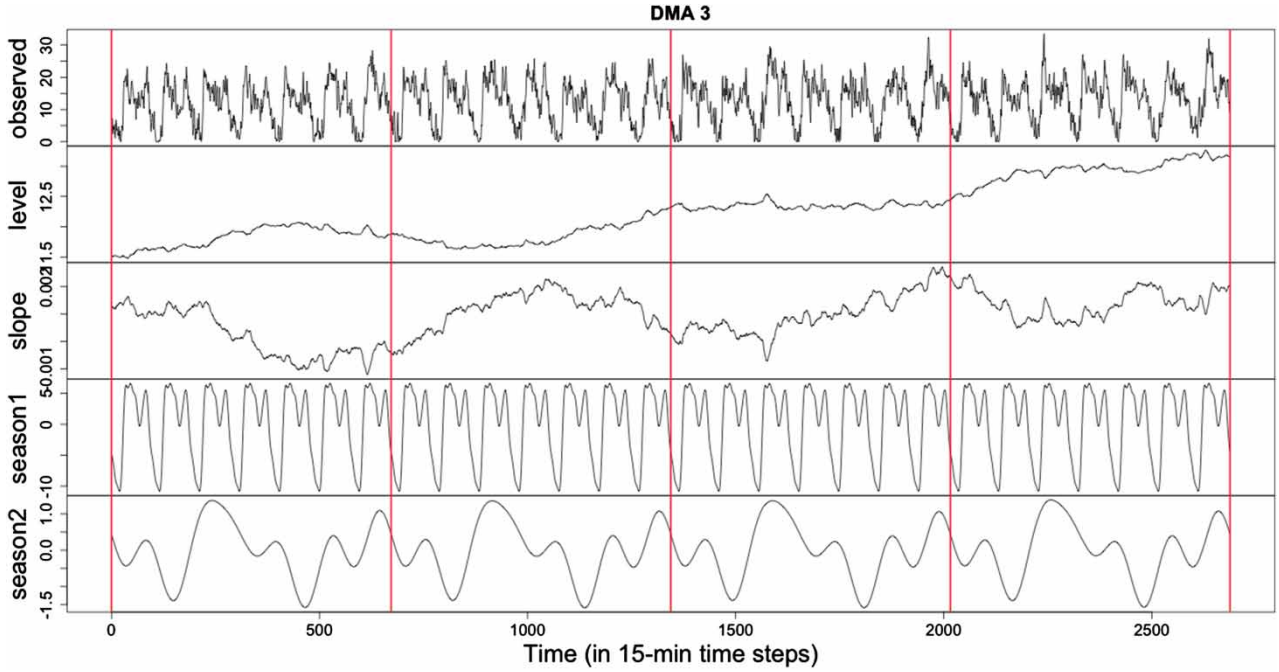

Figure 2 | TBATS model decomposition for DMA 3.

values, the slope (or trend) is the derivative of the level, season1 represents the daily seasonal component, and season2 represents the weekly seasonal component. The vertical lines indicate the window limits of each week. We observe that daily seasonality is highly prominent in the flow data when comparing the range of the daily seasonal component to the range of the observed values. On the other hand, weekly seasonality is less pronounced, although still present. Monitored DMAs correspond to network sectors where the domestic consumption is the most relevant component. Therefore, daily seasonality is generally more noticed, with lower consumption during the night period and significant consumption throughout the day, especially during the lunch and dinner periods. In terms of weekly seasonality, most DMAs show a distinct behaviour between working days, where people stay out of their homes most of the day, and weekends. This effect might be less pronounced in sectors where elderly people or inactive workers predominate and consumption habits are very similar among the week days (Loureiro *et al.* 2016a, 2016b).

Test 1: impact of seasonal effects on model forecasts

In this test we compared the prediction accuracy of the Forecast Method with three different forecasting models: a classic ARIMA model, a daily seasonal TBATS model, and a daily and weekly seasonal TBATS model. Other models found in the literature were used mainly for daily, weekly or even monthly water values, as opposed to the intra-day values of the flow data provided for this study. The model selected for this test was a daily seasonal ARIMA model which is expressed in factored form by ARIMA-$(2, 0, 0)(0, 0, 1)_{96}$, chosen based on a stepwise selection criterion and AIC (Hyndman & Khandakar 2007). Given a time series $\{y_t; t \in \mathbb{Z}\}$, the model is formulated as follows:

$$y_t = \upsilon + \frac{1 - \theta_{96,1}B^{96}}{(1 - \phi_1 B - \phi_2 B^2)}\varepsilon_t,\ t \in \mathbb{Z} \tag{14}$$

where υ is the expected value term, $\theta_{96,1}B^{96}$ is the seasonal moving average part, $(1 - \phi_1 B - \phi_2 B^2)$ is the seasonal autoregressive part, B is the usual backshift operator and $\{\varepsilon_t; t \in \mathbb{Z}\}$ is a Gaussian white-noise process with variance σ_ε^2.

The three models were fitted on a window size of 3 weeks, the forecast window was set to 1 week and the process was repeated for each DMA.

In Figure 3, we present the forecasts obtained from each model for DMA 2. We note that in Figure 3 only the 2 most recent weeks of the training set are represented, in order to better view the results.

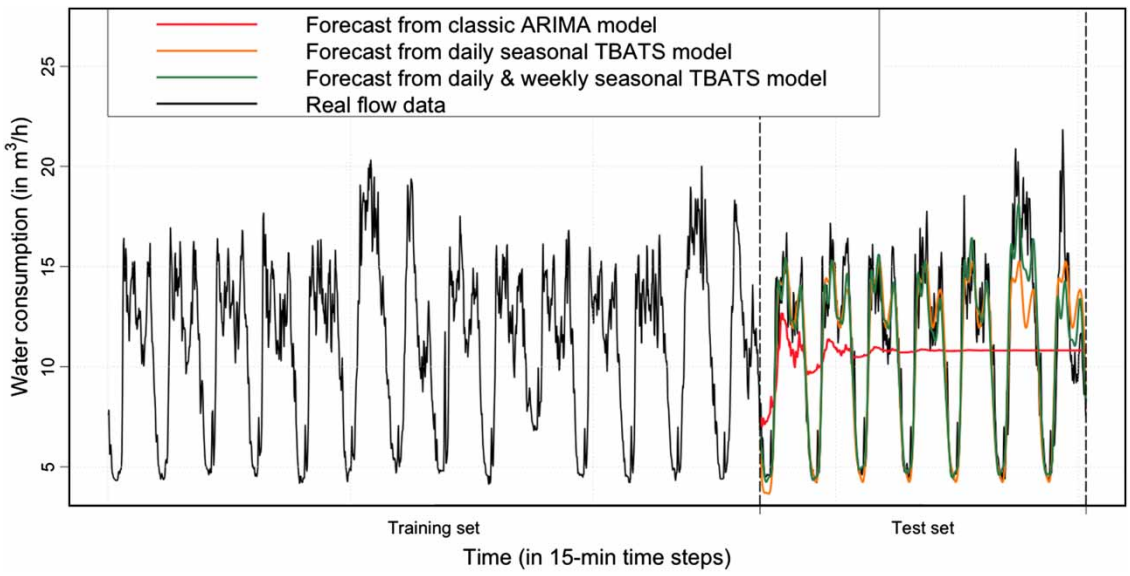

Figure 3 | Forecasts from ARIMA, daily TBATS, and daily and weekly TBATS models (in DMA 2).

In Figure 4 we present the NRMSE (Figure 4(a)) and MASE (Figure 4(b)) of each model forecast, for each DMA. We note that the ARIMA model forecast presented the highest errors in every case. The daily seasonal TBATS model and the daily and weekly seasonal TBATS model both presented low errors in every case, while the daily and weekly seasonal TBATS model presented the lowest errors in general. These results indicate that the ARIMA model is unsuitable to this type of data. Furthermore, there is evidence that a model incorporating both daily and weekly seasonalities is better suited to the flow data, as opposed to incorporating daily seasonality alone.

Test 2: impact of window size for daily and weekly seasonal TBATS model fitting

This test focuses on the assessment of the effect of the window size assigned for fitting the forecasting model when conducting the Forecast Method (the result of this test also holds for the Backcast Method). Results obtained from Test 1 indicated that the daily and weekly seasonal TBATS model is the most suitable to the flow data. Therefore, this test was conducted for that model only.

The test was performed for each DMA as follows. The test set was fixed and assigned a window size of 1 week. Since the daily and weekly seasonal TBATS model incorporates weekly seasonality, the minimum window size for fitting was 1 week. The window size for fitting was then iteratively increased by 1 week, reaching a maximum of 4 weeks.

In Figure 5 we present the NRMSE (Figure 5(a)) and MASE (Figure 5(b)) of each forecast, for each DMA. The NRMSE plot indicates a slight tendency towards lower errors with the increase of the window size. The MASE plot shows a comparatively high error for a window of 1 week in DMA 2.

In general, results indicate that there is virtually no difference in terms of prediction error by selecting either 2, 3 or 4 weeks for model fitting, with a tendency toward lower errors as the window size increases. Therefore, we conclude that the reconstruction algorithm should consider the maximum length of complete data available for fitting each model. We note that the window size considered for this test is somewhat limited, and future tests should include a finer granularity of window size.

Test 3: impact of location of anomalous event in training set

In this test we studied the robustness of the daily and weekly seasonal TBATS model: artificial anomalous events were created at various times in the data used for fitting the model (training set), and the change in the prediction error of the Forecast Method was analysed.

The test was performed for each DMA as follows. The daily and weekly seasonal TBATS model was fitted on a window size of 3 weeks, and the forecast window was set

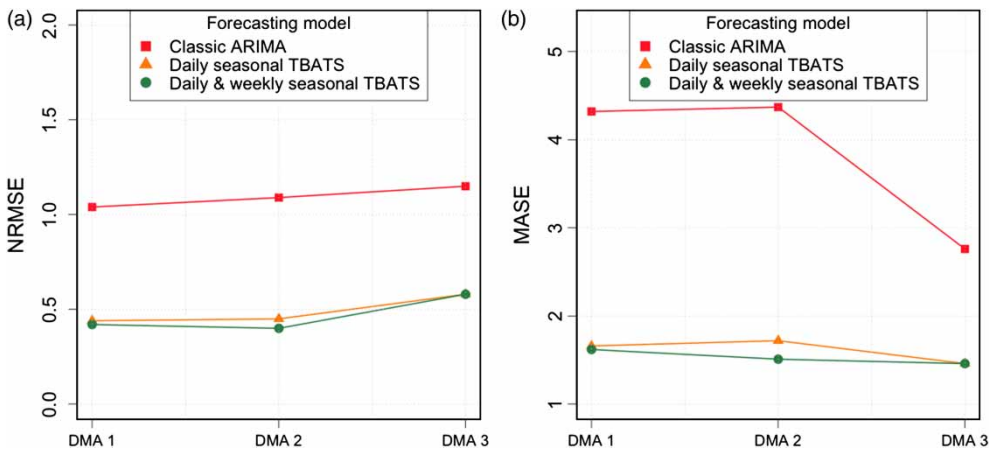

Figure 4 | NRMSE (a) and MASE (b) of forecasts from ARIMA, daily TBATS, and daily and weekly TBATS models.

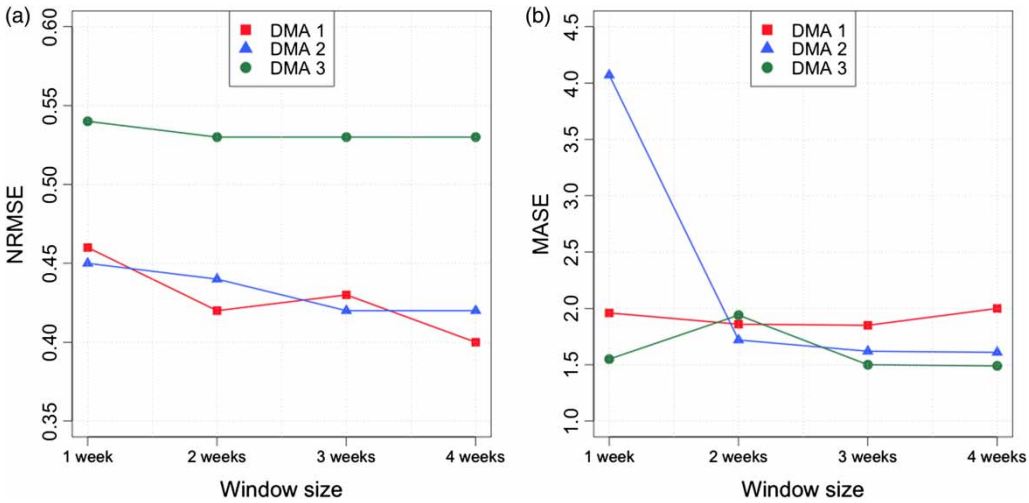

Figure 5 | NRMSE (a) and MASE (b) of forecasts from daily and weekly TBATS model, by window size of training set.

to 1 week. As indicated by Test 2, it would be best to use the largest number of weeks possible. However, the data are frequently faulty and a trade-off had to be reached between having enough data for model fitting and minimizing the prediction error.

An artificial anomalous event was created in the training set. In the context of network flow data, several types of anomalous events may take place (e.g., pipe bursts, atypical consumptions due to anomalous water uses, infrequent tanks or pumping stations operational conditions). For this study, a rule of thumb was adopted to simulate a reported pipe burst – usually characterized by a moderate to high flow rate whose effect may last a few hours before detection (Loureiro *et al.* 2016a). Therefore, an event with a 6-hour duration was simulated in the following way: a section of flow data corresponding to a time window of 6 hours was selected, and multiplied by a factor of 2. The model was then fitted on the new data, and the forecast was generated. The impact of the anomalous event was studied by iteratively placing the event in successive days preceding the test set (at the same time of day) and calculating the error of each forecast.

In Figure 6 we present the MASE of each forecast, for each DMA. Results from the NRMSE plot indicated that

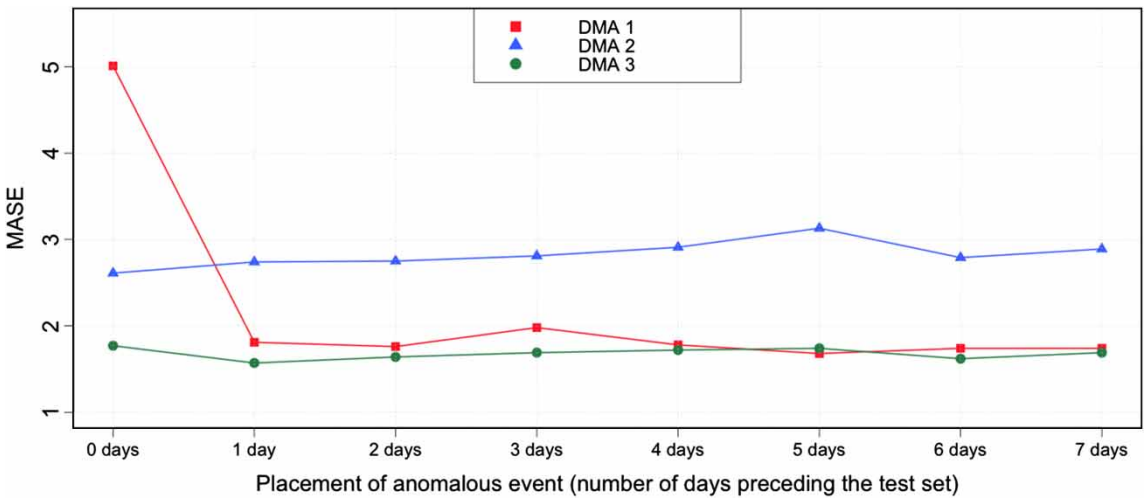

Figure 6 | MASE of forecasts from daily and weekly TBATS model, by placement of anomalous event.

there is generally very little difference in prediction error regardless of the placement of the anomalous event, and therefore only the MASE plot is presented. In Figure 6, the MASE for DMA 1 indicated a higher error when the event immediately precedes the test set (0 days preceding the test set), which suggests that the model assigns more weight to the most recent observations of the training set.

In Figure 7 we present the plot of the forecast resulting from the training set without anomalous events (Figure 7(a)), and the training set with the anomalous event placed at 0 days preceding the test set (Figure 7(b)). Only the most recent week of the training set is represented for a better visualization. It is clear from these plots that the forecast adequately resembled the real data when no anomalous event was placed in the training set. When the anomalous event was placed, the resulting model forecast took generally higher values, as the model interpreted the event as an increase in the level of the time series, instead of an irregular occurrence. Additionally, future work should include testing the robustness of the model with real anomalous events, identified and validated by the water utilities.

Test 4: impact of Combined Method on robustness

This section focuses on dealing with the lack of robustness issue following the application of the Forecast Method on DMA 1, as illustrated in Test 3. After applying the Forecast Method, the Backcast Method and the Combined Method were applied in order to generate predictions for the same test set.

In Figure 8, we present the plot of the training and test sets with the predictions of the Forecast and Backcast Methods (Figure 8(a)), as well as the plot of the test set with the estimates generated by all three reconstruction

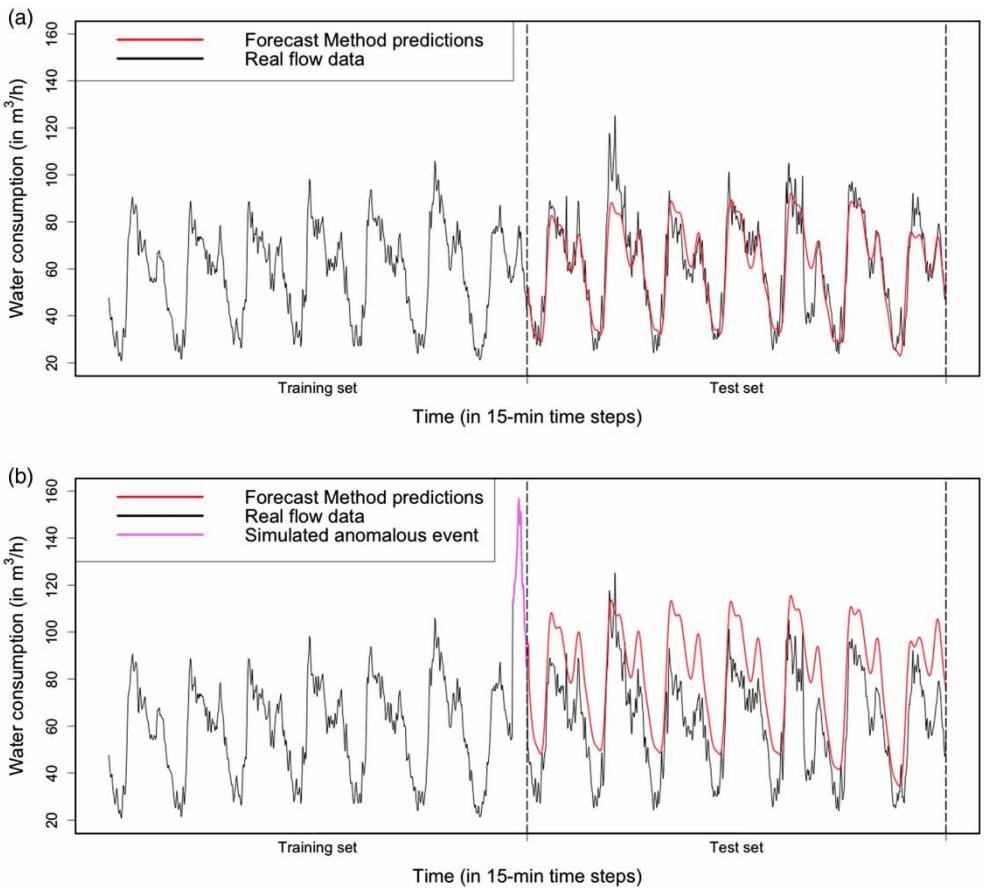

Figure 7 | Forecasts of daily and weekly TBATS model, with and without anomalous event (in DMA 1).

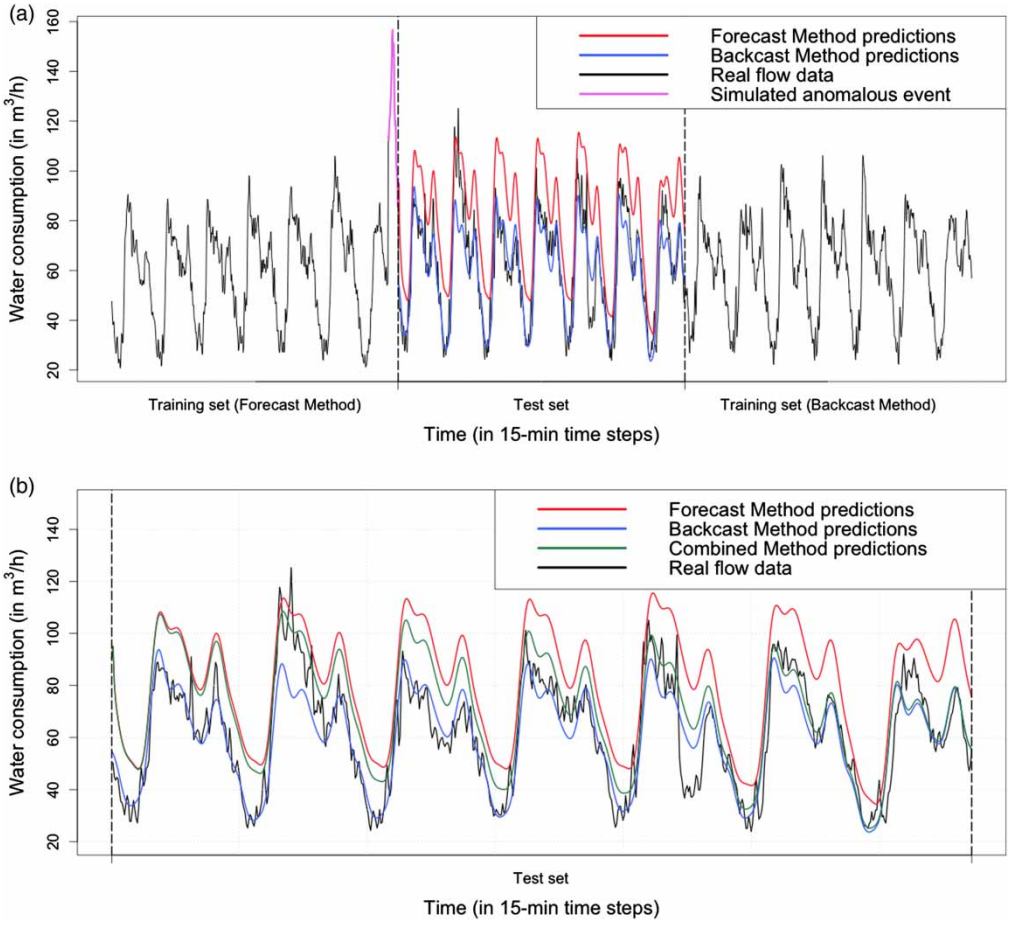

Figure 8 | Forecast and Backcast Method predictions (a) and Combined Method predictions (b) with anomalous event.

methods (Figure 8(b)). Only the most recent week of the training set of the Forecast Method is represented in Figure 8(a) for a better view of the results.

The largest amount of complete flow data available for computing the backcast was equivalent to a window size of 1 week, as opposed to the window size of 3 weeks for the forecast. Nevertheless, it is apparent from Figure 8(a) that the Backcast Method generated more accurate predictions than the Forecast Method. Indeed, the RMSE was equal to 21.41 m^3/h for the Forecast Method, 9.1 m^3/h for the Backcast Method and 13.8 m^3/h for the Combined Method.

In practice, if a detection step of anomalous events is not performed on the data before the data reconstruction, it is not possible to know whether an event will fall on the training set of the Forecast Method or the Backcast Method. Had the anomalous event been placed in the training set of the Backcast Method, the Forecast Method would have performed better. Therefore, the reconstruction method that would make sense as a compromise for an automated process would be the Combined Method, which would attenuate the impact of the anomalous event regardless of its location.

Test 5: comparison of prediction performance

The aim of this test is to compare and decide which of the data reconstruction methods selected for each approach is more suitable to the flow data provided. In this study, we focus on accurately reconstructing missing data in the short term. In this test, we perform a reconstruction of flow data with a test set window size equivalent to 1 day. For each day of the week, a section of flow data corresponding to that day is selected at

random as the test set. Then, the Forecast, Backcast and Combined Methods are applied, generating three sets of predictions for each day of the week. The performance measures are determined, and the process is repeated for each DMA.

In Figure 9 we present the RMSE of each set of predictions, for each day of the week and for each DMA. We note that the error of the predictions generated by the Combined Method is always the lowest or second-lowest of the three, except for one case (Wednesday for DMA 3, in Figure 9(c)). We conclude that the Combined Method successfully reduces the error of the least accurate reconstruction method (whether that method is the Forecast Method or the Backcast Method), and in several cases generates the most accurate predictions of all three.

CONCLUSIONS

Flow data reconstruction in water distribution systems is an essential step towards improving the billing system and network operation, namely water loss control, and is achieved through the imputation of missing values with accurate predictions.

In this paper, a new method for filling missing values was developed and tested, which comprised a combination of forecast and backcast values generated by TBATS and ARIMA models. An extensive set of tests that evaluated the suitability and robustness of the method was carried out, which yielded effective results and highlighted the advantages of the Combined Method for offline data reconstruction, over a simple forecast or backcast approach.

In summary, models for flow data reconstruction should incorporate daily and seasonal components for more accurate prediction; the window size used for forecast and backcast should comprise between 1 and 4 weeks, which reflects a compromise between the typical length of datasets with continuous and complete records available and the accuracy gain. Since the Combined Method uses two disjoint training sets to generate flow predictions, it is more robust to anomalous events than are other existing methods. However, in order to better assess the adaptability of the proposed Combined Method, it should be tested on a larger number of flow data time series. Furthermore, as

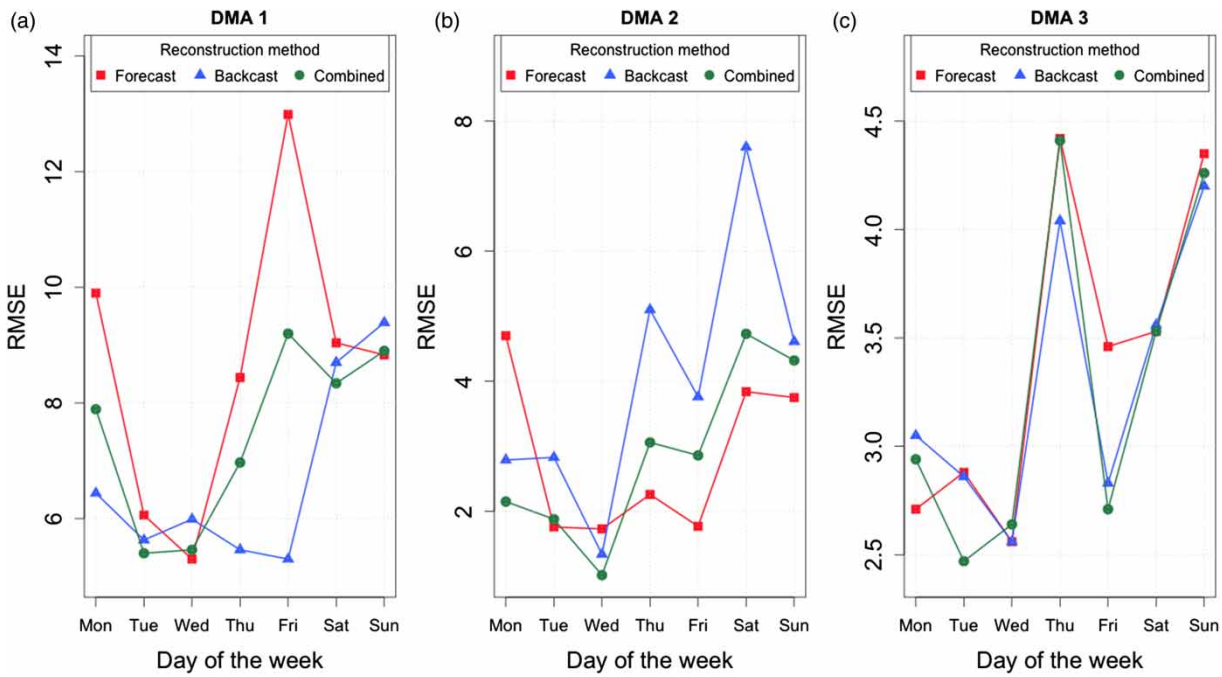

Figure 9 | RMSE of reconstruction method predictions by day of the week for DMA 1 (a), DMA 2 (b), and DMA 3 (c).

part of ongoing research, we are comparing the performance of the proposed methods with alternatives from the literature.

It is also worth noting that the Combined Method benefits from the so-called 'offline approach' to data reconstruction, since it makes use of historical flow data in order to generate predictions. Nevertheless, the proposed method is very flexible and for online flow data reconstruction only the Forecast Method should be applied.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to the iPerdas project and the water utilities that participated for providing the data used in this study.

REFERENCES

Akaike, H. 1973 Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov & F. Csáki (eds), *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR*, September 2–8, 1971, Akadémiai Kiadó, Budapest, Hungary, pp. 267–281.

Alegre, H., Baptista, J. M., Cabrera Jr, E., Cubillo, F., Duarte, P., Hirner, W., Merkel, W. & Parena, R. 2006 *Performance Indicators for Water Supply Services.* IWA Publishing, London, UK.

Alvisi, S., Franchini, M. & Marinelli, A. 2007 A short-term, pattern-based model for water-demand forecasting. *J. Hydroinform.* **9** (1), 39–50.

Caiado, J. 2009 Performance of combined double seasonal univariate time series models for forecasting water demand. *J. Hydrol. Eng.* **15** (3), 215–222.

De Livera, A. M., Hyndman, R. J. & Snyder, R. D. 2011 Forecasting time series with complex seasonal patterns using exponential smoothing. *J. Am. Stat. Assoc.* **106** (496), 1513–1527.

de Marinis, G., Gargano, R. & Tricarico, C. 2008 Water demand models for a small number of users. In: *Water Distribution Systems Analysis Symposium 2006*, Cincinnati, OH, USA.

Farley, M. & Trow, S. 2003 *Losses in Water Distribution Networks: A Practitioner's Guide to Assessment, Monitoring and Control.* IWA Publishing, London, UK.

Firat, M., Turan, M. E. & Yurdusev, M. A. 2010 Comparative analysis of neural network techniques for predicting water consumption time series. *J. Hydrol.* **384** (1), 46–51.

Hassan, S., Ahmad, M. & Mohamed, N. 2012 A comparison of the forecast performance of double seasonal ARIMA and double seasonal ARFIMA models of electricity load demand. *Appl. Math. Sci.* **6** (135), 6705–6712.

Herrera, M., Torgo, L., Izquierdo, J. & Pérez-García, R. 2010 Predictive models for forecasting hourly urban water demand. *J. Hydrol.* **387** (1), 141–150.

Hyndman, R. & Khandakar, Y. 2007 Automatic time series forecasting: the forecast package for R 7, 2008. https://www.jstatsoft.org/article/view/v027i03 (accessed 24 February 2016).

Lambert, A. & Hirner, W. 2000 *Losses From Water Supply Systems: Standard Terminology and Recommended Performance Measures.* IWA Blue Pages. IWA, London, UK.

Li, W. & Huicheng, Z. 2010 Urban water demand forecasting based on HP filter and fuzzy neural network. *J. Hydroinform.* **12** (2), 172–184.

Loureiro, D., Amado, C., Martins, A., Vitorino, D., Mamade, A. & Coelho, S. T. 2016a Water distribution systems flow monitoring and anomalous event detection: a practical approach. *Urban Water J.* **13**, 242–252.

Loureiro, D., Mamade, A., Cabral, M., Amado, C. & Covas, D. 2016b A comprehensive approach for spatial and temporal water demand profiling to improve management in network areas. *Water Resour. Manage.* **30** (10), 3443–3457.

Mamade, A. 2013 *Profiling Consumption Patterns Using Extensive Measurements*. MSc Thesis, Universidade de Lisboa, Lisbon, Portugal.

Mohamed, N., Ahmad, M., Ismail, Z. & Suhartono, S. 2010 Double seasonal ARIMA model for forecasting load demand. *Matematika* **26**, 217–231.

Quevedo, J., Puig, V., Cembrano, G., Blanch, J., Aguilar, J., Saporta, D., Benito, G., Hedo, M. & Molina, A. 2010 Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract.* **18** (6), 640–651.

Team, R. C. 2014 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2012.

Wei, W. 2006 *Time Series Analysis: Univariate and Multivariate Methods,* 2nd edn. Pearson Addison Wesley, New York, USA.

Witten, I. H. & Frank, E. 2005 *Data Mining: Practical Machine Learning Tools and Techniques,* 3rd edn. Morgan Kaufmann, San Francisco, CA, USA.

First received 8 September 2015; accepted in revised form 6 October 2016. Available online 17 December 2016