A comparison between pattern-based and neural network short-term water demand forecasting models
F. Gagliardi, S. Alvisi, M. Franchini and M. Guidorzi

ABSTRACT

In this paper, two models are set up in order to forecast hourly water demands up to 24 h ahead and are contrasted with each other. The first model (hereinafter referred to as the Patt model) is based on the representation of the periodic patterns that typically characterize water demands, such as seasonal and weekly patterns of daily water demands and daily patterns of hourly water demands. The second model is based on artificial neural networks (hereinafter referred to as ANN models). Both the models have been applied to three case studies, representing water distribution systems managed by HERA S.p.A., characterized by very different numbers of users served, and consequently very different average water demands, ranging from 900 L/s for the first case study (CS1) to about 8 L/s and 1.5 L/s for the second (CS2) and third (CS3) case studies, respectively. The results show that in general, both the models, Patt and ANN, provide good accuracy for the CS1. The performances of both the models tend to decrease for CS2 and, particularly, for CS3. In particular, in the validation phase, the Patt model is more accurate than the ANN model for the CS1; for the CS2, the accuracy of the two models are very similar, and for the CS3 the accuracy of the ANN model is slightly higher than that of the Patt model.

Key words | forecast, neural network, pattern, water demand

INTRODUCTION

Short-term water demand forecasting is a useful tool for water distribution system management. In fact, an accurate prediction of water demand of a network, or a part of it, can support network devices scheduling or real time control, such as pumping stations or valves. In the last decades, several models have been proposed for short-term water demand forecasting (e.g. Maidment & Miaou 1986; Zhou et al. 2000, 2002; Alvisi et al. 2007; Adamowski 2008; Bakker et al. 2013; Arandia et al. 2016; see also Donkor et al. (2014) for a complete review of water demand forecasting models developed since 2000). These models typically provide the water demand forecast on a time horizon variable from a few hours ahead (e.g. Zhou et al. 2002; Alvisi et al. 2007; Romano & Kapelan 2014; Arandia et al. 2016) up to 1–7 days ahead (e.g. Zhou et al. 2000; Adamowski 2008) assuming a short-time step, variable from 15 minutes (e.g. Bakker et al. 2013; Arandia et al. 2016) up to 1 day (e.g. Zhou et al. 2000; Cutore et al. 2008; Ghiassi et al. 2008). Most of these models use as inputs only previously observed water demands (e.g. in the previous 24 h or in the previous week), whereas in some cases other exogenous variables such as climatic factors are also taken into account. It is worth noting that these exogenous factors, when referred to the forecast period, should be forecast as well. However, in most of the models developed in the past, the weather data referred to the forecast period are assumed as observed or perfectly forecast data. A few examples exist in which exogenous inputs referred to the future are forecast, such as Tian et al. (2015), which integrates a water demand forecasting
model (based on the auto-regressive integrated moving average, ARIMA technique) with weather data forecasting models (NWP).

When reference to the structure is made, a short-term forecasting model can be subdivided into two main classes, that are data-driven and pattern-based models. Data-driven models use different linear and non-linear data-driven techniques in order to provide short-term water demand forecasting, such as time series analysis, regression processes, artificial neural network (ANN), ARIMA and support vector machine (SVN). ANN, in particular, is proved to be one of the most efficient techniques in water demand forecasting compared to other techniques. Many different ANN-based models have been developed in the last decades, using different network structures and inputs, and compared to other techniques. Bougadis et al. (2005), for example, compared different ANN-based models with several time series and regression based models and discovered that ANN-based models outperform the others. Similar results have been found by Adamowski (2008), in which a comparison between 39 ANN models and several other data-driven models was conducted, and by Jain et al. (2001) where two ANN-based models, respectively containing one and two hidden layers, were compared with time series analysis and regression based models. ANNs have been coupled with different algorithms in order to calibrate the parameters of the network, such as the SCEN-UA algorithm in Cutore et al. (2008) and the Evolutionary Algorithm in Romano & Kapelan (2014).

The second class includes the so-called pattern-based models. These models are based on the observation that water demand time series are generally affected by periodicities that can be used to support the forecast. Maidment & Miaou (1986) developed a model which removes trends from water demand time series using statistical regression and also analyses the response of daily water demand to rainfall and air temperature variations. Zhou et al. (2000, 2002) identified, in water demands, a base component and a seasonal component developing models respectively applied to daily and hourly water demand data. Alvisi et al. (2007) developed a model capable of simulating periodicity affecting daily and hourly water demands, in order to accurately forecast hourly demand. In Caiado (2010), different models such as Holt-Winter, ARIMA and generalized auto-regressive conditional heteroskedasticity (GARCH) for pattern recognition are compared and also combined; it is then proved that short-term forecasting accuracy increases using combined models. Sub-hourly water demand forecasting models have been provided by Bakker et al. (2013) and Arandia et al. (2016). In the first work, a pattern-recognition model based on day factors and daily demand patterns for each day of the week and for different deviant day types is proposed and applied to forecast 15-min water demand. In the second work a model recognizing weekly and daily patterns is developed and applied to daily, hourly and sub-hourly (15-min) water demands.

The objective of this study is to compare the effectiveness of two short-term forecasting models belonging to the two classes previously mentioned (i.e. data-driven and pattern-based) when applied to very different real observed water demand time series. In more detail, the two models considered are the pattern-based forecasting model (Alvisi et al. 2007) (hereinafter referred to as the Patt model) and an ANN model (Alvisi & Franchini 2017). Both the models are set up in order to forecast hourly water demands in a 24 h horizon, using only historical system water demand data as inputs. The models are applied to three case studies referred to network and districts, managed by HERA S.p.A., characterized by very different numbers of users served. The paper is organized as follows. The applied models are described in the following section and then the case studies are presented. Afterwards, the results of the application of both the models to the case studies are analysed and finally the conclusions are drawn.

MODELS: PATT AND ANN

Patt model

The Patt model (Alvisi et al. 2007) is a short-term forecasting model based on the periodicities that typically characterize water demand time series observed at the level of water networks or districts. Indeed, daily water demands are usually characterized by seasonal and weekly patterns; hourly water demands are generally affected by a daily pattern. The seasonal pattern usually consists in an increasing demand due to the rising temperatures during summer and
autumn (i.e. from May to October). The weekly periodicity depends upon the type of households supplied by the network and the habits of the users; for example, for residential users the average daily values of working days can be different from those of non-working days. The daily pattern reflects the consumers’ habits in terms of water demand as well; in the case of residential users, the hourly demand generally reaches the highest values in the early morning and evening, whereas it is low during the night and variable during the day. The Patt model takes into account these periodic behaviours to forecast future water demands. Accordingly, the model should be used to forecast the demand at the level of networks or districts (i.e. medium/high spatial aggregation level), whereas it cannot be effectively used to forecast demands at household level since at such fine spatial level water demand, time series are less characterized by patterns.

As regards the structure of the model, it is divided into two modules, a daily module (DM) and a hourly module (HM), as shown in Figure 1. In both the modules, two components are taken into account: a periodic one based on the simulation of the patterns and a short-term persistence, which makes a correction based on the errors the model makes in the previous steps.

In the DM, a seasonal periodic component \( Q_{m,s}^{d} \), a weekly periodic correction \( \Delta_{i,j}^{d} \) and a short-term persistence correction \( \delta_{d}^{m} \) are taken into account in order to estimate the average daily water demand \( Q_{m}^{d,for} \) for the \( m \) Julian date:

\[
Q_{m}^{d,for} = Q_{m,s}^{d} + \Delta_{i,j}^{d} + \delta_{d}^{m}
\]  

(1)

The seasonal periodic component, \( Q_{m,s}^{d} \), is modelled using a Fourier series:

\[
Q_{m,s}^{d} = a_{0} + \sum_{f=1}^{f} \left[ a_{f} \cos \frac{2\pi f}{365} m + b_{f} \sin \frac{2\pi f}{365} m \right], \quad m = 1, 2, \ldots, 365
\]  

(2)

with \( a_{f} \) and \( b_{f} \) Fourier coefficients, \( a_{0} \) the mean value of the seasonal cycle and \( f \) is the number of harmonics considered. The Fourier coefficients are determined as follows:

\[
a_{f} = \frac{2}{365} \sum_{m=1}^{365} Q_{m}^{d,obs} \cos \frac{2\pi m f}{365}
\]  

(3)

\[
b_{f} = \frac{2}{365} \sum_{m=1}^{365} Q_{m}^{d,obs} \sin \frac{2\pi m f}{365}
\]  

(4)

where \( Q_{m}^{d,obs} \) are the observed average daily water demands. The weekly pattern is considered through the term \( \Delta_{i,j}^{d} \), which represents a correction factor taking into account that different days of the week are typically characterized by different average daily water demands, and is determined as:

\[
\Delta_{i,j}^{d} = Q_{i,j}^{d} - Q_{j}^{w}
\]  

(5)

where \( Q_{i,j}^{d} \) is the mean value of the average daily water demands observed on day \( i \) of the week (\( i = 1, \ldots, 7 \), Monday, …, Sunday), in season \( j \) (\( j = 1, \ldots, 4 \), winter, spring, summer, autumn) and \( Q_{j}^{w} \) is the mean value of the average weekly water demands in season \( j \). The persistence component \( \delta_{d}^{m} \) is modelled using an autoregressive process AR(1) (Box et al. 1994). This term represents the deviation between the mean observed daily water demand \( Q_{m}^{d} \) and the mean value calculated using only the periodic components \( Q_{m,s}^{d} \) and \( \Delta_{i,j}^{d} \). Thus, it is calculated as:

\[
\delta_{d}^{m} = \Phi_{1} \delta_{d}^{m-1}
\]  

(6)
where $\Phi_1$ is a parameter calibrated on the basis of the observed deviations:

$$e^{d,obs}_m = Q^{d,obs}_m - (Q^{d,s}_m + \Delta^d_{ij})$$  \hspace{1cm} (7)

The second module, named hourly module (HM), provides the forecast of the hourly water demands for $k$ hours ahead (with $k = 1, 2, \ldots, 24$). The forecast value is given by the sum of the forecast average daily water demand $Q^{d,for}_m$, a daily periodic correction $\Delta^h_{n,ij}$ and a persistence component $\epsilon_{t+k}$:

$$Q^{h,for}_{t+k} = Q^{d,for}_m + \Delta^h_{n,ij} + \epsilon_{t+k}$$  \hspace{1cm} (8)

The first term of the sum, $Q^{d,for}_m$, represents the output of the DM. The correction component based on the daily pattern is computed as:

$$\Delta^h_{n,ij} = \bar{Q}^h_{n,ij} - \bar{Q}^d_{n,ij}$$  \hspace{1cm} (9)

where $\bar{Q}^h_{n,ij}$ represents the mean value of the average hourly water demands observed in the $n$-th hour ($n = 1, \ldots, 24$, is the hour of the day) of the day $i$ in season $j$ and $\bar{Q}^d_{n,ij}$ is the mean value of the average daily water demands observed on day $i$ in season $j$. The short-term persistence $\epsilon_{t+k}$ is correlated to the deviations occurring 1 h and 24 h earlier so that it is modelled as:

$$\epsilon_{t+k} = \psi_{t+k-1}\epsilon_{t+k-1} + \psi_{t+k-24}\epsilon_{t+k-24}$$  \hspace{1cm} (10)

where $t$ is the forecasting instant, $k$ the lag time and $\psi_{t+k-1}$ and $\psi_{t+k-24}$ regression coefficients. This coefficients are variable depending on the hour of the day and are calibrated using observed errors $\epsilon^{obs}_t$:

$$\epsilon^{obs}_t = Q^{h,obs}_t - (Q^{d,obs}_m + \Delta^h_{n,ij})$$  \hspace{1cm} (11)

**ANN model**

ANN is a computational processing technique inspired by biological neural networks. The ANN-based model is defined as a data-driven model since it is able to receive, analyse and manipulate information using mathematical functions. Many different ANNs can be defined; one of the most common is the multilayer perceptron, which organizes the neurons in layers: the network receives the input data via the *input layer*, transfers the information to one or more *hidden layers* using different transformation functions and finally provides the final signals in an *output layer*. In this paper a three-layer feed-forward ANN featuring one single *hidden layer* has been used.

A transformation is performed between every layer using pre-defined functions and weights and biases. The inputs, stored into a vector $p$, are multiplied by a weight matrix $W_1$ and then added to a bias vector $b_1$ computing a vector called $n1$ which is then transformed in the hidden layer by a log-sigmoidal function:

$$a_{1i} = \frac{1}{1 + e^{-n_{1i}}}$$  \hspace{1cm} (12)

where $n_{1i}$ is the $i$-th component of the $n1$ vector and $a_{1i}$ is the $i$-th component of the $a1$ vector. Vector $a1$ is multiplied by the weights’ matrix $W_2$ and then added to the bias vector $b_2$ in order to obtain a vector $n2$ which is transformed into the final output vector by a pure linear function in the final layer:

$$a_{2i} = n_{2i}$$  \hspace{1cm} (13)

where $n_{2i}$ is the $i$-th component of the $n2$ vector and $a_{2i}$ is the $i$-th component of the $a2$ outputs vector.

The ANN-based model applied in this paper is aimed at forecasting hourly water demands up to 24 h ahead. The model is thus set up in order to receive as input the water demands observed in the last 24 h and to provide, as outputs, the forecast water demands for the next 24 h. The number of hidden neurons is set equal to 10 and is fixed in the calibration phase looking for the smallest number of neurons that can be used without excessively penalizing the model’s performance (Hsu et al. 1995; Zealand et al. 1999). The log-sigmoid and linear transfer functions are used in the hidden and output layers, respectively. The network parameters, weights and biases, are estimated in the training phase by using the Levenberg Marquardt algorithm (Hagan & Menhaj 1994). In order to prevent overfitting and
improve the robustness of the model, the technique of early stopping is used within the selected calibration period (ASCE 2000; Demuth & Beale 2000).

**CASE STUDIES**

The two models are applied to the hourly water demand time series of three different case studies in order to compare their results in terms of forecasting accuracy. It is worth noting that, for all three case studies, the water demand considered represents the system total demand, inclusive of users’ water demand and leakages. Thus, the water demand time series here considered represent the entire flow entering three different portions of a network or districts placed in northern Italy and managed by HERA S.p.A., characterized by different sizes. These differences allow us to verify the models’ effectiveness when applied to variable sized areas.

In more detail, the first study case CS1 is referred to a portion of a network supplying the city of Ferrara and eleven villages of the area, which count around 120,000 users and 2,500 km of pipes. With an average (system) hourly water demand, inclusive of water leakages, around 900 L/s, it is the biggest portion of network out of the three analysed case studies. CS2 and CS3 are referred to two districts (Polinago and Marano sul Panaro, Province of Modena), each of which presents a single interconnection with the rest of the network, where the water flow has been measured. The areas supplied in these cases are smaller than that of CS1 as well as the number of users (around 50 for CS2 and 20 for CS3), so that the average hourly water demands inclusive of water leakages are around 8 L/s and 1.5 L/s respectively for CS2 and CS3. For all the case studies, the users are mainly residential. All the time series consist of data collected during 2014 and 2015 with 15 min time intervals and have been aggregated in order to obtain hourly data since this is the time step typically used for operational purposes. It is worth noting that by using shorter time steps (in the order of minutes) and given the small sizes of the CS2 and CS3 districts, the water demand time series to be forecast might be affected by several values equal to 0 (or equal to rather constant leakage values) (Buchberger & Nadimpalli 2004; Gargano et al. 2016). Working with hourly demands, the considered models do not have to deal with this kind of issue. In the original time series it is possible to find various anomalous data, or outliers, such as negative terms and non-plausible flow values due to different causes, such as instrumental problems or damage or maintenance activities on the network. Once identified, each outlier has been replaced with the average hourly demand computed taking into account the demands occurred in the same hour of the same day of the week during the same month. This operation did not affect significantly the whole dataset, as the number of replaced data was around 1.5% of the total data in the case of CS2 and less for the other two cases. Data pertaining to year 2014 were used for the calibration of the models and data of year 2015 for the validation. Concerning the calibration of the ANN model, for the application of the early stopping procedure, the calibration dataset was divided in two sets containing, respectively, the first 80% of data for training the network, and the remaining 20% of data for testing.

**Pattern characterization**

Preliminary to the application of the Patt model, for each case study, an analysis of the water demand time series has been performed in order to characterize the periodicities of the water demand, and in particular, seasonal and weekly patterns in daily water demand and a daily pattern in hourly water demand, which represent the basis of the model. Analysis of the time series of the three case studies highlighted that the bigger the portion of a network/district is, the more the data are influenced by periodicities. In fact, the high number of consumers, as, for example, in the case of CS1, balances out all the different behaviours of the single users in terms of water demand, making the time series quite smooth and the patterns easily recognizable. As the number of users decreases (i.e. small districts such as CS2 and CS3), the time series can be highly influenced by anthropic occurrences happening in the network/district, leakages or breaks in the pipes and by the single user behaviours. Indeed, in the CS2 and CS3 cases it is more difficult to ascertain periodicities. In fact, considering daily water demands, CS1 time series show not only a seasonal pattern characterized by higher
demand during the summer season, but also a clear weekly pattern, characterized by higher demands on Saturday and lower demands on Sunday; on the contrary, in the CS2 and CS3 time series, the seasonal and weekly patterns of the daily demands are definitely less marked and even not present at all, as shown, in Figures 2 and 3. Similar considerations apply to the daily pattern of hourly water demands, as shown in Figure 4.

ANALYSIS OF RESULTS

The forecasting accuracy is analysed as a function of the time horizon, therefore considering separately the forecasts for 1, 2 up to 24 h ahead, using the Nash–Sutcliff index $N_S$, defined as:

$$\text{NS} = 1 - \frac{\sum_{i=1}^{nd} (x_{\text{obs}}^i - x_{\text{for}}^i)^2}{\sum_{i=1}^{nd} (x_{\text{obs}}^i - \bar{x}_{\text{obs}}^i)^2}$$

(14)

where $nd$ is the number of observed data, that is the number of hours in year 2014 for the calibration phase and in year 2015 for the validation phase, $x_{\text{obs}}^i$ the observed data, $\bar{x}_{\text{obs}}^i$ the observed data mean value and $x_{\text{for}}^i$ the forecast data.

Figure 5 shows the results obtained by the Patt and ANN models application in both calibration and validation phases. Firstly, it is worth noting that, as expected, the most accurate forecast, for both the models applied to all case studies, is the 1-hour-ahead forecast. The accuracy declines along with the time horizon: in particular, it tends to slightly decrease for lag-times from 1 to 3 and then to remain steady up to 24 h ahead. Furthermore, in general, it can be observed that the performances of both the models tend to decrease from CS1 to CS3. In fact, considering both the models and both the calibration and validation phases, for CS1 the $N_S$ coefficient for the 1-hour-ahead forecast ranges from 0.99 to 0.96, for CS2 from 0.96 to 0.91 and for CS3 from 0.83 to 0.77. In more detail, it is possible to observe that the Patt model performs better than the ANN model in the calibration phase of
each case study, whereas in the validation phase the accuracy provided by this model decreases more than that of the ANN model from CS1 to CS3. Indeed, considering the validation results, it is possible to observe that for CS1 (Figure 5(a)) the Patt model produces better results than ANN, for CS2 (Figure 5(b)) the models show similar performances and
finally, for CS3 (Figure 5(c)), ANN’s performance is better than Patt’s.

The different performances of the two forecasting models can be explained considering that as the number of users decreases, the time series tend to be more variable and less influenced by periodicities, as illustrated in Figure 6, where hourly water demands observed in the same working day (Monday) of each week of the winter season (from November until February) in 2014 for every case study are shown. CS1 water demands (Figure 6(a)) at each hour of the day are included in a narrow range, CS2 data (Figure 6(b)) show an increasing variability, as well as an increasing width of the range at each hour, even though a daily pattern is still recognizable. Finally, CS3’s water demands (Figure 6(c)) are highly variable and the range at each hour of the day gets wider. Furthermore, the increasing variability from CS1 to CS3 and the corresponding lack of demand periodicity makes the patterns estimated according to calibration data less representative of the validation data (see Figure 7). On the whole, the increasing data variability leads to a decrease in the forecasting accuracy, and in particular, the Patt model shows a remarkable loss in accuracy from the calibration to the validation phase for CS2 and CS3. In fact, as the time series become more variable and less influenced by periodicities, the information provided by the patterns, on which the Patt model is based, become less useful to forecast future demands, and thus the pattern-based forecasting approach becomes less effective than a pure data-driven approach.

CONCLUSIONS

In this paper two short-term water demand forecasting models based on different forecasting techniques, are compared. The models are applied to the water demand time
series, inclusive of users’ demands and leakages, of three case studies, characterized by different dimensions and number of users. Both the models perform medium–high forecasting accuracy, with NS coefficients variable between 0.98 and 0.68, although, as the number of users decreases, both the models’ performances decline due to the increasing variability of data. In particular, as the number of users decreases, periodic behaviours become less evident, the patterns become less representative of water demand time series, which tend to be more affected by anthropic occurrences and leakages. Thus, as the number of users decreases both the models lose forecasting accuracy and in particular the Patt model performs less accurate forecasts (with an average NS slightly lower than 0.7) than the ANN model (with an average NS slightly higher than 0.7) in the case study where the number of users is around 20. Concluding, when applied to parts of network/districts including a large number of users, the pattern-based model tends to be more efficient than the ANN-based one. For smaller districts, with a lower number of users and an increasing variability in water demands, the pattern-based model tends to be outperformed by the ANN-based model since it is less influenced by patterns.

ACKNOWLEDGEMENTS

This study was carried out as part of the ongoing PRIN 2012 project no. 20127PKJ4X ‘Tools and procedures for an advanced and sustainable management of water distribution systems’ funded by MIUR, and FIR2016 project ‘Metodologie gestionali innovative per le reti urbane di distribuzione idrica’ funded by University of Ferrara.

REFERENCES


First received 23 December 2016; accepted in revised form 9 March 2017. Available online 24 March 2017