Comparison of reconstruction methods for water supply systems flow rate time series

Carlos Ascensão [1[0000-0002-3227-890X]], Bruno Ferreira [1[0000-0002-2863-7949]], Raquel Barreira [1[0000-0002-8326-1593]], Nelson Carriço [1[0000-0002-2474-7665]]

[1] INCITE, Barreiro School of Technology, Polytechnic Institute of Setúbal, Setúbal, Portugal
carlosfpascensao@gmail.com

Abstract. The purpose of this paper is to compare the performance of five univariate models for the reconstruction of flow rate time series. Errors in the measurements may occur due to problems in the sensor or in the communication system with data logger, thus generating missing data in the flow rate time series. The presence of missing values in flow rate data restricts its use in network operation processes. The performance of seasonal ARIMA, Standard and double seasonality Holt-Winters, and original and improved Quevedo approach are assessed. The analysis is made considering a real Portuguese case study and 1-month of flow rate data at 1-hour and 10-minute period. The holidays compared to the weekdays show great differences in consumption patterns. For this reason, the effect of forecasting holidays is assessed. Obtained results evidence that the improved Quevedo model can cope with different time step intervals and type of day being forecasted, with a reduced computation time.

Keywords: Flow rate, forecasting, reconstruction methods, time series, water supply systems.

1 Introduction

Flow rate monitoring is an increasingly recurring practice in water utilities, due to the larger accessibility and availability of telemetry equipment and remote management systems. The stored time series data can be used for many tasks in operating and monitoring systems (e.g., demand forecasting, burst detection). The measured data are acquired by sensors and stored in the data logger, which communicates remotely to the management system [1]. Errors in the measurements can happen and may be caused by problems in the sensor or in the communication system with the data logger due to power failures, storage limitations or working outside the operational range generate the missing data in time series [2].

The treatment of flow rate time series is a challenging task for water utilities. The validation processes are based on simple heuristics. Usually non-validated data is replaced using reconstruction methods by predicting the measures with multivariate or univariate statistical models for flow rate time series [3]. More advanced techniques

such as machine learning may be applied to forecast water demand in water supply systems which consider air temperature, precipitation, and flow rate. However, multivariate models that require many variables represent a greater challenge in their application and data acquisition, thus making the operationalization of the models a hard task. Also, the application of advanced techniques, in terms of system monitoring, requires real-time operation and multivariate models do not provide good results when applied in real-time [1]. For these reasons water utilities search for simple forecasting models with low difficulty and complexity in required data, model application and operationalization.

Flow rate time series may show great evidence of daily and weekly cycles that must be considered by forecasting models [4]. Literature shows that autoregressive models and exponential smoothing models with components and parameters that express seasonality are able to obtain reasonable results [1, 9]. Caiado [6] assessed the performance of three different univariate models for water demand forecasting, namely, Holt-Winters, ARIMA and GARCH model and results suggest that all the univariate time series models can be quite useful for short-term forecasting. Quevedo *et al*. [2] developed a short-term forecasting methodology with the purpose of reconstructing missing data in water supply systems. This methodology considers an aggregate daily flow model based on ARIMA models and a 10-minute model based on distributing the daily flow using a 10-minute demand pattern. Cugueró-Escofet *et al*. [7] applied a methodology for reconstruction of missing flow rate data using a double seasonal Holt-Winters.

This paper compares the performance of five forecasting models, namely, a seasonal ARIMA, a seasonal and double seasonal Holt-Winters, a method proposed in Quevedo *et al.* [2] and our improvement of Quevedo approach. The performance assessment is carried out for a complete forecasted day using the root-mean-squared-error (RMSE) and was applied to a real Portuguese case study.

2 Reconstruction methods

2.1 Seasonal ARIMA

The ARIMA models are derived from the family of Auto Regressive Moving Average (ARMA) models. Their difference is the integration component that allows differentiating the series to be possible to apply to non-stationary time series. In order to forecast, ARIMA models use a polynomial of the previous values together with the previous prediction errors. Seasonal ARIMA models considers an additional polynomial for the seasonal component [1].

The function of the ARIMA models can be represented by the degrees of the model (p,d,q), where *p* represents the number of autoregressive terms, *d* represents the number of differentiations and *q* the number of lagged forecast errors in the prediction equation. The polynomial function dedicated to the seasonal component works only with a periodicity. Similar to the polynomial of the regular component it can also be represented by (P,D,Q)s where P, D and Q represent the degrees of the model and *s* represents the number of seasonal periods.

The degrees of the seasonal ARIMA model (p,d,q)(P,D,Q)s can be selected based on the Bayesian information criterion and assessing the fitted values. For the current study, and since 1-hour and 10-minute time intervals are considered, we used the seasonal ARIMA parameters $(2,0,0)(2,0,0)_{24}$ and $(1,0,0)(1,0,1)_{144}$, respectively.

2.2 Quevedo approach

The present approach shows the implementation and improvements to the model for the reconstruction of time series data from water supply systems presented in Quevedo *et al.* [8].

The procedure for reconstructing missing data consists of two modules. The first module gives the prediction of aggregated daily flow based on the seasonal ARIMA models, denominated as the aggregated daily flow model. This model requires a historic record of daily aggregate flow to be able to predict the daily volume taking advantage of the main components of the ARIMA model. For the selection of the polynomial degrees (p,d,q)(P,D,Q)s it was based on the Bayesian information criterion [9] evaluating the set of models generated by $0\leq p\leq 3$, $0\leq P\leq 1$, $0\leq q\leq 3$ e $0\leq Q\leq 1$.

The second module determines a set of flow distribution patterns at 10-minute intervals, consisting of 144 average flow rate values for each pattern. Distribution patterns consider the variation in measurements between weekdays and weekends. For this reason, patterns must be determined for the days of the week (Monday to Friday), and of the weekend (Saturdays and Sundays). Holidays should also be considered for the impact they have on the analysis. Quevedo *et al.* [8] determined that consumption habits for holidays are the same as on Sundays. However, by considering ARIMA as the aggregate daily flow forecast model it is not possible to take into account the effect of a holiday during a weekday.

Improvements to the Quevedo approach are proposed to estimate the daily aggregate flow of a holiday. When initializing the model, it is verified that the date to be forecasted is within the subset with holiday date. If the date does not coincide, the model runs according to the initial approach and estimates the aggregate daily flow with the seasonal ARIMA model. If the date coincides, the aggregate daily flow estimation starts with a simple exponential smoothing model, for which the input is a subset with the aggregate daily flow for Sundays. Estimation for the aggregate daily flow of a holiday is carried as a new Sunday.

To reconstruct flow rate time series, Quevedo *et al.* [8] distributes the aggregate daily flow estimate by the distribution pattern of the day to be reconstructed.

2.3 Exponential smoothing

Holt Winters method is an exponential smoothing method considering trend and seasonal components [10]. In the urban water sector, exponential smoothing methods are well known and have been used in automatic forecasting models [1]. The main characteristic is its simplicity, considering it can be optimized only with the least squares.

Holt-Winters method is based on level, trend and seasonality [11] and models can be divided into two versions based on seasonality patterns, namely, additive or

multiplicative seasonality. Depending on the type of seasonal pattern presented in the data, one of the reference versions can be chosen [12]. In additive seasonality, the difference in seasonal fluctuation between successive is constant, while in multiplicative seasonality the variation is a percentage [12]. In this article, only the Holt Winters models with multiplicative seasonality are considered.

Standard Holt-Winters

Initial values of the components (i.e., level, trend and a seasonal index) are required to start the Standard Holt-Winters multiplicative seasonality model. According to [10] initial level is obtained by averaging the observations from the first seasonal period. The estimated initial trend uses a moving average of the first seasonal period and the seasonal indices are estimated using the average of first seasonality period.

The model components are based on three smoothing parameters: α, β and δ. These parameters represent the level parameter, trend parameter, the seasonal parameter for the seasonal cycle (daily in our case), respectively. These parameters can be estimated by minimizing the RMSE and are usually restricted to lie between 0 and 1.

Double seasonal Holt-Winters

The double seasonal Holt-Winters accommodates two seasonal periods, in this case daily and weekly. To consider the effect of weekly seasonality, the double seasonal Holt-Winters model requires one more seasonality component than the standard model.

The double model requires initial values for initializing – trend, level, daily seasonality, and weekly seasonality index. Taylor [13] formulation was used to represent the initial values, using a s1-period cycle for the daily seasonality and s2-period cycle for weekly seasonality. According to Taylor [13] the initial trend, was chosen as the average of (1) 1/s2 of the difference between the mean of the first s2 and second s2 observations and (2) the average of the first differences for the first s2 observations. The initial level was chosen as the mean of the first two s2 observations minus s2 and half times the initial trend. The initial values for the daily seasonal index are defined by the average of the ratios of actual observation to s1-point centered moving average, taken from the corresponding half-hour period in each of the first 7 days of the time series. The initial values for the weekly seasonal index were set as the average of the ratios of actual observation to s2-point centered moving average, taken from the corresponding half-hour period on the same day of the week in each of the first 2 weeks of the demand series, divided by the initial value of the smoothed within-day seasonal index.

The model components are based on four smoothing parameters α, β, Υ and δ. The first three parameters are similar to the ones previously presented for standard Holt Winters. In addition, a seasonal parameter for bigger seasonal cycle (weekly in our case) is considered. Similarly, these parameters can be estimated by minimizing the RMSE and are usually restricted to lie between 0 and 1.

2.4 Performance assessments

A complete day of missing data (i.e., long gap duration) was assumed to assess the performance of the presented models. For this reason, models will be trained with historical data with a duration no smaller than a month. The last day will be forecasted using the trained models to assess its performance.

The different forecasting models' parameters were calibrated by minimizing the RMSE of the fitted model. The performance of the models' predictions was assessed using the RMSE between real and predicted measurement.

3 Applications and discussion

In this section, an analysis to the performance of the five flow rate time series reconstruction models is carried.

The flow rate time series of the case study is collected using an impulse flow rate meter installed at the inlet of the water distribution network of a residential area with up to 3,000 inhabitants and was provided by a water utility located in Lisbon metropolitan area. This case study considers a 1-month of historical flow rate data recorded in intervals of 1-hour and 10-minute. As such, the model with 1-hour intervals will predict 24-steps and the model with 10-minutes intervals the model will predict 144-steps.

Initially, forecasts were carried considering a weekday and 1-hour intervals. Figure 1 presents the obtained results for each model as well as the real flow rate data.

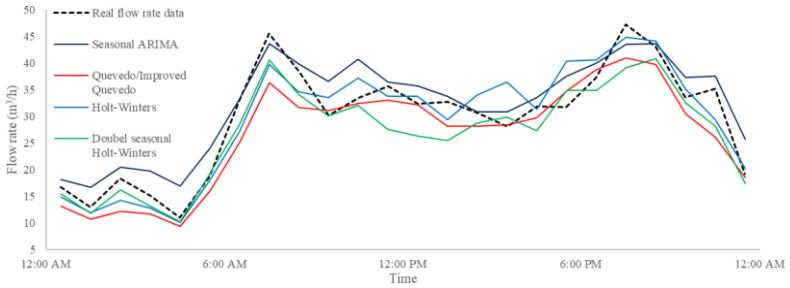

Fig. 1. Comparison for five forecasting techniques considering weekday with hourly intervals.

All models performed reasonably well predicting a day in the series at 1-hour intervals for all models. The seasonal ARIMA model presented the best values (RMSE=3,78), followed by the Holt-Winters (RMSE=3,89), double seasonal Holt-Winters (RMSE=4,09), and lastly the Quevedo (RMSE=4.41). The computation time of each model was assessed, and it was concluded that all models were relatively fast (i.e., less than 40 seconds). Table 1 shows a summary of computational time (in seconds) and the RMSE obtained for all models.

Time series with 1-hour intervals have a very limited use in real-time water supply systems operation. Application of advanced techniques, such as machine learning requires time series with shorter time intervals to operate water supply system in real-

time. As such, the same day of the week is predicted with the series of 10-minute intervals. The obtained results for each model are presented in Figure 2 and Table 1.

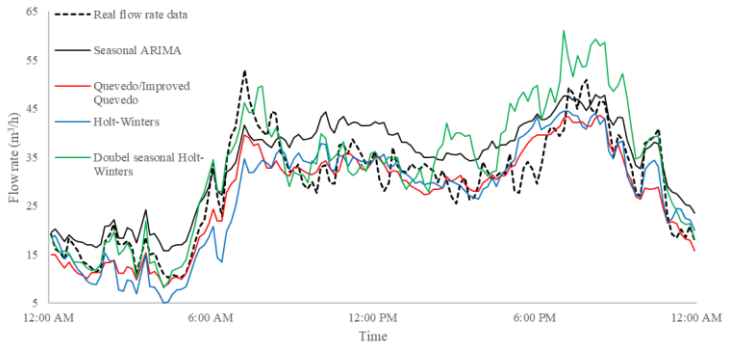

Fig. 2. Comparison for five forecasting techniques considering weekday (10-minute intervals).

All models predicted reasonably well, with the Quevedo presenting the best overall results (RMSE=4,69), followed by Holt-Winters (RMSE=5,97), seasonal ARIMA (RMSE=6,10) and lastly double seasonal Holt-Winters (RMSE=6.76). Nonetheless, the difference in computation time amongst methods is quite significant, with the Quevedo (t=1s), Holt-Winters (t=108s), Double seasonal Holt-Winters (t=286s) and lastly seasonal ARIMA (t=591s).

Problems may arise when forecasting holidays during weekdays since the expected behavior of the water demand of a holiday is usually related to the behavior of the Sundays (in opposition to the behavior of weekday) [14]. As such, a holiday during a weekday is predicted with the series of 10-minute intervals. In addition to the four methods already compared, a fifth is considered with the Quevedo improvement. The obtained results for each model are presented in Figure 3 and Table 1.

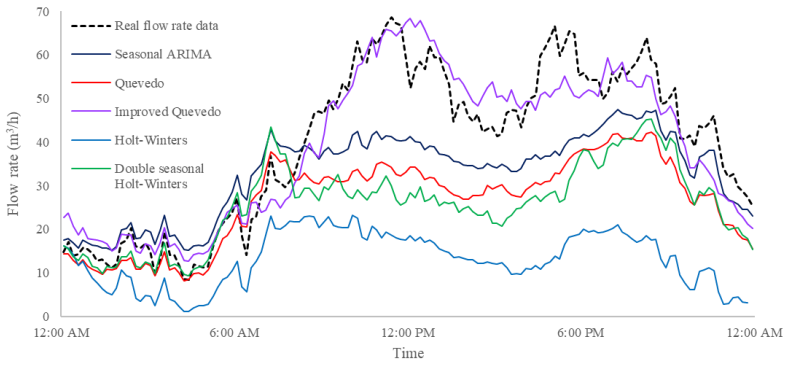

Fig. 3. Comparison for five forecasting techniques considering holiday with 10-minute intervals.

Based on Figure 3 and Table 1 it is possible to conclude that, overall, the improved Quevedo performed reasonably well (RMSE=5,84). The remaining models failed to capture the holiday variations. The autoregressive and exponential smoothing models make predictions based on seasonality periods and cannot "look" outside the seasonality periods. Similarly, the original Quevedo approach forecast the aggregate daily flow using an autoregressive model (i.e., seasonal ARIMA).

Table 1. Comparison of model's performance and computational time

Models	Weekday (1-hour)		Weekday (10-minute)		Holiday (10-minute)	
	RMSE	Computation time (s)	RMSE	Computation time (s)	RMSE	Computation time (s)
Seasonal ARIMA	3.78	12	6.10	591	12.50	624
Quevedo	4.41	1	4.69	1	16.91	1
Holt-Winters	3.89	38	5.97	108	30.53	101
Double seasonal Holt-Winters	4.09	12	6.76	286	19.21	284
Improved Quevedo	4.41	1	4.69	1	5.84	1

Conclusions

Missing data from flow rate time series resulting from a validation process must be reconstructed to apply advanced techniques that requires validated data. Usually, the reconstruction of the flow rate time series is performed by forecasting models. This paper presents a comparison of five methods to the reconstruction of flow rate time series, namely, the seasonal ARIMA, the standard and double seasonal Holt-Winters, the original and improved Quevedo approach. The comparison is based on 1-month of historical flow rate time series at 1-hour and 10-minute intervals of a real Portuguese case study. A complete day was forecasted and analyzed for weekdays and holidays. In weekdays, forecasts with 1-hour intervals obtained a reasonable RMSE result for all methods. Similarly, reasonable RMSE results were obtained for all models considering weekdays with the 10-minute intervals. Nonetheless, great difference in computation time were observed amongst methods. On the other hand, and when forecasting a holiday, only the improved Quevedo approach produced reliable results.

Future research may include the forecast of flow rate data when in absence of reliable historical data, for instance, due to changes in patterns motived by recent lockdowns.

Acknowledgement

The authors want to acknowledge Fundação para a Ciência e a Tecnologia, (grant number DSAIPA/DS/0089/2018) through the Data Science and Artificial Intelligence in Public Administration Programme for supporting WISDom project. The authors also acknowledge the participating water utilities for their contribution.

References

1. Puig V., Ocampo-Martínez C., Pérez R., Cembrano G., Quevedo J., and Escobet T.: *Real-time Monitoring and Operational Control of Drinking-Water Systems*. Springer International Publishing, Cham (2017).
2. Quevedo J., Puig V., Cembrano G., Blanch J., Aguilar J., Saporta D., Benito G., Hedo M. and Molina A.: Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract. 6(18)*, 640–651 (2010).
3. Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A.: Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* 48, 1–14 (2012).
4. De Marinis, G., Gargano, R. & Tricarico, C.: Water demand models for a small number of users. In: *8th Annu. Water Distrib. Syst. Anal. Symp* (2007)
5. Taylor, J. W., de Menezes, L. M. & McSharry, P. E.: A comparison of univariate methods for forecasting electricity demand up to a day ahead. *Int. J. Forecast.* 22, 1–16 (2006).
6. Caiado, J.: Performance of combined double seasonal univariate time series models for forecasting water demand. *J. Hydrol. Eng.* 15, 215–222 (2010).
7. Cugueró-Escofet M., García D., Quevedo J., Puig V., Espin S., and Roquet J.: A methodology and a software tool for sensor data validation/reconstruction: Application to the Catalonia regional water network, *Control Eng. Pract.*, vol. 49, pp. 159–172, (2016)
8. Quevedo J., Puig V., Cembrano G., Blanch J., Aguilar J., Saporta D., Benito G., Hedo M., Molina A.: Validation and reconstruction of flow meter data in the Barcelona water distribution network. *Control Eng. Pract.* 18, 640–651 (2010).
9. Schwarz, G.: Estimating the Dimension of a Model. *Ann. Stat.* 6, 461–464 (1978).
10. Spyros, M., Steven C, W. & Rob J, H.: *Forecasting: Methods and Applications*. Wiley, (1997).
11. Hyndman, R. J., Koehler, A. B., Ord, J. K. & Snyder, R. D. *Springer Series in Statistics Forecasting with Exponential Smoothing*. (2008).
12. Galvas G.,: Time series forecasting used for real-time anomaly detection on websites. Vrije Universiteit, Amsterdam (2016).
13. Taylor, J. W.: Short-term electricity demand forecasting using double seasonal exponential smoothing. *J. Oper. Res. Soc.* 54, 799–805 (2003).