Data reconstruction of flow time series in water distribution networks

Rui Miguel de Sousa Guerra Barrela
r.guerra.barrela@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

June 2015

Abstract

The presence of missing values in flow data severely restricts the use of these data for billing/customers' management and network control in water distribution systems. Missing values are frequent due to problems with metering, acquisition and flow data storage. In order to reconstruct the missing data, a new procedure is developed, based on a forecasting model which can accommodate the several seasonality cycles present in the data. This procedure takes advantage of the large set of historical data available by computing predictions based on data preceding the missing values (forecasts) but also on the data succeeding them (backcasts), subsequently creating a combination of both sets of predictions. Additionally, a two-level reconstruction procedure that has shown good results in the Barcelona water network flow data is implemented and improved upon. A number of tests are conducted in order to justify the proposed procedures and improvements, including robustness tests and prediction accuracy comparison tests.
Keywords: Data reconstruction. Flow measurements. Forecasting models. Multiple seasonality. Water distribution systems.

1. Introduction

The partitioning of water distribution systems into a set of sub-systems, in which the respective inflows are monitored continuously, constitutes a good method for operational management. By performing data analysis on these water consumption flow data, it is possible to manage water loss, detect atypical consumption behaviour, and control the operation of pumps and valves. Therefore, the existence of a reliable and complete consumption data history is fundamental. In order to properly monitor a complex water distribution network in real time, existing SCADA or telemetry systems collect and validate data taken regularly from flow meters and sensors placed throughout the network. Each flow meter is associated to a District Metering Area (DMA).

The flow data are often faulty (e.g., missing, duplicate or out of range values). Missing data in particular usually result from problems with the equipment's power supply, with the communication system between the sensors and data loggers, or with the exiting data storage and processing. In order to estimate useful statistics, such as the instantaneous, daily or monthly water consumption value, it is necessary to have complete and accurate data. Besides the importance for billing and customers management, complete and accurate flow data are crucial to improve water loss (e.g., night flow analysis) and demand management (e.g., large consumers consumption analysis). Hence the aim of this study is flow data reconstruction with accurate values.

The estimation of missing values in time series is essentially achieved through predictions generated by a model, which is based on the available data. As such, the subject of forecasting is key, and is found frequently throughout the literature. Also, since this type of intra-day flow data usually shows strong evidence of daily and weekly cycles [5, 1, 8], the focus will be mainly on models that can accommodate multiple seasonalities. We define seasonality in a univariate time series as a regular pattern of changes that repeats over a number of time periods. In this study, only daily and weekly seasonalities were considered. Since the complete extent of the flow data provided for testing consists of one year, the annual seasonality effect was not considered.

Multiple seasonality models are used, for instance, in electricity load demand forecasting. [9] investigated the use of a double seasonal ARIMA model for electricity load demand forecasting in the context of electric power planning. The data is regular with half-hour intervals and the model includes both daily and weekly seasonalities. [7] com-

pared double seasonal ARIMA and double seasonal ARFIMA models for forecasting half-hourly electricity load demand, with the ARFIMA model producing slightly better results. [12] proposed a new exponential smoothing formulation for forecasting time series with daily and weekly cycles, which produced good results when compared to other model forecasts of electricity load data. [6] presented a simple neural model with local learning for forecasting time series with multiple seasonal cycles, applying it to electrical load forecasting and comparing the results with ARIMA and exponential smoothing approaches. Double seasonal models were also applied to mobility network traffic prediction.

In the context of water demand forecasting, various models and techniques were explored. [1] developed a short-term forecasting procedure of hourly water demand based on two modules: a daily demand module, which included annual and weekly seasonalities, and an hourly demand module, which incorporated the intra-day patterns. [2] examined the forecasting performance of several univariate time series models (Holt-Winters, ARIMA and GARCH) in the prediction of daily water consumption. Weekly and yearly cycles were incorporated into the procedure in order to take into account both seasonalities, and different combinations of the forecasts were computed in the interest of improving accuracy. [3] presented a system for demand analysis and forecasting in water supply systems which included a daily demand forecasting model as well as standard load profiles of hourly consumption. [10] employed a two-level method in order to validate and reconstruct missing and false flow meter data of a water distribution network: a daily model based on ARIMA time series and an intra-day model based on demand patterns.

[4] introduced a model incorporating Box-Cox transformations, Fourier representations with time varying coefficients, and ARMA error correction. The model enables the forecasting of complex seasonal time series such as those with multiple seasonal periods, and was applied to electricity demand, gasoline and call centre data.

In this study, new robust and automatic methods are proposed to fill missing values in multiple seasonality flow time series. Specifically, two different methods are constructed based on the work found on [10] and on [4], respectively.

This paper is structured in the following way. In Section 2, we present two approaches for flow data reconstruction and indicate the tests that were conducted in order to select the best reconstruction method for each approach. In Section 3, we present the results from the tests, and in Section 4, we draw the main conclusions from the study.

2. Methodology for data reconstruction

2.1. General overview

This Section is dedicated to the formulation of flow data reconstruction methods according to two different approaches: the TBATS approach (see Section 2.2), and the JQ approach (see Section 2.3). Several reconstruction methods are presented and discussed, and the most suitable method is selected in each approach through comparative analysis and robustness tests, which are described throughout. The general training/testing procedure and the performance measures used are described in Section 2.4.

2.2. The TBATS approach

2.2.1 Overview

The TBATS approach refers to the application of the time series forecasting model TBATS (as described in [4]) in data reconstruction methods. In Section 2.2.2, the TBATS model is formalised. In Section 2.2.3, we describe and formalise the proposed data reconstruction methods that incorporate the TBATS model: the Forecast Method, the Backcast Method and the Combined Method.

2.2.2 The TBATS model

TBATS is an acronym for the key features of the model: Box-Cox transformation, ARMA errors, Trend, and Seasonal components (the first T standing for Trigonometric, as in trigonometric representation of seasonal components).

The model can be formalised as follows. Consider a realisation 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:

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

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 smoothing parameters are given by α and β, while ϕ represents the damping parameter, and $m_1, ..., m_T$ denote the T 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, with $i = 1, \ldots, T$.

Parameter estimation In order 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 ([11]) of the model automatically handles these estimates, as well as optimal model selection through AIC ([4]).

2.2.3 Application in data reconstruction methods

The Forecast Method In order to perform data reconstruction, the initial idea 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 the scope of this study, this procedure is referred to as the Forecast Method.

Since the flow data provided for this study consist of 15-minute regular time steps, there are 4 time steps in a window of 1 hour, and a window of 1 day is composed of $4 \times 24 = 96$ time steps. Therefore, daily and weekly seasonalities are accounted for with a 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 then to apply the Forecast Method (with Daily/Weekly Seasonal TBATS model).

On the other hand, the Forecast Method can also be applied with other forecasting models. In order to assess the need for a double seasonal model, the Forecast Method was applied with a classic ARIMA model, and also with a TBATS model that accommodates the daily seasonality only (with $T = 1$ seasonal component containing 96 time steps). A comparative analysis of the prediction accuracy of the Forecast Method with these three models is addressed in Section 3.2 as Test 1.

Another concern is defining the window size of flow data for fitting the model at each iteration. The selection of the window size for training the TBATS model is addressed in Section 3.3 as Test 2.

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 Since we are working with sufficiently large historical data, it is possible to compute backcasts in addition to forecasts. In [13], the notion of backcasting is introduced as a means to back-forecast the unknown past values. It is possible to apply this concept 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 in case 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.

The Combined Method We consider that the uncertainty of the predictions generated by a time series forecast model should increase with the size of the prediction window. As such, given a sequence of missing values, the corresponding predictions generated with the Forecast Method are progressively less reliable as the prediction window extends. Analogously, the corresponding predictions generated with the Backcast Method are progressively more reliable. 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 predictions for a given sequence of missing values, and is constructed as follows:

$$c_i = \delta_i \times fc_i + (1 - \delta_i) \times bc_i, \quad i = 1, \ldots, l \quad (9)$$

with

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

where l is the length of the sequence of missing values, fc_i and bc_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.

On the occasion that the Forecast Method (resp. the Backcast Method) is unable to generate predictions due to lack of data, the Combined Method consists in the application of the Backcast Method (resp. the Forecast Method) alone. A comparative analysis of the Forecast Method, the Backcast Method, and the Combined Method is addressed in Section 3.4 as Test 3.

The Combined Method may also help to attenuate the influence of anomalous data on the predictions, since the values generated by the Combined Method rely on predictions generated by two models fitted on disjoint sets of data. The analysis of this effect is addressed in Section 3.6 as Test 5.

2.3. The JQ approach

2.3.1 Overview

The JQ approach refers to the implementation, adaptation, and proposed improvements of the flow data reconstruction method first described in [10].

For simplicity, the original method is referenced throughout the text as the JQ Method, after the initials of the first author of [10] and consists of a two-level procedure. First, an aggregate daily flow model, built on the basis of an ARIMA model, is applied to reconstruct the aggregate daily flow data. Then, a set of daily flow distribution patterns is determined, which takes into account the intra-day variation. Finally, the two levels are combined in a flow model.

In Section 2.3.2, the original aggregate daily flow model of the JQ Method is formalised, and an alternate method for reconstructing the aggregate flow data is proposed. In Section 2.3.3, we discuss the construction of the daily flow distribution patterns of the JQ Method, and propose a more robust procedure. In Section 2.3.4, we formulate the resulting flow model.

2.3.2 Level 1: Aggregate daily flow data reconstruction

Aggregate daily flow data reconstruction is conducted by applying the Forecast Method to aggregate daily flow data, with forecasting models that are suitable to the aggregate data.

Original ARIMA-based aggregate daily flow model This model was created after a time series analysis on daily aggregate data consistently showed a weekly seasonality and deterministic periodic components [10]. Let $y_p(k)$ be the prediction for day k. The model structure is given by

$$y_p(k) = -b_1 y(k-1) - b_2 y(k-2) - b_3 y(k-3) - b_4 y(k-4) - b_5 y(k-5) - b_6 y(k-6) - b_7 y(k-7), \quad (11)$$

where

$$\begin{aligned} b_1 &= a_1 - \eta, \\ b_2 &= a_2 - \eta a_1 + \eta, \\ b_3 &= a_3 - \eta a_2 + \eta a_1 - 1, \\ b_4 &= a_4 - \eta a_3 + \eta a_2 - a_1, \\ b_5 &= -\eta a_4 + \eta a_3 - a_2, \\ b_6 &= \eta a_4 - a_3, \\ b_7 &= -a_4 \end{aligned}$$

and $\eta = 2\cos(2\pi/7) + 1$.

The Least Squares Method is used in order to adjust the model parameters. Furthermore, historical data free of faults is required.

For simplicity, the resulting reconstruction method for aggregate daily flow data is referred to as the Forecast Method (with ARIMA-based model), denoting the original aggregate data reconstruction procedure of the JQ Method.

Proposed Weekly Seasonal TBATS model for aggregate data reconstruction Assuming that the aggregate daily data retains the weekly seasonality of the original flow data, it is possible to accurately reconstruct the aggregate daily flow by applying the Forecast Method with a TBATS model that accommodates weekly seasonality only. The proposed model is then a TBATS model with $T = 1$ seasonal component containing 7 time steps, since in aggregate daily data each time step corresponds to 1 day.

For simplicity, the resulting reconstruction method for aggregate daily flow data is referred to as the Forecast Method (with Weekly Seasonal TBATS model).

The aim is then to evaluate and compare the prediction accuracy of both reconstruction methods for aggregate daily data. This comparative analysis is addressed in Section 3.5 as Test 4.

2.3.3 Level 2: Daily flow distribution patterns

The daily flow distribution patterns refer to the typical daily water consumption patterns, depending on the month and day of the week. In the scope of one year of data, there are 7 different days of the week in each of the 12 months, and therefore $7 \times 12 = 84$ patterns are determined as follows. For each month, the days are grouped by day of the week. In each of the resulting groups, a typical pattern for that day of the week is constructed by applying a measure to bind every corresponding time step, producing an estimate for each of the 96 values of the pattern.

In the original JQ approach, the measure used to estimate the typical value at each time step is the mean [10]. In the present study, we will determine whether replacing the mean with a robust measure, such as the median, results in more accurate predictions when the flow data contains anomalous data. For simplicity, the flow data reconstruction methods are referred to as the JQ Method (with mean), the JQ Method (with median), respectively. This test is addressed in Section 3.6 as Test 5.

2.3.4 15-min flow model

The original JQ approach was applied to regular flow data measured in 10-min intervals, which resulted in daily distribution patterns with 144 values [10]. However, the flow data provided for the present study is measured in 15-min intervals, resulting in patterns with 96 values. The original flow model was then adapted to the resulting 15-min model, which takes into account the daily/month variation and is formalised as follows:

$$y_{p15}(k+i) = \frac{y_{pat}(k,i)}{\sum_{j=1}^{96} y_{pat}(k,j)} y_p(k), \quad i = 1, ..., 96 \tag{12}$$

where $y_p(k)$ is the predicted flow for day k as described in equation (11), and $y_{pat}(k,i)$ is the prediction provided by the 15-min flow pattern considering the flow pattern class day of week/month of the actual day k.

Parameter estimation The JQ Method and the proposed modifications were implemented in statistical software package R [11] according to the formulations in this Section. The parameters for the aggregate ARIMA-based model were estimated for each DMA case study using the Least Squares Method and data with no missing values, as indicated in the original approach in [10].

2.4. 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 [14]. The use of a set of independent data (test set), but with the same distribution of the training set, avoids overfitting and allows to obtain 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.

In the TBATS approach, 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). In the JQ approach, the daily distribution patterns are constructed from flow data enclosed in each month. Therefore, considering that the test set is contained in a given month, the training set corresponds to the flow data that is not assigned to the test set in that month.

The accuracy of the predictions is determined by the following performance measures: the root-mean-square error (RMSE), a normalised 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 [14]. As scale invariant measures, the NRMSE and the MASE are used when the test requires comparison between predictions on flow data from several 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

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}} \tag{13}$$

and

$$\text{NRMSE} = \sqrt{\frac{\sum_{t=1}^{n} \left(\frac{\hat{y}_t - \hat{\mu}}{\hat{\sigma}} - \frac{y_t - \mu}{\sigma} \right)^2}{n}} \tag{14}$$

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,

$$\text{MASE} = \frac{\sum_{t=1}^{n} |y_t - \hat{y}_t|}{\frac{n}{n-1} \sum_{t=2}^{n} |y_t - y_{t-1}|} \tag{15}$$

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.

3. Results

In this Section we will present the results from each test, beginning with an exploratory analysis of the data provided for this study.

3.1. Exploratory analysis

The data provided consist of three flow time series belonging to different DMAs. The flow data correspond to the year 2013 and were measured in m^3/h, with regular 15-min time steps.

As a full-view example, we present the aggregate daily medians for DMA 3 (see Figure 1).

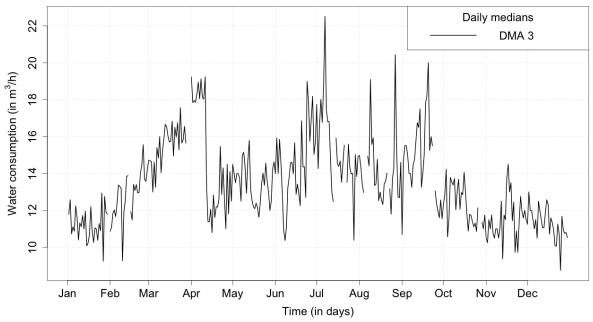

Figure 1: Flow data from DMA 3 (2013).

Furthermore, by grouping each of the 96 intra-day values of every day of the year, it is possible to illustrate the overall daily pattern present in the flow data (see Figure 2).

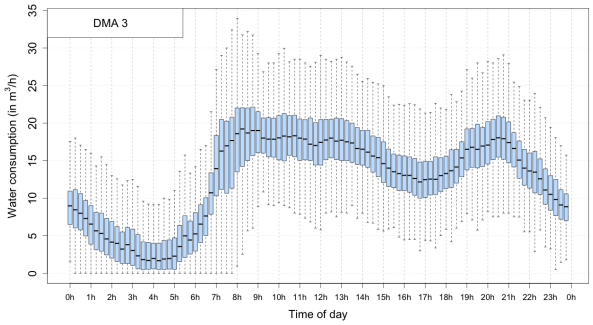

Figure 2: Grouped intra-day values of DMA 3.

Moreover, by aggregating these intra-day values by day of the week and then calculating the corresponding medians, it is possible to depict the typical daily pattern by day of the week (see Figure 3). The patterns corresponding to Saturday and Sunday clearly represent consumption habits that are different from the consumption habits on workdays. The five patterns corresponding to the workdays nearly coincide throughout the day, indicating very similar consumption habits. Furthermore, the patterns corresponding to Saturday and Sunday are also dissimilar from each other.

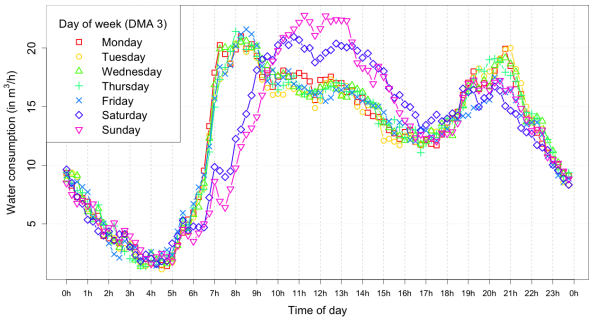

Figure 3: Patterns of days of the week of DMA 3.

3.2. Test 1: Impact of seasonal effects on model forecasts

We compared the RMSE of the forecasts generated by a classic ARIMA model, a Daily (D) Seasonal TBATS model, and a Daily/Weekly (D/W) Seasonal TBATS model (see Table 1). The ARIMA model selected was an $\text{ARIMA}(2,0,0)(0,0,1)_{96}$ model. The models were all fitted on an window size of 3 weeks beginning in November 2013 of DMA 2, and the forecast window was set to 1 week.

Table 1: Performance of classic ARIMA, TBATS (Daily), and TBATS (Daily/Weekly) models.

	ARIMA	TBATS (D)	TBATS (D/W)
RMSE	3.99	1.81	1.61
RMSE % increase	-	-55%	-60%

The results in Table 1 indicate that the ARIMA model is unsuitable to this type of data, especially in comparison to the TBATS models, as there is a decrease in RMSE of 55% when using the TBATS model with daily seasonality, and a decrease of 60% when using the TBATS model with both seasonalities, when compared to the ARIMA model. Therefore, there is evidence that a model incorporating both seasonalities is better suited to the flow data.

3.3. Test 2: Impact of window size for TBATS (D/W) 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 (and the Backcast Method). The results obtained from Test 1 (see Section 3.2) indicated that the Daily/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/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.

The overall means of the scale invariant measures (NRMSE and MASE) were then determined for each of the 4 predictions (see Table 2).

Table 2: Overall NRMSE and MASE of model forecasts, by model fitting window size.

	1 week	2 weeks	3 weeks	4 weeks
NRMSE	0.48	0.46	0.46	0.45
MASE	2.53	1.84	1.66	1.70

The overall NRMSE indicates that the model fitted on 1 week of flow data generally produces the least accurate predictions, and there is a slight progressive decrease in the overall NRMSE as the window size increases. The overall MASE reflects this as well, although the MASE increases again when the window size reaches 4 weeks. Therefore, there is evidence that in order to generate the most accurate predictions, the Daily/Weekly Seasonal TBATS model should be fitted on flow data with window size greater than 1 week, and preferably equal to 3 weeks. However, the overall means of the NRMSE are very similar, which indicates that the accuracy of the predictions is not greatly influenced by the window size for model fitting.

3.4. Test 3: Comparative analysis between methods in TBATS approach

This test aims at studying the suitability of the Daily/Weekly Seasonal TBATS model in the Forecast Method, the Backcast Method, and the Combined Method, and was conducted as follows for each DMA. A test set was fixed and assigned a window size of 1 week. The window size for fitting the model was selected according to the results obtained from Test 2 (Section 3.3), which indicated a window size of 3 weeks, for both Forecast Method and Backcast Method. The two reconstruction methods were applied, and a third set of predictions was then determined by applying the Combined Method.

The performance measures were determined for each reconstruction method of the TBATS approach and for each DMA (see Table 3).

From Table 3 we gather that the Combined Method effectively generates a set of predictions that is often more accurate than a simple Forecast Method or Backcast Method. For DMA 1 in particular, all prediction errors decreased to their lowest values when conducting the Combined Method. For DMA 3, the NRMSE for the Combined Method is the lowest from all methods as well.

Therefore, the Combined Method is considered a successful improvement by ultimately generating either the lowest or second-lowest error in every case.

Table 3: Performance measures of reconstruction methods of TBATS approach.

		Forecast	Backcast	Combined
DMA 1	RMSE	7.64	7.70	7.39
	NRMSE	0.39	0.39	0.37
	MASE	1.78	1.75	1.71
DMA 2	RMSE	1.60	1.77	1.72
	NRMSE	0.40	0.45	0.44
	MASE	1.79	1.93	1.85
DMA 3	RMSE	3.81	4.69	3.97
	NRMSE	0.58	0.58	0.57
	MASE	1.46	1.94	1.57

3.5. Test 4: Selection of aggregate daily flow model for the JQ approach

The aim of this test is to evaluate and compare the original ARIMA-based aggregate daily flow model from [10], and the proposed Weekly Seasonal TBATS model for aggregate daily data reconstruction. Both models are described in Section 2.3.2.

The original ARIMA-based model was implemented in statistical software R [11] for this study, in order to ultimately compare its prediction accuracy with other aggregate daily flow models. In order to reconstruct the aggregate time series data, the Forecast Method is applied with the selected aggregate data model.

The test was conducted as follows. First, the aggregate daily time series were determined for each DMA. A test set was fixed and assigned a window size of 8 days. Then, both models were applied on the aggregate data preceding the test set, and both forecasts were generated. In Table 4 we present the performance measures regarding both sets of forecasts for each DMA.

Table 4: Performance measures for the aggregate data reconstruction methods.

		ARIMA-based	Weekly TBATS
DMA 1	RMSE	75.63	52.88
	NRMSE	0.96	0.53
	MASE	1.02	0.70
DMA 2	RMSE	29.16	11.91
	NRMSE	1.37	0.63
	MASE	1.16	0.43
DMA 3	RMSE	1149.50	24.95
	NRMSE	1.33	1.01
	MASE	31.43	1.45

It was found that, in every case, the Weekly Seasonal TBATS model for aggregate daily flow data

generated the most accurate forecasts. By employing the TBATS model, the RMSE was successfully decreased by 30% in DMA 1, by 59% in DMA 2, and by 98% in DMA 3.

3.6. Test 5: Assessment of robustness in TBATS and JQ approaches

The aim of this test is to assess the impact of anomalous events in the accuracy of predictions generated by the TBATS approach and the JQ approach. The reconstruction methods analysed for the TBATS approach were the Forecast Method, the Backcast Method, and the Combined Method, with the Daily/Weekly Seasonal TBATS model. The reconstruction methods analysed for the JQ approach included the aggregate daily flow model selected according to the results obtained in Test 4 (see Section 3.5) for the first level (the Weekly Seasonal TBATS model), and for the second level the measure for constructing the daily distribution patterns was the mean (JQ Method with mean), and the median (JQ Method with median).

In order to assess model robustness, an artificial anomalous event was introduced in sections of data used for training, and the impact in terms of prediction accuracy is subsequently analysed. In the context of network flow data, several types of anomalous events exist. For this study, a reported pipe burst event with duration of 4 hours is simulated in the following way. A section of flow data corresponding to a time window of 4 hours is selected, and multiplied by a factor of 2.5.

The data reconstruction methods are applied to the flow data before and after the placement of the anomalous event. By calculating the prediction accuracy before and after, it is possible to assess the influence of this particular event. In the following tests, the test set corresponds to a section of flow data with window size of 1 week (in November 2013, from DMA 2). An artificial anomalous event was placed nearby, at a distance of four days following the test set.

Robustness of JQ approach We determined the RMSE of the resulting four sets of predictions (see Table 5).

Table 5: RMSE of JQ Methods, before and after placement of anomalous event.

	Before	After	% increase
JQ with mean	1.26	1.60	27%
JQ with median	1.26	1.33	6%

We observe that before adding the anomalous event to the flow data, both JQ Methods generated predictions with the same error. After placing the event, the original method (JQ Method with mean) generated predictions with a 27% error increase. On the other hand, the proposed adaptation to the JQ Method (JQ with median) withstood only a 6% error increase.

Therefore, there is evidence that the proposed adaptation of the JQ Method is more robust than the original procedure, as the influence of the anomalous event is attenuated by the use of a robust measure (the median). As such, the adapted JQ Method (with median) is preferable, and it will be the method used in subsequent comparison tests, instead of the original JQ Method (with mean).

Although these tests should be further explored with different types of anomalous events, the results indicate that the influence of these events on the adapted JQ Method to reconstruct network flow data is reduced. Therefore, it is not necessary to detect and remove outliers before data reconstruction, which simplifies remarkably the data processing approach.

Robustness of TBATS approach In this case, it is useful to remember that the predictions of the Forecast Method correspond to the forecasts generated by a Daily/Weekly Seasonal model, which is fitted on a window size of 3 weeks, and that the anomalous event was not placed therein. Therefore, the predictions of the Forecast Method will not be affected by the placement of the anomalous event.

By calculating the RMSE of the resulting six sets of predictions (see Table 6), we confirm that the RMSE for the Forecast Method remains unchanged. Furthermore, the anomalous event caused a 43% increase in RMSE for the Backcast Method, which indicates that the TBATS model is weak in terms of robustness.

The RMSE for the Combined Method, however, increased by only 30%. By combining the predictions generated by the Forecast Method and the Backcast Method, the Combined Method successfully attenuated the effect of the anomalous event, even producing the lowest RMSE of the methods, after placement of the event.

Table 6: RMSE of TBATS Methods, before and after placement of anomalous event.

	Before	After	% increase
Forecast	2.28	2.28	0%
Backcast	1.67	2.39	43%
Combined	1.75	2.27	30%

In conclusion, since anomalous events could occur in flow data that could be used for training either the Forecast Method or the Backcast Method, it is

not recommended to always choose one method over the other in the context of an automated flow data reconstruction procedure. There is evidence that the Combined Method reduces the overall RMSE, thus increasing the robustness of the procedure.

3.7. Test 6: Comparative analysis of the TBATS and JQ approaches

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 comparative analysis, we perform two kinds of short-term reconstruction: day-long (reconstruction of flow data with test set window size equal to 1 day) as well as week-long (reconstruction of flow data with test set window size equal to 1 week, or 7 consecutive days). Both tests are performed on each DMA.

The aim of day-long reconstruction tests is also to verify whether the reconstruction methods generate accurate predictions regardless of the day of the week. Therefore, the test is conducted as follows. For each day of the week, a section of flow data corresponding to that day is randomly selected as the test set. Then, the selected methods from the TBATS and JQ approaches are applied, thus generating two sets of predictions for each day of the week. The performance measures are determined, and the process is repeated for each DMA.

For week-long reconstruction, a random week is selected as the test set. Then, the selected methods from the TBATS and JQ approaches are applied, thus generating two sets of predictions for each day of the week. The performance measures are determined, and the process is repeated for each DMA.

For this test, we used two flow data reconstruction methods. The first method corresponds to the more suitable method of the TBATS approach, as determined by the results of Test 1 (see Section 3.2), Test 2 (see Section 3.3), Test 3 (see Section 3.4), and Test 5 (see Section 3.6): the Combined Method (with Daily/Weekly Seasonal TBATS model).

The second method corresponds to the more suitable method of the JQ approach, as determined by the results of Test 4 (see Section 3.5), and Test 5 (see Section 3.6): the JQ Method (with aggregate Weekly Seasonal TBATS model and median).

Day-long reconstruction By calculating the mean NRMSE and mean MASE of the predictions for each day of the week and for all DMAs (see Figure 4), we observe that on the weekends both NRMSE and MASE are much higher for the TBATS approach than for the JQ approach.

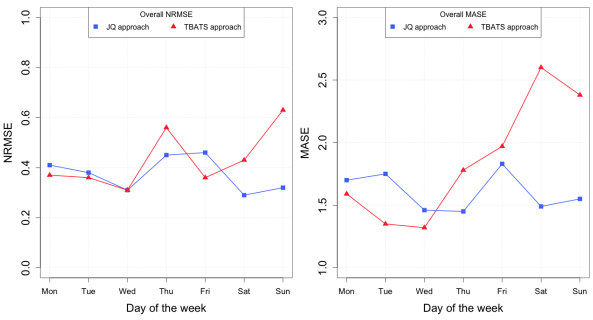

Figure 4: Overall NRMSE and MASE of predictions for each day of the week.

On the other hand, the TBATS approach generated relatively better predictions overall for Mondays, Tuesdays and Wednesdays. However, the JQ approach ultimately generates results that are more consistent throughout the week.

Week-long reconstruction The performance measures resulting from the application of the JQ approach and the TBATS approach on the selected week (for each DMA) are presented in Table 7.

Table 7: Performance measures of JQ and TBATS approaches for the selected week.

		JQ approach	TBATS approach
DMA 1	RMSE	4.94	7.75
	NRMSE	0.22	0.34
	MASE	1.26	1.94
DMA 2	RMSE	1.21	1.72
	NRMSE	0.30	0.44
	MASE	1.38	1.85
DMA 3	RMSE	3.77	3.79
	NRMSE	0.56	0.61
	MASE	1.49	1.45

From Table 7 we observe that the JQ approach generated ultimately better predictions in each case, although in DMA 3 the errors were very similar. We conclude that for longer gaps in flow data, the JQ approach is preferable to the TBATS approach.

4. Conclusions

In this study we have implemented and modified an existing two-model reconstruction method that has been shown to generate good results in water flow data [10], as well as devised a new reconstruction method based on the multiple seasonality forecasting model TBATS [4].

The results obtained from comparative analysis indicated that a double seasonal TBATS model is

better suited to flow data than models that incorporate only the daily seasonality. We concluded that this model should be fitted on windows of flow data with size greater than 1 week. Furthermore, the proposed Combined Method of forecast and backcast predictions effectively generated good results, and was shown to reduce the influence of anomalous events on the predictions. This finding enables the application of a new approach for data reconstruction and is especially useful for off-line analysis.

Regarding the JQ approach, two modifications to the original two-level procedure were proposed. The proposed Weekly Seasonal TBATS aggregate model effectively generated predictions that were more accurate than the original aggregate model, in every case. Furthermore, the results from the robustness test indicated that the measure used to build the daily distribution patterns could be improved upon, by replacing the mean with the median.

The most suitable method in each approach was selected based on the tests: the Combined Method (with Daily/Weekly Seasonal TBATS model) for the TBATS approach, and the JQ Method (with Weekly Seasonal TBATS model and median) for the JQ approach. The results obtained from the comparative analysis indicated that the JQ approach generated predictions that are more accurate overall than the predictions of the TBATS approach.

Regarding future work, it would be interesting to employ these data reconstruction methods toward data validation and detection of anomalous events. Although the proposed modifications to the reconstruction methods effectively increased their robustness to anomalous events, it is likely that the prediction accuracy would improve by performing a data validation step beforehand.

Furthermore, besides being used for reconstructing historical flow data, the Forecasting Method can be applied to online flow data reconstruction, in which the model is fitted to the most recent flow data available.

Additionally, the methods described in this dissertation can also be applied to other kinds of time series data, such as electricity load, waste water, and energy data, which present similar properties to water consumption flow data.

References

[1] S. Alvisi, M. Franchini, and A. Marinelli. A short-term, pattern-based model for water-demand forecasting. *Journal of Hydroinformatics*, 9(1):39–50, 2007.

[2] J. Caiado. Performance of combined double seasonal univariate time series models for forecasting water demand. *Journal of Hydrologic Engineering*, 15(3):215–222, 2009.

[3] S. Coelho. A system for demand analysis and forecasting in water supply systems. Master's thesis, University of Newcastle Upon Tyne, Newcastle, UK, 1988.

[4] A. M. De Livera, R. J. Hyndman, and R. D. Snyder. Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496):1513–1527, 2011.

[5] G. De Marinis, R. Gargano, and C. Tricarico. Water demand models for a small number of users. In *8th Annual Water Distribution Systems Analysis Symposium*, 2006.

[6] G. Dudek. Forecasting time series with multiple seasonal cycles using neural networks with local learning. In *Artificial Intelligence and Soft Computing*, pages 52–63. Springer, 2013.

[7] S. N. Hassan, M. H. Ahmad, and N. Mohamed. A comparison of the forecast performance of double seasonal arima and double seasonal arfima models of electricity load demand. *Applied Mathematical Sciences*, 6(135):6705–6712, 2012.

[8] A. Z. Mamade. Profiling consumption patterns using extensive measurements. Master's thesis, Universidade Técnica de Lisboa, Lisbon, Portugal, 2013.

[9] N. Mohamed, M. H. Ahmad, Z. Ismail, et al. Double seasonal arima model for forecasting load demand. *Matematika*, 26:217–231, 2010.

[10] J. Quevedo, V. Puig, G. Cembrano, J. Blanch, J. Aguilar, D. Saporta, G. Benito, M. Hedo, and A. Molina. Validation and reconstruction of flow meter data in the barcelona water distribution network. *Control Engineering Practice*, 18(6):640–651, 2010.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[12] J. W. Taylor and R. D. Snyder. Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. *Omega*, 40(6):748–757, 2012.

[13] W. W.-S. Wei. *Time series analysis*. Addison-Wesley publ, 1994.

[14] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.