Herrera Fernández, AM.; García-Díaz, JC.; Izquierdo Sebastián, J.; Pérez García, R. (2011). Municipal water demand forecasting: Tools for intervention time series. Stochastic Analysis and Applications. 29(6):998-1007. doi:10.1080/07362994.2011.610161.

MUNICIPAL WATER DEMAND FORECASTING: TOOLS FOR INTERVENTION TIME SERIES

M. Herrera[1*], J. C. García-Díaz[2], J. Izquierdo[1], R. Pérez-García[1]

[1]*Fluing – Instituto de Matemática Multidisciplinar, Universitat Politècnica de València, Edificio 5C, C. de Vera s/n, 46022 Valencia*

[2]*Centro de Gestión de la Calidad y del Cambio, Universitat Politècnica de València, Edificio 7A, C. de Vera s/n, 46022 Valencia*

Abstract

This paper introduces some approaches to common issues arising in real cases of water demand prediction. Occurrences of negative data gathered by the network metering system, and demand changes due to closure of valves or changes in consumer behavior are considered. Artificial neural networks (ANNs) have a principal roll modeling both circumstances. First, we propose the use of ANNs as a tool to reconstruct any anomalous time series information. Next, we use what we call interrupted neural networks (I-NN) as an alternative to more classical intervention ARIMA models. Besides, the use of hybrid models that combine not only the modeling ability of ARIMA to cope with the time series linear part, but also to explain nonlinearities found in their residuals, is proposed. These models have shown promising results when tested on a real database, and represent a boost to the use and the applicability of ANNs.

Keywords: water demand; ARIMA models; intervention analysis; neural networks; hybrid models

[*]*Corresponding author. Tel.: +34 96 387 98 90 Fax: + 34 96 387 79 81*
Email address: mahefe@gmmf.upv.es (Manuel Herrera)

1. Introduction

The most important consideration for the planning and the operation of a water distribution system is to satisfy consumer demands. Thus, it is imperative to provide consumers with quality water in adequate amounts, at reasonable pressure and at all the times, ensuring some degree of reliability into the water distribution system. Efficient operation and management of an existing water supply system requires short-term water demand forecasts as a crucial input. The estimation of future municipal water demand is central to the planning of a regional water supply (Zhou *et al.*, 2002), becoming an essential tool for design, operation and management of the system. These predictions are fundamental in taking decisions in water management issues such as pricing policies and planning new developments or system expansions, and estimating the size and operation of reservoirs and pumping stations (Bougadis et al., 2005). Thus, short-term demand projections help water managers to make more informed water management decisions and to balance the needs of water supply of residential and industrial demands (Jain and Ormsbee, 2002).

In this paper, our analysis is based on hourly water demand data in a village of south-east Spain. Water demand around the Mediterranean basin is growing at an alarming rate. Like many Mediterranean regions, south-east Spain is suffering from large stress in the use of groundwater from its major aquifers. This region is a prime example of an area where the aquifers are under pressure and groundwater supplies essential needs for agriculture and tourism. The time series data we consider here have different singularities and appear with certain frequency in Hydraulics. One characteristic is that negative demands are observed during the period for which the time series of water demand is under study. We propose the use of an artificial neural network (ANN) to interpolate these anomalies. Also, we address the occurrence of some changes affecting the level of the demand average. This issue suggests the use of intervention analysis to model the series. We study two options: an intervened ARIMA and a new concept of *interrupted* ANN. Finally, to analyze the series we apply a hybrid methodology that improves the achieved results. This is performed by fitting an intervened ARIMA to the data and using an ANN to explain its residuals.

The paper is organized as follows. In Section 2, the forecasting procedures are described. As regards time series analysis, the considered techniques include ANN and ARIMA models. Transfer functions and intervention analysis also are introduced because they are useful to represent the impact of special events on

water demand. In Section 3, a case study based on water demand historical hourly data for six months is analyzed in detail. This instance serves as a working line to propose different alternatives to classical issues of the considered forecasting tools. Section 4 summarizes the conclusions of this paper.

2. Water demand forecasting methods

2.1 ARIMA models

Water demand can be considered as a random process $\{X(t)\}$ in continuous time, from which a time series $\{x_t\}$ is obtained by sampling at discrete times. The symbol t denotes the time. The stochastic nature of water demand as a function of time has frequently been modeled with seasonal ARIMA models (Yamada et al., 1992).

The multiplicative seasonal ARIMA or SARIMA(p,d,q)(P,D,Q)$_s$ model is given by

$$\varphi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D(x_t - c)=\theta_q(B)\Theta_Q(B^s)a_t \qquad (1)$$

where x_t is the water demand in period t; c is a constant term; s is the number of periods in the seasonal cycle; B is the lag operator $(B^j X_t = X_{t-j})$; ∇ is the difference operator $\nabla = (1 - B)$; ∇_s is the seasonal difference operator $\nabla_s = (1 - B^s)$; d an D are the orders of differencing; a_t is a zero-mean white noise error term; and $\varphi_p, \Phi_P, \theta_q, \Theta_Q$ are polynomial functions of orders p, P, q and Q, respectively (Peña et al., 2000).

The Box-Jenkins methodology (Box and Jenkins, 1976) provides a step-by-step procedure for ARIMA analysis to find the best fit of a time series to past values of this time series, in order to give forecasts.

Fuzzy logic, rough sets and decision rules (Fernández-Baizán, 2000) are proposed as an alternative to the ARIMA models, but the data requirements are very large and different variables such as wind direction, air temperature or humidity need to be measured; on the other hand, in ARIMA just the time sequence is necessary to achieve suitable results. Other predictive alternatives are discussed by Herrera et al., 2010.

2.2 Transfer function and intervention analysis methodology

Univariate ARIMA models are useful for analysis and forecasting of single time series. In such situations,

one can only relate the series to its own past and does not explicitly use the information contained in other pertinent time series. However, a time series is not only related to its own past, but may also be influenced by the present and past values of other related time series. This class of models is referred to as transfer function models by Box and Jenkins (1976). A special case of transfer function models is called *intervention model* class, which can be used to evaluate the effect of the external events, or to incorporate the interventions into a time series model to possibly improve parameter estimates or forecasts (McLeod and Vingilis, 2005). This class of models is typically used as a means to assess the impact of a discrete intervention on a time series. In the following we shall assume that the time t at which the change (or intervention) occurs is known.

Usual transfer functions are instantaneous or gradual and can have permanent or transitory effects. The time series analyzed in this paper (see Section 3) is affected by a step function, in view of its permanent effects. If the step jump is sharp, then it will be of instantaneous character. The effect of a step function on the y_t series, that follows an ARIMA model, can be represented by the intervention model: $y_t = f P_t^h + N_t$, where f is the transfer function, N_t is a stationary model, and P_t^h is a pulse function. If x_t and y_t are two stationary time series, x_t being the input or cause variable, and y_t the output or response variable; and N_t is an uncorrelated noise with the input series x_t that is modeled as ARIMA, then the final model is expressed by:

$$Y_t = \sum_{j=0}^{\infty} \tau_j X_{t-j} + N_t \qquad (2)$$

The operator $VB = \sum_{j=0}^{\infty} \tau_j X_{t-j}$ is called transfer function and the weights τ_i are the system pulse weights, as see in expression (2).

This model undergoes the problem of infinite parameters. One alternative is the *V(B)* polynomial truncating, that is: $y_t = \tau_0 x_t + \tau_1 x_{t-1} + \ldots + \tau_h x_{\tau-h} + N_t$, where h is chosen so that the effect of the not considered posterior delays is negligible. Nevertheless, truncating this series it is a difficult decision that can be avoided by establishing *V(B)* as the next polynomial rate:

$$V(B) = \frac{W(B)}{\delta(B)} B^b \qquad (3)$$

where $W(B) = w_0 + w_1 B + \ldots + w_s B^s$; $\delta(B) = 1 - \delta_1 B - \ldots - \delta_r B^r$. In this form the number of parameters

diminishes and the model is:

$$y_t = \frac{W(B)}{\delta(B)} B^b x_t + N_t \qquad (4)$$

with $N_t = \frac{\theta(B)}{\phi(B)} a_t$ being a white noise process.

If disturbance does not exist, that is, if $N_t = 0$, the model becomes: $(1- \delta_1 B -...- \delta_r B^r) y_t = (w_0 + w_1 B +... + w_s B^s) x_t$. That is to say, an ARIMA model is a particular case of a dynamical model of transfer function in which the input is a white noise. In addition, disturbance does not exist and the function of transference is the quotient of two polynomials corresponding to the moving average (MA) part (the numerator) and to the auto-regressive (AR) part (the denominator). Thus, the conditions of a stationary and invertible process for the transfer function are the same than those for the ARIMA models, namely, the roots of the polynomials $w(B)$ and $\delta(B)$ fall out of the unit circle.

Box and Tiao, 1976, introduced a model for intervention analysis that has the same form as the transfer function model (eq. 2), except that the input $\{X_t\}$ and the coefficients $\{t_j\}$ are chosen in such a way that the changing level of the observations of $\{Y_t\}$ is well represented by the sequence $\sum_{j=0}^{\infty} \tau_j X_{t-j}$. Then, for $\{Y_t\}$ with $EY_t = 0$ for $t \leq T$ and $EY_t \to a$ as $t \to \infty$, a suitable input series is the step function of equation 2.

Having chosen an appropriate form for X_t and possible values of b, q and p by inspection of the data, the estimation of the parameters and the fitting of the model for $\{N_t\}$ can be carried out by using the next steps:

1. Input model with $r = s = 0$ where the least square estimation of w_0 is obtained. The residuals are saved as the first $\{N_t\}$ estimators.

2. Fit a seasonal ARIMA model to the residuals data (with no mean correction and using maximum likelihood estimators).

3. Run step 1 using the estimate $\hat{w}_0$ and the $\{N_t\}$ found in 2.

4. Test the new residual series for whiteness and, after passing all the tests, conclude that the model in equation 4 is satisfactory.

2.3 Artificial neural networks

An artificial neural network (ANN) is an interconnected group of artificial neurons that uses a mathematical model or computational model for information processing based on nodes interconnected in a series layers. In most cases, an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network (Bishop, 1995). In more practical terms, neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data by exploiting its main advantage, namely, the capability of the network to self-learn. Knowing the inputs and desired output(s), the ANN model will try to reproduce the observed outputs through a series of iterations. The most common ANN network is the feedforward network, which uses the back-propagation algorithm for training.

Multi-layer networks use a variety of learning techniques, the most popular being back-propagation (Bougadis *et al.*, 2005). Usually, a typical three-layer feedforward model is used for forecasting purposes (Lingireddy and Ormsbee, 1998). On the one hand, the input nodes may contain both important co-variables at the current time and the previous time series lagged observations. On the other hand, the output provides the forecast for the future values. Hidden nodes with appropriate non-linear transfer functions are used to process the information received by the input nodes. Finally, the model can be written as (Zhang and Qi, 2005):

$$Y_t = \alpha_0 + \sum_{j=1}^{h} \alpha_j f\left(\sum_{i=1}^{p} \beta_{ij} y_{t-i} + \beta_{0j}\right) + \varepsilon_t$$

where p is the number of input nodes, h is the number of hidden nodes, f is a sigmoid transfer function; α_j, with $j = 0, 1, ..., h$, is the vector of the weights from the hidden to the output nodes and β_{ij}, with $i = 0, 1, ..., p$ and $j = 1, 2, ..., h$, are the weights from the input to hidden nodes. α_0 and β_{0j} are the weights of the arcs leaving from the bias terms.

3. Forecasting tools applied to water demand predictions

3.1 Case study

This paper introduces a case study of a water demand time series for a small town close to the city of Murcia (located in south-east Spain). It has a population of about 5000 inhabitants and an extension close to 8 km². The water demand data was taken as the average difference between flows entering and leaving this town. A failure on the metering stations originated that the data showed negative demands throughout several consecutive days (see Fig. 1). This fact led us to propose the use of an ANN as a tool for interpolation and reconstruction of such anomalous information. Once purified, the data series is analyzed by the methodology ARIMA, in this case by intervention, since from April the average changes sharply. This intervention is due to the modified record in the testing pressures of the pumping stations. Field measurements were conducted at the treatment and control sites from January 2005 to June 2005.

[FIGURE 1]

The database can be used to supply the average water consumption on a hourly basis for a sample of costumers, using the information on a six-month time period. To estimate the model parameters we use 6 months of hourly data: from January 1, 2005 to May 31, 2005. The last month is reserved to evaluate our forecasts. The data suggests a simple level change that can be modeled using intervention analysis.

3.2 Interpolation data using ANN

An ANN model is developed for interpolating water demand (Fariñas and Pedreira, 2002) along 3 days in April, namely 10, 11 and 12. During these days negative demands were observed (Fig. 1). The sample data are divided into training part (first 2275 items) and validation part (next 100 items). The interpolation lags are items from 2376 to 2447. The validation part is the last 10 days observations before interpolation time. The training sample is used to estimate the parameters for any specific model architecture, and the validation sample is then used to select the best among all the considered models. Because of the number of input data (and then the training samples) is so long, there is no problem with over-training. In this case the training is the so-called “asymptotic training”.

The proposed ANN is a feedforward three-layer perceptron (Zealand *et al.*, 1999; Wang *et al.*, 2006). The

input layer is composed by three nodes: hour, week day and weekend (this last node gives the information whether the day is a holiday or not). The tested hidden nodes vary from 4 to 14, with a logistic transfer function for them, and the identity function for the output node. Two learning rates were probed each time: 0.01 and 0.1. Finally, the best architecture in the sense of root mean square error (RMSE), and mean absolute error (MAE), is a 3-8-1 scheme, with learning rate of 0.1 (Kneale *et al*., 2001). The errors obtained were 7.53 m^3 in RMSE terms and 5.59 m^3 in MAE. These results are reinforced when the parameters number is considered, as in the Akaike (AIC) and Bayesian (BIC) criteria (see Table 1).

[TABLE 1]

3.3 Modeling data using intervention analysis

Once the data has been properly interpolated, it is important to seek a time series stationary condition: constant mean and variance. A transformation of the original data is usually applied to attain a more variance-stable data, while simple and seasonal difference operators can be used to obtain a more stable mean. In this case, the logarithmic transformation was used for achieving stationarity. Besides of this, the data proposes a simple level change, which can be modeled using intervention analysis. The jump in the average suddenly happens at noon on April 22nd. The consequences directly affect the data average, which exceeds the 19 m^3 to become stable around 32 m^3. Their corresponding variances also grow from 8.4 to 16. The suitable transfer function in this case will be classified as instantaneous and their effects will be permanent.

In this intervention process, we first take w_0 (defined in subsection 2.2) equal to 3.47 (32 m^3, in the original units), the *Log* of the series mean after the jump, as an initial estimation of w_0. Next, we perform a least squares regression of the series output, and we get the estimated value of 0.63. The noise of this series estimates $\{N_t\}$ in the model. *Maximum likelihood* estimation gives the SARIMA$(1,1,1)(1,1,1)_{24}$. Their residuals are white noise and offers initial estimations for the final model. Then, using the found estimate of w_0 and the given model for N_t leads to the model: $Y_t = 0.57\, I_t^{2676} x_t + N_t$, which we specify in Table 2.

[TABLE 2]

3.4 Proposing alternative models

There are various ways to formalize the validation about the previously obtained results. The Peña and Rodríguez portmanteau test (Peña and Rodríguez, 2002; Peña and Rodríguez, 2006), the classical Ljung and Box test (Ljung and Box, 1978), and the Monti and Hong tests (Monti, 1994) are known examples. In our case-study we do not validate the intervention model using the first Peña and Rodríguez test, but we obtain that the residuals are white noise using the support of Ljung-Box test and a graphical inspection. In addition, the results will depend on the working sample and on the validation set. This ambiguity motivates the introduction of some bootstrap time series methodologies to avoid bias and achieve robustness in the proposed models (Herrera et al., 2010). Next subsections show alternative solutions on the study of an intervention time series process.

3.4.1 Interrupted neural networks

A first proposed alternative is forecasting with a neural network (it is arranged under similar conditions that the proposed ANN in Subsection 2.3 and used as an interpolation tool in 3.2). On one hand, the case of approaching an intervened time series suggests the use of two neural networks. But on the other hand, although there are two different time series, both have a common origin. Thus, we propose what we call an *Interrupted* Neural Network (I-NN). It consists in arranging two different ANN, one for each part of the series (pre and post-intervened). The I-NN particularity is that the second ANN automatically inherits the architecture of the first one, and it is initialized with the weights of the final solution of this pre-intervened network. In this way, we can achieve better algorithmic convergence.

Turning back to our case-study, we apply a 3-8-1 architecture in the I-NN proposed. Using the model to predict the water consumption in the 24 hours of the first day of June, we obtain a RMSE of 11.62 m^3 and MAE of 9.74 m^3. We can see a summary of these results in Table 3.

3.4.1 Hybrid models

This proposal exploits the strengths of traditional ARIMA time series approaches and artificial neural networks (Wang et al., 2006; Zhang, 2005). The idea consists in using an ARIMA model to analyze the linear

part of the problem and an ANN to model their residuals. The joint model captures different forms of relationship in the time series data, maintaining the ARIMA interpretation but modeling the nonlinear patterns that can appear. The proposed hybrid model is then composed by a linear and a nonlinear component as see in equation 5:

$$\hat{y}_t = L_t + Nl_t \quad (5)$$

where y_t denotes the original time series, L_t denotes the linear component and Nl_t denotes the nonlinear component. L_t is estimated by an ARIMA model and residuals obtained from the ARIMA model, $e_t = \hat{y}_t - L_t$, are estimated by an ANN.

We test this methodology in our case-study to predict the water demand of the first day of June. Then, adding to our intervened ARIMA model (see Subsection 3.3) a 3-8-1 ANN to fit their residuals, we obtain errors such as a 7.69 m³ in RMSE and 5.18 m³ in MAE. Table 3 summarizes the results.

[TABLE 3]

The best results are achieved with the analysis by a hybrid model. Then, the final model is a SARIMA$(1,1,1)(1,1,1)_{24}$, for the linear part and an additional 3-8-1 ANN modeling their residuals. The forecasts are shown in Figure 2.

[FIGURE 2]

4. Conclusions

This paper has analyzed a model to forecast the municipal water demand in a town of Spain using the ARIMA models. To cope with the difficulties found on these data analysis we have proposed different tools, including some interesting conceptual approaches using artificial neural networks (ANN).

For one thing, we have worked with some important quantity of negative data of water demand, which we need to interpolate. The use of artificial neural networks as tools for this interpolation fits perfectly in the

philosophy of the ARIMA models, restoring automatically the necessary data to prolong the time series analysis. And for another, a simple level change found on the data has been modeled using intervention analysis. Nevertheless, the paper introduces a new alternative approach based on interrupted neural networks. This breed of ANN takes advantage of the a-priori model of the pre-intervened series to perform more suitably the model after the intervention.

Other working line is the use of hybrid models in time series. These are based in linear ARIMA models, but they are able to gather possible nonlinearities by studying their residuals by an ANN. In a comparative process on our case-study, the obtained model by this hybrid approach exhibit higher predictive power.

Acknowledgements

This work has been supported by project IDAWAS, DPI2009-11591, of the Dirección General de Investigación of the Ministerio de Ciencia e Innovación of Spain, and ACOMP/2010/146 of the Consellería de Educación of the Generalitat Valenciana. Besides, we will thanks to "*Aguas de Murcia*" for the collaboration in this work and for the availability of the data.

References

Bishop, C. M. (1995) "Neural Networks for Pattern Recognition", Oxford University Press

Bougadis, J., Adamowski, K. and Diduch, R. (2005) "Short-term Municipal Water Demand Forecasting", Hydrological Processes, 19, pp. 137-148

Box, G. E. and Jenkins, G. (1976) "Time Series Analysis, Forecasting and Control", San Francisco: Holden-Day

Box, G. E. and Tiao, G. C. (1976) "Intervention Analysis with Applications to Economic and Environmental Problems", Journal of the American Statistical Association, vol. 70, pp. 70 – 79

Fariñas, M. and Pedreira, C. E. (2002) "Missing Data Interpolation By Using Local-Global Neural Networks" International Journal of Engineering Intelligent Systems for Electrical Engineering and Communications, vol. 10, 2, pp. 85-92

Fernández-Baizán, C. *et al.* (2000) "Mining Time Series of Meteorological Variables Using Rough Sets – A Case Study", Binding Environmental Sciences and Artificial Intelligent (BESAI 2000 – Germany)

Herrera, M., Torgo, L., Izquierdo, I. and Pérez-García, R. (2010) "Predictive Models Forecasting Hourly Water Demand", Journal of Hydrology, 387 (1-2), pp. 141 – 150

Jain, A. and Ormsbee, L. E. (2002) "Short-term Water Demand Forecasting Modelling Techniques—Conventional Versus AI", Journal AWWA, 94 (7): 64-72.

Kneale, P., See, L. and Smith, A. (2001) "Towards Defining Evaluation Measures for Neural Network Forecasting Models", Proceedings of the Sixth International Conference on GeoComputation, University of Queensland, Australia.

Lingireddy, S. and Ormsbee, L.E., (1998) “Neural Networks in Optimal Calibration of Water Distribution Systems,” Artificial Neural Networks for Civil Engineers: Advanced Features and Applications. Ed. I. Flood, and N. Kartam. American Society of Civil Engineers, p277

Ljung, G. M. and Box G. E. P. (1978). "On a Measure of a Lack of Fit in Time Series Models". *Biometrika* 65, pp. 297 – 303

McLeod, A. I. and Vingilis, E. R. (2005) “Power Computation for Intervention Analysis” , American Association and American Society for Quality, Technometrics, vol. 47, number 2, pp.174 – 181

Monti, A. C. (1994) “A proposal for residual autocorrelation test in linear models”, Biometrika 81, pp. 776 – 780

Peña, D. and Rodríguez, J. (2002) “A Powerful Portmanteau Test of Lack of Fit for Time Series” Journal of the American Statistical Association vol. 97, n. 458 pp. 601 – 610

Peña, D. and Rodríguez, J. (2006) “The Log of the Determinant of th Autocorrelation Matrix testing goodness of Fit in Time Series” Journal of Statistical Planning and inference 136, pp. 2706 – 2718

Peña, D., Tiao, C. G. & Tsay, R. S. (2000) “A Course in Time Series Analysis”, Wiley

Wang, W., Van Gelder, P. H., Vrijiling, J. K. and Ma, J. (2006) “Forecasting Daily Streamflow Using Hybrid ANN models”, Journal of Hidrology 324 (1-4), pp. 383 – 399

Yamada, R., Zhang, S. P. and Konda, T. (1992) “An Application of Multiple *ARIMA* Model to Daily *Water Demand* Forecasting”, Annual Reports of NSC, vol. 18, number 1

Zealand, C. M., Burn, D. H. and Simonovic, S. (1999) “Short Term Streamflow Using Artificial Neural Networks”, Journal of Hydrology 214, pp. 32 – 48

Zhang, G. P. and Qi, M. (2005) “Neural Network Forecasting for Seasonal and Trend Time Series”, European Journal of Operational Research 160, pp. 501 – 514

Zhou, S. L. *et al*. (2002) “Forecasting Operational Demand for an Urban Water Supply Zone”, Journal of Hydrology 259, pp. 189 – 202

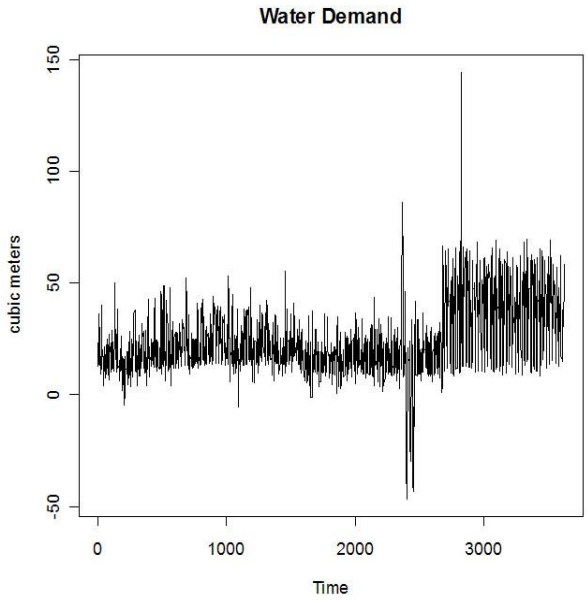

Fig. 1. Original data of water demand in the case study

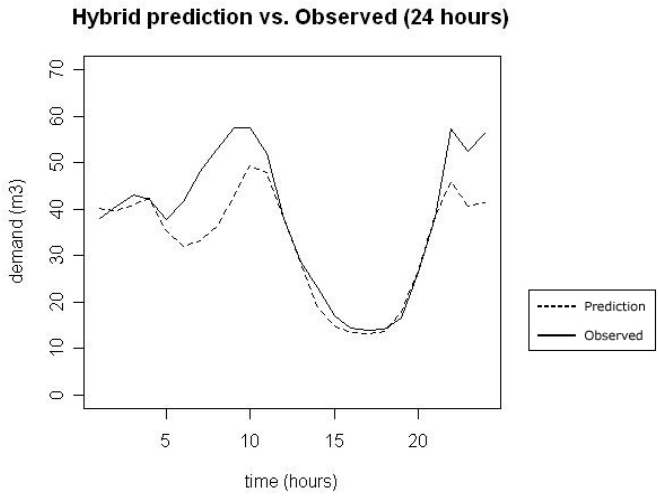

Figure 2. Forecasting with the hybrid model along the first day of June in the case-study

Table 1. AIC and BIC table with the three best results of the architectures analyzed

Architec.	Alpha	RMSE	MAE	AIC	BIC	M patterns	N weights
3-7-1	0.1	8.9	7.3	2120.9	2158.4	2175	28
3-8-1	0.1	7.5	5.6	1967.3	2010.0	2175	32
3-9-1	0.1	8.1	6.4	2047.9	2096.1	2175	36

Table 2. Parameters estimated for the final model of the intervened ARIMA

	AR1	MA1	SAR1	SMA1
	0.75	-0.98	0.11	-0.94
st. error	0.01	0.01	0.02	0.16

Table 3. RMSE and MAE of the forecasting methods used

	RMSE	MAE
intervened ARIMA	8.45	6.03
interrupted ANN	11.62	9.74
hybrid model	7.69	5.18