Flow meter Data Validation and Reconstruction using Neural Networks: Application to the Barcelona Water Network

Hector Rodriguez[1] and Vicenç Puig[1] and Juan J. Flores[2] and Rodrigo Lopez[3]

Abstract— The use of false or erroneous data can lead to wrong decisions when operating a system. In case of a water distribution network, the use of incorrect data could lead to errors in the billing system, waste of energy, incorrect management of control elements, etc. This paper is focused on detecting flow meters reading abnormalities by exploiting the temporal redundancy of the demand time series by means of artificial neural networks (ANN). Communication problems with the sensor generate missing data and bad maintenance service in the flow meters produce false data. In this work, a methodology to detect the false data (validate) and replace the missing or false data (reconstruct) is proposed. As a core methodology, ANNs are used to model the time series generated from the water demand flow meters, and use the confidence intervals to validate the information. To illustrate the proposed methodology, the application to flow meters in the water distribution network of Barcelona is used.

I. INTRODUCTION

The purpose of a water distribution system is to deliver water to consumers with appropriate quality, quantity and pressure. The water supply system collects, stores, treats, and distributes the water for homes, commercial establishments, industry, and irrigation, as well as for such public needs as firefighting, street flushing, etc. The main goal of the water distribution system is to satisfy the consumers' demand (with appropriate quality). This implies the continuous delivering of quality water at a reasonable pressure to users. To maintain the quality of service, the water distribution system requires a real-time monitoring system. This system will facilitate finding faults in the water distribution system.

Among the activities of the water distribution system, it is to periodically measure the amount of water supplied to each user. In a complex water distribution network, a telecontrol system must acquire, store, and validate data from many flow meters and other sensors every few minutes to achieve accurate monitoring of the whole network in real-time. Frequent operation problems occurs in the communication system between the set of the sensors and the data logger, or in the telecontrol itself, generate missing data during certain periods of time. Missing data must, therefore, be replaced by a set of estimated data [1], [2].

Missing data is not the only problem, a second common problem is the lack of reliability of the flow meters (offset, drift, breakdowns, etc.), producing false flow data readings. These false data must also be detected and replaced by estimated data, since flow data are used for several network water management tasks, namely: planning, investment plans, operations, maintenance, and billing/consumer services and operational control [1], [2].

This paper is related to works in the area of Artificial Neural Network (ANN) for the task of modelling time series [3], [4], [5], [6], but differ in the way of creating the architecture of the ANN, and also the way to train it. In previous works ([7], [8], [9]), we have used Evolutionary Algorithms (EA) to define the architecture of the ANN and train it. The results show that the forecast accuracy achieved using EA presents a better performance than traditional gradient-based training algorithms.

In this paper, we propose a methodology to detect (validate) and reconstruct data of water demand time series by modelling the time series with an ANN. With this model, it is possible to detect anomalies and replace the missing data. One of the inputs of the ANN is the mode (or regime) of the next day to model. In previous works related to regime recognition, the identification of regimes behaviour has used qualitative information. Benmouiza and Cheknane [10] proposed the implementation of a global NAR (Non-Linear Autoregressive) neural network predictor to estimate the regimes associated to another local NAR neural network predictor for the hourly global solar radiation. Kumar and Patel [11] propose a predictive algorithm using data clustering and local training models that combined produce the forecast. Martinez Alvarez et al. [12] use clustering to group the days with similar patterns, regarding the variation of the electricity cost in working days and holidays. In [13], a daily seasonality ARIMA model with hourly pattern with the goal of working in daily and hourly scales is proposed. The seasonal ARIMA predicts total days consumption and a daily pattern is selected according to a calendar for distributing the consumption hourly along the day. The proposed case study, the Barcelona water network model has been already studied by [2] also with the focus on the validation and reconstruction of flow meter data.

The structure of the paper is the following: In Section II, the approach proposed for the modelling the flow meter data series is introduced. Section **??** describes the flow meter data validation and reconstruction approach. Section IV presents the case study used to illustrate the proposed approach. Section V summarizes the main results obtained in the considered case study. Finally, in Section VI draws the main conclusions.

[1] Hector and Vicenç are with the Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Carrer LLorens Artigas, 4-6, Barcelona (Spain), 08028. `hrodriguez@iri.upc.edu` and `vicenc.puig@upc.edu`

[2] Juan F. works at División de Estudios de Postgrado, Facultad de Ingeniería Eléctrica, Universidad Michoacana de San Nicolás de Hidalgo, Morelia, Michoacán, Mexico `juanf@umich.mx`

[2] Rodrigo is with IMT Institute for Advanced Studies, Lucca, (Italy). `rdglpz@gmail.com`

II. Modelling the Time Series

A. Background

Modelling a water demand time series is not something new, since water is one of the most basic non renewable natural resources to sustain life and ecosystems. The development of new forecasting models and strategies are strongly related to multidisciplinary novel research, mainly in physics, computing, and statistics.

In the literature regarding modelling time series (e.g. [14], [15]), we found that there is a strong effort on finding the best way to decompose time series in several simpler time series to better fit an accurate model. This is not an easy task since in many real cases there is not an available model that describes the dynamic fluctuation of the data. In the early successful stage in statistics the *divide and conquer* strategy has been used. For example, the decomposition on different basic components that might explain the general dynamics of the time series, such as trend, seasonal, random, and cyclical components that are integrated in the ARIMA methodology [16], [17].

Nowadays, with the growing of computational resources and the development of machine learning and pattern recognition algorithms (e.g. [18], [19], [20]), it is possible to model more complex time series. There are practical cases where the single linear modeling approach is not enough for systems that present different behaviors along time [10], [11], [12], [13]. These behavior changes might be produced by changes on dynamical regimes. Although the multi-modeling approach was born with the analysis of partially known real systems, the same ideas can be adopted for time series modelling where there is no knowledge about the system behind the dynamics. That is the case of water demand time series.

Regarding the water demand, different methodologies for modelling have been explored. In particular, revising the literature, several methodologies (e.g. based on Box Jenkins, ANNs, Holt-Winters, ARIMA, etc.) that deal with this problem [21], [22], [13], [23] already were satisfactorily applied.

B. Proposed approach

Recent research activities in ANNs have shown that they have powerful pattern classification and pattern recognition capabilities. Inspired by biological systems, particularly by research into the human brain, ANNs are able to learn and generalise from experience. Currently, ANNs are used for a wide variety of tasks in many different fields of business, industry, and science [24].

One major application area of ANNs is forecasting (modelling time series). ANNs provide an attractive alternative tool for both forecasting researchers and practitioners. Several distinguishing features of ANNs make them valuable and attractive for a forecasting task [24].

Determination of the optimal number of hidden neurons is a crucial issue. If the hidden layer is too small, the network can not possess sufficient information processing power, and thus yields inaccurate forecasting results. On the other hand, if it is too large, the training process will be very long.

Given a time series, we need to provide a neural model capable of producing an acceptable model that fits with the behavior of the time series. The ANN architecture used for this task is a Multi-Layer Perceptron (MLP). A MLP as a universal approximator, can learn any function given, as long as it has enough neurons in the hidden layer. That fact allows the network to capture the different forms of the function to be modelled. The network used in this work is a three layer MLP network. The past observations of the time serie are used as input of the ANN. The hidden layer has m neurons, and the output layer corresponds to the forecast ($\hat{y}_{t+1}$), and the sigmoid function as the activation function. The ANN that forecast $\hat{y}_{t+i}$ is defined as:

$$
\hat{y}_{t+1} = f_1(\sum_i^n w_i x_i)
$$
$$
x_i = f_2(\sum_j^h w_{ij} y_{t-j})
$$
(1)

where f_1 and f_2 are the activation functions, and w are the coefficients or the weight connections.

An example of this kind of architecture is depicted in Figure 1. The proposed network has k inputs, which are selected from the k previous measurements, plus the next mode. It is understood as next mode, the estimation of the next day qualitative behavior (e.g. labor day or holiday). It has m neurons in the hidden layer, one output (the corresponding hour prediction ($\hat{y}_{t+1}$)), and the sigmoid function as the activation function.
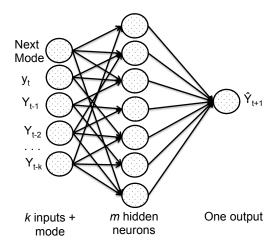

Fig. 1. MLP architecture with $k+1$ past observations plus the next mode, m hidden neurons in the hidden layer and one output ($\hat{y}_{t+1}$).

C. Regimes

Time series can present different dynamic behaviors that cannot be fitted with a single model. These different behaviors might be considered as changes of regimes that might be unknown. Examples of this kind of behaviors are found in natural processes, such as temperature, solar radiation, wind speed, etc., where the behavior changes according to some rules. The behavior in these cases is different during

the winter than during the summer, but regular in its own season. Also, behavioral changes can be found in time series corresponding to human activities such as water demand, electricity consumption, etc. In those cases the behavior of the series is reflected according to human activities (i.e. holiday or working days). Each behavior can be represented by a regime that is characterized by a qualitative behavior.

In [25], to predict the time series water consumption, it is decomposed in two kinds of time series: a quantitative and qualitative one. A quantitative time series that represents the total consumption of water corresponding to every day is considered. The qualitative time series represents the sequence of distribution patterns (corresponding to working and holy days) called modes to be used to distribute the predicted amount of water. The training requirements for both time series are different and consequently several kind of algorithms can be used. The mode defined in this work is used as an input to the ANN.

D. Training Process

For the training process, taking into account the results obtained in previous works ([7], [8], [9]) where we tested traditional methods (gradient-based [26]) versus evolutionary computation, we use Genetic Algorithms (GA) to define the architecture of the net and the training process for finding the best structure and weights of the ANN.

GA is an optimisation technique inspired on Darwin's principle of evolution. That is, it mimics a simplistic version of the process of biological evolution, which consists of creating a population of individuals, where each individual represents a prospective solution of the problem being solved. GA modifies this population using genetic operators: selection, mutation, recombination, etc. This stage, called a generation, repeats until a termination criterion is met. At the end of the process, the best individual (e.g. the one with best fit) found during the evolution is returned as the solution of the problem.

Determining the best ANN architecture for forecasting is an optimization problem; we use GA to find the optimal ANN architecture and its weights. The function to minimise is:

$$min(MSE) = \sum_{k}^{N} (y_k - \hat{y}_k)^2 \qquad (2)$$

where y_k is the real data, $\hat{y}_k$ is the estimate data, N is the number of points in the training set, and MSE (Mean Square Error) is the function to minimise.

GA defines the architecture of the ANN and the weights of the neurons connections. Each individual (the chromosome) is defined as a vector of real numbers. The vector is coded in three parts. The first part of the vector defines the architecture of the net (n and h of Equation 1). This part is a a vector of two elements: the first element is coded and defines the number of previous measurements to use as an input (n), and the second element, also coded, defines the number of neurons in the hidden layer (h). The second and third part

of the vector corresponds to the weights connections from the input neurons to the neurons in the hidden layer (second part); its size is $(n+1)*h$ (n is the number of inputs, h is the number of neurons in the hidden layer). And the third part corresponds to the weights from the neurons in the hidden layer to the output layer (third part)its size is h. Figure 2 shows the structure of the chromosome.
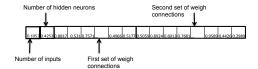

Fig. 2. The chromosome of the ANN represented as a vector of reals. The first two positions of the vector defines the structure of the net, and the following data corresponds to the weights connections.

As mention before, the first part defines the architecture of the ANN. In this part, the two elements are coded in real format, but they represent an integer number as follows

$$I = Integer(LimMax * realData) \qquad (3)$$

where $LimMax$ is the largest possible number defining the boundary of the architecture. We have one $LimMax$ to define the number of inputs, and another one to define the number of neurons in the hidden layer. I is an integer number between zero and $LimMax$; $realData$ is the value contained in the vector of reals, and the function $Integer$ returns the integer part of the product of $LimMax$ and $realData$.

III. DATA VALIDATION AND RECONSTRUCTION

To validate and reconstruct the data from the flow meter sensors of the distribution network, first we need to model the behavior of the corresponding time series. To model the water demand time series, as explained before we use an ANN. With the model of the time series, it is possible to start the process of validation and reconstruction of data. The process starts from the acquisition of data, and ends until they are stored in the operational database.

The information from flow meter is sent to the telecontrol center. If system does not receive the data, the process of reconstruction is started. If the data is received, the data goes to a validation process. If the data is out of the boundary of a valid range, the data is discarded and the process of reconstruction is started. After this processes the data are stored in the operational database. Figure 3 shows the flow of the validation process and reconstruction of information.

The validation process verify if the data is between a valid boundary. The valid boundary is defined by the model obtained with the ANN. The model produces an estimation, and with statistical methods a valid threshold can be determined. In this case, to quantify the model uncertainty, we use the confidence interval assuming that error follows normal distribution as follows:

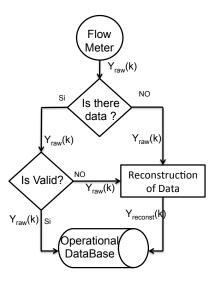$$\bar{x} - 1.96\frac{\sigma}{\sqrt{N}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{N}} \qquad (4)$$

Fig. 3. Validation and reconstruction process of the flow meter data sensor.

where μ is interval with a 95% confidence, $\bar{x}$ is the mean of the error, σ is the standard deviation, and N is length of the population. The valid range is defined as:

$$y(k) \ \epsilon \ [\tilde{u}_{sup}, \ \tilde{u}_{inf}] \tag{5}$$

where $\tilde{u}_{sup} = \bar{x} - 1.96\frac{\sigma}{\sqrt{N}}$ and $\tilde{u}_{inf} = \bar{x} + 1.96\frac{\sigma}{\sqrt{N}}$.

In the data reconstruction process, the ANN model is already defined (trained). According to the architecture of the ANN, n last observations are selected, and the estimation can be carried out. This estimation is sent to the operational database.

IV. CASE STUDY

The Barcelona network is managed by the company Aguas de Barcelona (AGBAR) [27]. This company does not only supply water to Barcelona city, but also to the metropolitan area. The network supplies 23 Municipalities in a 424 km^2 area with 4,645 km of pipes in order to meet about 3 million people water demand.

The sources of water are the rivers Ter and Llobregat that are regulated at the head by some dams with an overall capacity of 600 hm^3. There are four drinking water treatment plants which treat the underground flows. Also, there are several underground sources (wells) that can provide water through pumping stations.

The different water sources currently provide a flow of around 7 m^3/s. Currently, a desalter plant has just been put in production with a capacity of 60 $hm^3/year$. This plant is located at the end of the Llobregat river and produces drinking water by treating the sea water through a desalinization process. This plant will become of prime importance especially in drought periods, helping to maintain the water supply.

The complete transport network contains: 63 storage tanks, 3 surface sources and 7 underground sources, 79 pumps, 50 valves, 18 nodes, and 88 demand sectors. The network is controlled through a SCADA system with sampling periods

of 1 hour. For the predictive control scheme a prediction horizon of 24 hr is chosen. This record is updated at each time interval.

Due to the geographical topology of Barcelona and its surroundings, the water network supply in the metropolitan area is structured in pressure flows. Indeed, the topology of Barcelona, with a big difference in height between the sea level and the highest point to be supplied that is about 500 m above the sea level makes it necessary to homogenise the pressure by intermediate tanks and pumping stations.

The water supply system responds to changes in network topology (e.g. ruptures), typical daily profiles, as well as major changes in water demand, etc. The demand varies hourly, following a pattern that repeats every day. There are slight changes during weekends; additionally, there are changes in demand between winter and summer.

V. RESULTS

The considered water demand flow meter sensor is named $p10025$ of Barcelona water distribution network. The time serie of this sensor of one month in hourly scale is depicted in the Figure 4.
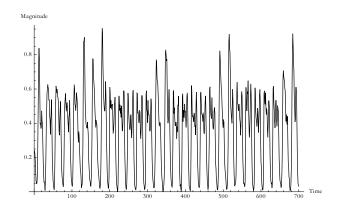

Fig. 4. The time serie of the sensor p10025 of one moth in hourly scale.

The first step is to model the dynamic behaviour of water demand by using an ANN. We use GA to learn the best ANN architecture. The parameters used to perform the experiments performed are shown in Table I. After the process of defining the best architecture of the net and the training process, an ANN capable of model the behavior of the time series is obtained. Figure 5 shows the results of the modelling process, by plotting the real data versus the prediction data resulted form the obtained ANN model using the GA optimization process.

The telecontrol system manages the information received from the flow meter sensor. If the system do not receive any data because problems with communications or the sensor; it starts the process of reconstruction. An example of this case is presented in the Figure 6 where a bar diagram is presented. Each bar represents the magnitude of each time interval. From the diagram we observe some gaps that correspond to the missing information.
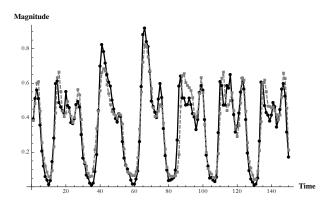
Fig. 5. Real versus predicted demand using the ANN obtained using the GA optimisation process.

TABLE I

PARAMETERS OF THE GA

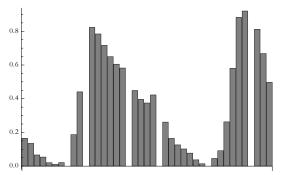| Concept	Value
Length of data	720
Length of the validation set	216
inputs of the ANN	35
neurons in the hidden layer	28
Individual mutation probability	35%
Gen mutation probability	0.05%
Crossover probability	70%

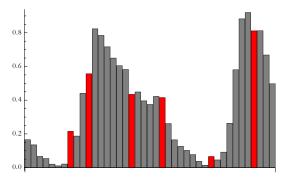Fig. 6. Gaps in the log history of the sector p10025.

Fig. 7. Missing data replaced by the reconstruction process. The grey bars represent the real data magnitude and the the red ones represent the reconstructed data.

To deal with this problem, the system uses the reconstruction process to replace the missing data. Reconstruction process uses the ANN model to obtain the prediction of the missing information. The model already trained uses a past information and produce the prediction. Figure 7 shows the result after applying the reconstruction process to the missing data shown in Figure 6.

A second problem is the false data produced by problems in the flow meters. Once the tele-control system has received the data, the system validates them. If its out of the boundary (5), a false data is detected. Then, the system starts the reconstruction process (explained before). And, after the reconstruction, the data is inserted in the operational database. The boundary (confidence intervals) of this problem are presented in Figure 8. In this figure, there are four lines, the black continuous one represents the real data; the dotted one represents the prediction produced by the ANN model and the two thinner dotted ones represent the valid boundary of the model.

An example of false data produced by a flow meter is depicted in the Figure 9. From this figure, two points leave the boundary of the CIs, these two anomalous measurements are replaced by reconstruction process.

VI. CONCLUSIONS

In this paper, we have presented a methodology to detect abnormalities in the flow meter reading, as missing data or false data. We model the time series dynamics with an ANN, where its architecture and learning process is based on solving an optimisation process using GA. The missing data are replaced by data obtained from the ANN model, and the false data are detected if it leaves the boundary defined by the CIs associated to the ANN model. Detection and correction of false data could avoid the water distribution company problems in the billing process, or the operational process, just to mention some examples. The proposed approach has satisfactorily assessed using some flow meters of the Barcelona water distribution network.

REFERENCES

[1] V. Puig, M. Witczak, F. Nejjari, J. Quevedo, J. Korbicz, A gmdh neural network-based approach to passive robust fault detection using a constraint satisfaction backward test, Engineering Applications of Artificial Intelligence 20 (7) (2007) 886–897.

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

[3] J. Liu, H. H. Savenije, J. Xu, Forecast of water demand in weinan city in china using wdf-ann model, Physics and Chemistry of the Earth, Parts A/B/C 28 (4?5) (2003) 219 – 224, water Resources Assessment for catchment management. doi:http://dx.doi.org/10.1016/S1474-7065(03)00026-3.
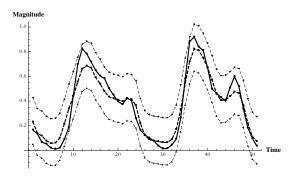
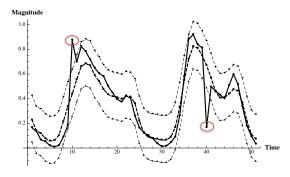Fig. 8.   Confidence intervals or valid boundary of the model.

Fig. 9.   False data detected by the boundary of the CIs

[4] A. Jain, A. Kumar Varshney, U. Chandra Joshi, Short-term water demand forecast modelling at iit kanpur using artificial neural networks, Water Resources Management 15 (5) (2001) 299–321. doi:10.1023/A:1014415503476.
URL http://dx.doi.org/10.1023/A

[5] A. Jain, A. K. Varshney, U. C. Joshi, Short-term water demand forecast modelling at iit kanpur using artificial neural networks, Water Resources Management 15 (5) (2001) 299–321.

[6] A. Bakirtzis, V. Petridis, S. Kiartzis, M. Alexiadis, A. Maissis, A neural network short term load forecasting model for the greek power system, Power Systems, IEEE Transactions on 11 (2) (1996) 858–863.

[7] J. J. Flores, M. Graff, H. Rodriguez, Evolutive design of arma and ann models for time series forecasting, Renewable Energy 44 (2012) 225–230.

[8] J. J. Flores, H. Rodriguez, M. Graff, Reducing the search space in evolutive design of arima and ann models for time series prediction, in: Advances in Soft Computing, Springer, 2010, pp. 325–336.

[9] J. J. Flores, R. Loaeza, H. Rodríguez, E. Cadenas, Wind speed forecasting using a hybrid neural-evolutive approach, in: MICAI 2009: Advances in Artificial Intelligence, Springer, 2009, pp. 600–609.

[10] K. Benmouiza, A. Cheknane, Forecasting hourly global solar radiation using hybrid k-means and nonlinear autoregressive neural network models, Energy Conversion and Management 75 (2013) 561–569.

[11] M. Kumar, N. R. Patel, Using clustering to improve sales forecasts in retail merchandising, Annals of Operations Research 174 (1) (2010) 33–46.

[12] F. Martinez Alvarez, A. Troncoso, J. Riquelme, J. Riquelme, Partitioning-clustering techniques applied to the electricity price time series, in: Intelligent Data Engineering and Automated Learning-IDEAL 2007, Springer, 2007, pp. 990–999.

[13] J. Quevedo, J. Saludes, V. Puig, J. Blanch, Short-term demand forecasting for real-time operational control of the barcelona water transport network, in: Control and Automation (MED), 2014 22nd Mediterranean Conference on Control and Automation (MED) University of Palermo. June 16-19, 2014. Palermo, Italy, 2014, pp. 990–995. doi:10.1109/MED.2014.6961503.

[14] J. D. Hamilton, Time series analysis, Vol. 2, Princeton university press Princeton, 1994.

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

[16] J. Contreras, R. Espinola, F. J. Nogales, A. J. Conejo, Arima models to predict next-day electricity prices, Power Systems, IEEE Transactions on 18 (3) (2003) 1014–1020.

[17] A. C. Harvey, Forecasting, structural time series models and the Kalman filter, Cambridge university press, 1990.

[18] R. T. Olszewski, Generalized feature extraction for structural pattern recognition in time-series data, Tech. rep., DTIC Document (2001).

[19] T. W. Liao, A clustering procedure for exploratory mining of vector time series, Pattern Recognition 40 (9) (2007) 2550 – 2562. doi:http://dx.doi.org/10.1016/j.patcog.2007.01.005.

[20] T. W. Liao, Clustering of time series data?a survey, Pattern recognition 38 (11) (2005) 1857–1874.

[21] M. S. Al-Hafid, G. H. Al-maamary, Short term electrical load forecasting using holt-winters method, Al-Rafadain Engineering Journal 20 (6).

[22] S. Zhou, T. McMahon, A. Walton, J. Lewis, Forecasting daily urban water demand: a case study of melbourne, Journal of Hydrology 236 (3?4) (2000) 153 – 164. doi:http://dx.doi.org/10.1016/S0022-1694(00)00287-0.

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

[24] G. Zhang, B. E. Patuwo, M. Y. Hu, Forecasting with artificial neural networks:: The state of the art, International journal of forecasting 14 (1) (1998) 35–62.

[25] R. Lopez, V. Puig, H. Rodriguez, An implementation of a multi-model predictor based on the qualitative and quantitative decomposition of the time-series, International work-conference on Time Series 1 (1).

[26] Neurolab, Neurolab a python plugin, Available: https://pypi.python.org/pypi/neurolab [Online].

[27] S. Sociedad General de Aguas de Barcelona, Annual corporate governance report (2008).