24-Hours Demand Forecasting Based on SARIMA and Support Vector Machines

Mathias Braun, Thomas Bernard, Olivier Piller, Fereshte Sedehizade

Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 89 (2014) 926 – 933

Procedia Engineering

www.elsevier.com/locate/procedia

16th Conference on Water Distribution System Analysis, WDSA 2014

24-Hours Demand Forecasting Based on SARIMA and Support Vector Machines

M. Braun[a],*, T. Bernard[a], O. Piller[b] and F. Sedehizade[c]

[a]*Fraunhofer IOSB, Fraunhoferstr. 1, D-76131 Karlsruhe (Germany)*
[b]*Irstea, 50, avenue de Verdun, F-33612 Cestas (France)*
[c]*Berliner Wasserbetriebe, Neue Jüdenstr. 1, D-10179 Berlin (Germany)*

Abstract

In time series analysis the autoregressive integrate moving average (ARIMA) models have been used for decades and in a wide variety of scientific applications. In recent years a growing popularity of machine learning algorithms like the artificial neural network (ANN) and support vector machine (SVM) have led to new approaches in time series analysis. The forecasting model presented in this paper combines an autoregressive approach with a regression model respecting additional parameters. Two modelling approaches are presented which are based on seasonal autoregressive integrated moving average (SARIMA) models and support vector regression (SVR). These models are evaluated on data from a residential district in Berlin.

Keywords: SARIMA, support vector machine, water distribution network, water demand, prediction, forecasting, calibration;

1. Introduction

Water distribution Networks (WDNs) are critical infrastructures that are exposed to deliberate or accidental contaminations. Water suppliers are installing sensors for water quality and water quantity throughout their networks. These sensors create a huge amount of data that allows to develop a system that is able to monitor and protect WDNs in real-time. The objective of the SMaRT-Online^{WDN} management toolkit [5] is to give fast and high quality decision support in case of a contamination. The online security management toolkit is based on sensor measurements

* Corresponding author. Tel.: +45-721-6091-251.
E-mail address: mathias.braun@iosb.fraunhofer.de

doi:10.1016/j.proeng.2014.11.526

(hydraulic and water quality sensors) and online hydraulic and transport models and allows an estimation of the localization of the contamination source and the simulation of short-time future scenarios in order to estimate the impact of a contamination source. In order to make an accurate short-term prediction for contamination scenarios a good prediction for the hydraulic state of the network is needed. The hydraulic state is strongly influenced by the water demand in the observed area that poses the need for an accurate prediction and calibration of the water demand [6, 7]. Due to the fact that water demand is subject to high seasonality it is prudent to use the tools of time series analysis to model the water demand and predict the consumption in a district or consumption nodes of the network.

The objective of the models presented in this paper is the prediction of the water demand for the pressure zone östliche Hochstadt in Berlin. The method allows to predict the water demand in hourly steps from one hour up to 24 hours in the future. Further, it is possible to determine the demand pattern for the next 24 hours.

The paper is structured as follows. The first chapter gives a short view on literature, the area for the case study and on common features in water demand prediction. Chapter 2 presents the two modelling approaches used in this paper followed by the detailed description of the modelling process. In chapter 3 the results of the demand forecasting procedures are presented and evaluated. Chapter 4 gives a conclusion and a look on further prospects for the research.

1.1. Literature

In the context of water demand forecasting, a wide variety of methods have come to use ranging from the classical time-series approach with ARIMA type models, that are commonly used in statistics and econometrics, to a number of algorithms from statistical and machine learning. In [1] Zhou et al. present a time series model for the prediction of the daily water consumption in Melbourne, Australia. The model is formulated by a set of equations that consider four main factors on water demand given by long term trend, seasonality, climatic correlation and autocorrelation. Preis et al. present a different approach with the M5 model tree algorithm in [7]. The objective of the model is the pre-estimation of the water demand in 24 hours. This forecast is used in a predictor-corrector procedure for the online hydraulic state estimation in urban water networks. The main factors considered for modelling the demand prediction are past demand values that consider the daily and the weekly demand pattern. In [8] Ischmael et al. apply several models from statistical and machine learning theory to the task of water demand forecasting. They use two different artificial neural networks and support vector regression to predict the daily water demand on the basis of the previous daily water demands and the annual estimated population of the province. Aijun et al. [4] concentrate on generating prediction rules from observed data for the demand prediction. Since the evaluation of interviews with experienced operators lead to inadequate results for the forecast, alternative methods for knowledge acquisition have been applied through data mining. The paper present an application of rough-set approach for the automated discovery of rules from measurement data.

Due to the increasing popularity of methods from statistical learning theory in the prediction and forecasting of water resource variables Maier and Dandy published a paper modelling issues and application [9]. They outline the major steps that should be followed in the development of such models to help in the development of guidelines for the application of such models.

1.2. Data and Features

The dataset used for the development of the demand prediction models presented in this paper has been supplied by the Berliner Wasserbetriebe (BWB). The BWB supplies drinking water to the city area of Berlin through a water distribution network with a length of 9,500 kilometers. To keep a constant pressure throughout the city area it is divided into the 4 pressure zones *"Nördliche Hochstadt", "Östliche Hochstadt (with pressure zone Buch)", "Tiefstadt" and "Südliche Hochstadt"* [10].

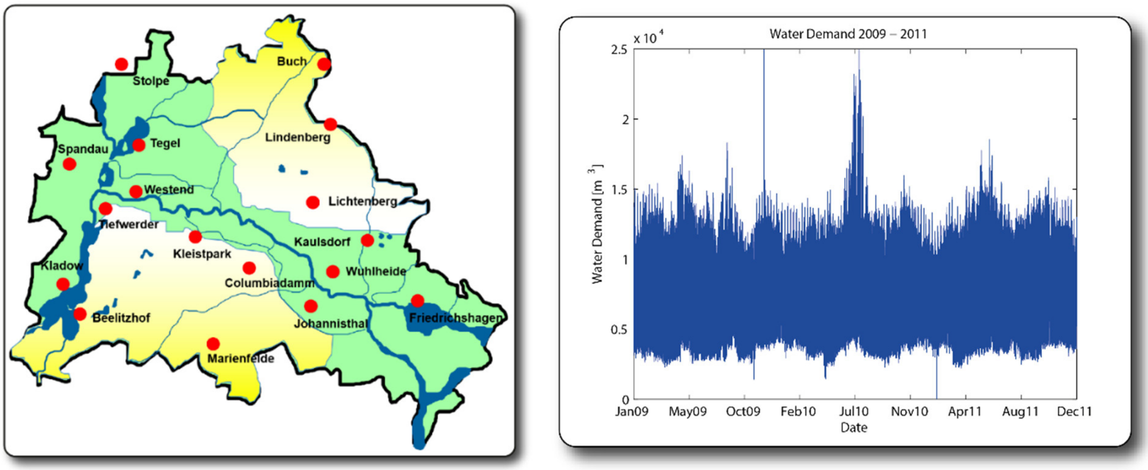

Fig. 1. (a) City area of Berlin with main pressure zones; (b) Agglomerated water demand for östl. Hochstadt from 2009 until 2011.

For this study, the inflow and outflow values of the pressure zone östliche Hochstadt were available. This zone contains the districts Stolpe, Buch, Lindenberg and Lichtenberg. The data is given over a period of nearly 3 years. Fig. 1(b) displays the agglomerated water demand for the area for the years 2009, 2010 and 2011. The demand data is sampled with a sample rate of 1/hour. From these values the net value of the agglomerated water demand for the district are calculated. In effect, we have a dataset of over 25,000 samples for the modelling process at our disposal. The water demand in an urban area is determined by industrial, commercial, public and domestic costumers. Their consumption is influenced by factors like the hour of the day, day of the week, seasonal variations, whether a day is a holiday or school holiday and the weather conditions. A selection of these factors that are used as regression factors in addition to the past demand values is given by the hour of the day, day of the week, the month and an indicator for weekends and holidays.

2. Forecasting Models

2.1. Modelling Approaches

In this section we give a basic look on the modelling approaches presented in this paper. These are the classical time series approach with the ARIMA model and Support Vector Regression representing machine learning methods. ARIMA models are a general class of models used to fit time series data for the better understanding and forecasting of a non-stationary process. They have been used in a wide variety of applications ranging from statistic and econometrics to demand prediction in WDNs. Although the method itself was developed much earlier the systematic approach for applying the technique has been published by Box and Jenkins in 1976 [11]. ARIMA stands for Auto-Regressive (AR) Integrated (I) Moving-Average (MA). The autoregressive part describes correlations between the current and past values. Non stationary effects in a signal can be modelled by the integrated part and the moving average models dependencies on the errors of past values. There are a number of extensions to the method one of which are SARIMA models that consider seasonal effects by using the values from past periods. In this paper the SARIMA model gives a basic performance to compare other models against. More information on statistical time series modelling is given in [12].

The second method used for water demand forecasting presented in this paper is based on Support Vector Machines (SVMs). Support Vector Machines are supervised learning algorithms from statistical learning theory that are used to recognize patterns in data for classification. Support Vector machines for function estimation are also referred to a Support Vector Regression (SVR). In the work presented in this paper the ε-SVR is used. The goal of ε-SVR is to find a function that has a maximum distance of ε from the actual measurement points y_i and at the same time is as flat as possible. For points that do not fall within the margin ε an additional cost is added proportional to its distance from the margin. These points define the support vectors. The model has two independent parameters *C* and ε that have to be adjusted for fitting the model to the data. The cost factor $C > 0$ which determines the trade-off between the flatness of the regression line and the distance to the margin ε that is tolerated. The SVR is linear process, however by applying kernel functions the feature space can be transformed and the same method can be used for nonlinear problems. The kernel function can be defined by polynomial functions, Gaussian radial basis, hyperbolic tangent and many more. In the following the radial basis Gaussian kernel will be used.

For the results presented in this paper have been generated using a library for support vector machines called *"libsvm"* [13] that is available for a number of different platforms. For further information on the subject of Support Vector Regression the reader is referred to the paper of Smola and Schölkopf [14] that gives a detailed introduction to the method.

2.2. Modelling ARIMA

ARIMA models consist of three basic parts each of which has its own modelling procedure. Since ARMA models are modelling stationary processes the first step is to test the time series for stationarity. In statistics the Dickey-Fuller-Test is used to test if a time series is stationary. It tests the null hypothesis of a stochastic process with a unit root against an alternative process without a unit root. This makes it possible to determine if the process has integrating properties. For the water demand data from the Berlin area shown in Fig. 1(b) no non-stationary behavior is detectable. To make sure even small effects do not influence the prediction model a log-transformation is performed on the data.

After ensuring that the time series is stationary the next step is to determine the AR and MA factors. Therefore two helpful tools are given by the autocorrelation function (ACF) and the partial autocorrelation function (PACF). The autocorrelation function is the correlation of the signal with itself and can be used to find reoccurring patterns in a signal. It is given by:

$$ACF(k) = \rho_k = \frac{Cov(y_t, y_{t-k})}{Var(y_t)} \tag{1}$$

Fig. 2(a) shows the autocorrelation factors for lag values up to t -168 which represents a full week. The ACF has three major modes of behavior concerning ARIMA models. In AR models the ACF tails off with a factor of γ^k. Here γ is the value of the first factor from the ACF and k is the lag. For MA models the ACF cuts off after the relevant factors for the time series. For non-seasonal models the behavior shown in Fig. 2(a) could indicate a non-stationary time series. Since we made sure in the first step that our signal is stationary the ACF confirms the intuition that the data is strongly influenced by daily and weekly seasonality.

The partial autocorrelation function (PACF) is a second important tool for the model identification in ARIMA modelling. The partial autocorrelation at a lag *k* is the autocorrelation between y_t and y_{t-k} that is not explained by the lags in between. The PACF is defined as proportion of the conditional covariance of y_t and y_{t-k} and their conditional variances as follows:

$$PACF(k) = \rho_k^* = \frac{Cov(y_t, y_{t-k} | y_{t-1}, y_{t-k+1})}{\sqrt{Var(y_t | y_{t-1}, y_{t-k+1}) Var(y_{t-k} | y_{t-1}, y_{t-k+1})}} \tag{2}$$

Fig. 2(b) shows the result for the PACF performed on the data set. The maximum lag is again chosen to be t - 168. It is apparent that apart from the direct dependence on the last few hours especially the past demands from one day ago and a week ago have a significant correlation to the actual demand. Based on the analysis of the time series in this chapter a first step for the demand prediction is a $SAR(2)_{(24)x(168)}$ model. Similar to the model in [7] it is based on two past demand values for the current hour, a day before (24 hours) and the week before (168 hours).

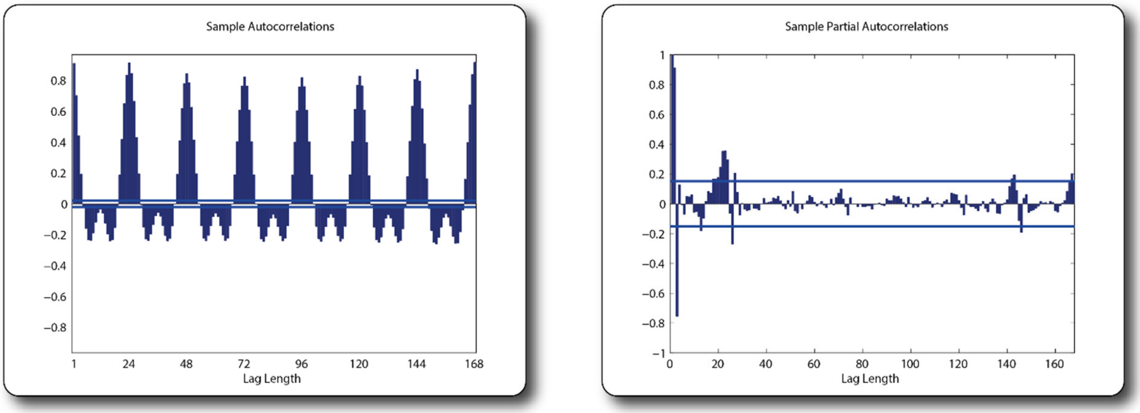

Fig. 2. (a) Auto-correlation factors for the water demand time series; (b) Partial auto-correlation factors for the water demand time series.

2.3. Modelling Support Vector Regression

This section covers the steps taken for the modelling process of the SVR model. Following the results of the analysis done for the ARIMA model the first attempt for modelling the SVR is done with the same feature set. A guideline for the development steps of models from machine learning theory is given by [9]. An important phase is model validation. Therefore the dataset is divided into a training set that contains about 60% of the samples in a randomized order and a validation set that contains the other 40% of the samples. All subsequent models are trained with data from the training set and tested on the validation set to evaluate its accuracy and generalizing properties. The first modelling step is to determine the values of the model parameters the margin ε and the cost factor *C*. To determine these the model is trained with the training set for a selection of values for $\varepsilon_p = C_p = [0.001\ 0.01\ 0.1\ 1\ 10\ 100]^T$. The results are evaluated with the validation set and the procedure is performed again in a smaller area for the parameters. The parameters determined for the model in this paper are given by $C = 9.4$ and $\varepsilon = 0.001$. The standard deviation of the kernel function is set to a value of $\gamma = 1/N_{features}$.

The next step is the evaluation of the fitting and generalization of the model. Fig. 3(a) shows the learning curves for the 1 hour prediction models of the SVR. The learning curves Fig. 3(a) and (b) show the mean relative error of the training set and the validation set as a function the number of training samples used for the training of the model. The training set is used to fit the SVR to the data. The validation set is used to test the generalizing properties of the model. For a small size of the training set we can see that the training error is small or even zero since in this case the model is over fitting the data. As a result the model has bad generalization for the validation set. With a higher number of training samples the generalization is getting better which can be seen by the decreasing validation error but as an effect the training error rises.

The model inFig. 3(a) uses five autoregressive water demand parameters with the lags *t*-1, *t*-24, *t*-25, *t*-168 and *t*-169. As we can see the model has a high learning rate in the beginning but the learning rate decreases fast and both the training error and the validation error show minimal change for more than 5000 training samples. In general this is a sign that the model has converged. One possibility to improve the performance at this point is to add further features to the model.

As we have seen in the PACF in Fig. 2(b) the autoregressive demand values selected for the model inFig. 3(a) are not the only significant lags. Due to the generous amount of data samples a second SVR model is set up that has 168 autoregressive features. This means that all demand values from the past week are used in the model.Fig. 3(b) shows the learning curve for the extended model and shows that the learning curve does not converge as fast as for the smaller model, but results in a smaller training and validation error. Still the learning curve shows no further improvement for a training set with more than 10.000 samples. This means additional features could be applied to improve the model.

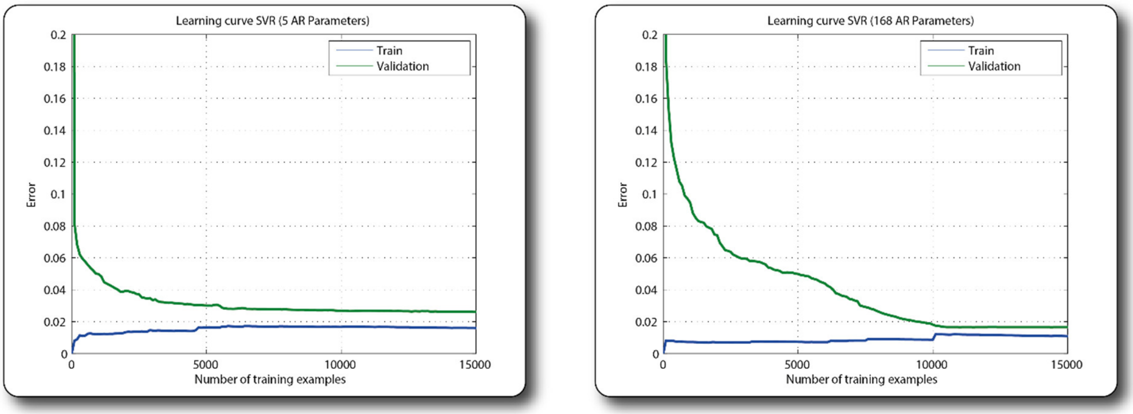

Fig. 3. (a) Learning curve for SVR model with 5 autoregressive parameters; (b) Learning curve for SVR model with 168 autoregressive parameters.

3. Results

The results discussed in this section come from four different models. They are a SAR and a SVR model for the demand prediction 1 hour in advance and a SAR and a SVR model for the prediction 24 hour in advance. The evaluation is done on the relative error of the water demands forecasts. A more significant evaluation of the performance of the 1 hour prediction models is given in Fig. 4. Fig. 4(a) shows the histogram of the relative errors of the hourly demand prediction for all points from the year 2011 from the linear regression model. It indicates that 67% of the estimates fall within the relative error band of ±8% and 95% within a range of ±20%. Fig. 4(b) is the histogram for the relative errors of the SVR model. From the Fig. 4(b) it seems that the errors are smaller than for the regression model. The percentiles show that 67% of the estimates fall within the relative error band of ±5% and 95% within a range of ±13%.

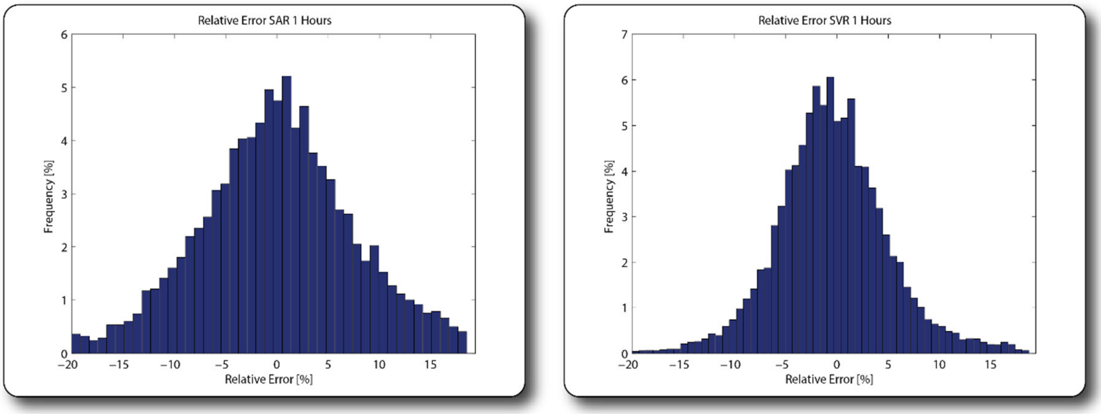

Fig. 4. (a) Relative errors for the SAR model with a 1 hour prediction span; (b) Relative errors for the SVR model with a 1 hour prediction span.

Errore. L'origine riferimento non è stata trovata. gives the relative error histograms for the SAR and SVR prediction models with a prediction span of 24 hours again for the data from summer 2011. Fig. 5(a) shows the results for the regression model and it is apparent that the errors increased with the prediction span. 67% of the estimates still fall within the relative error band of ±8% but the error band for 95% expand to a range of about ±23%. For the 24 hour span SVR the performance decreases as well. The percentiles show that 67% of the estimates fall within the relative error band of ±7% and 95% within a range of ±19% but the support vector regression still outperforms the linear regression model.

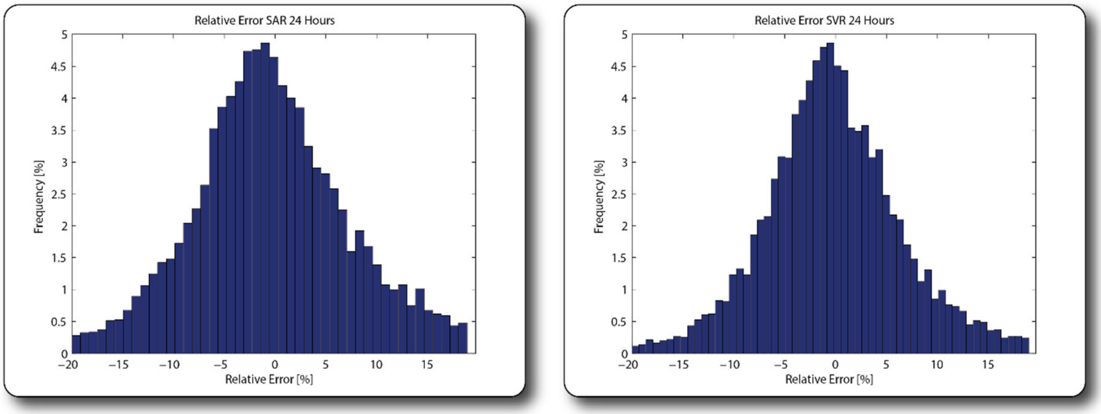

Fig. 5. (a) Relative errors for the SAR model with a 24 hour prediction span; (b) Relative errors for the SVR model with a 24 hour prediction span.

The performance differences can be explained by the methods itself since the SAR model is a linear modelling approach and the SVR models nonlinear effects with a Gaussian kernel transformation. Table 1 gives the exact values of the percentiles of the different models. In addition, it shows the correlation coefficient between the actual demand values and the models. The relative errors are also evaluated by the ACF and the PACF with the result that the errors of none of the model are white noises. This indicates that there are still effects in the data that are not explained by the selected features.

4. Conclusions and further work

As the results in section 3 show, both procedures for the prediction of the water demand give good results. The models are adaptable for different prediction spans reaching from 1 hour to 24 hours in the future. The method chosen in the paper allows for even longer prediction spans. Expectedly, the prediction error rises for longer prediction spans since information for the immediate past strongly influence the water demand of the subsequent hours. In general, the support vector regression model performed better than the seasonal autoregressive model. This is in part because the SAR model is based on linear regression and the SVR model uses Gaussian kernels to model nonlinear effects. The modelling process for the SVR showed that through the addition of further autoregressive features the accuracy of the prediction was improved by about 1%. Still the evaluation of the relative errors of all models indicated that the errors are not made up of white noise, so there is still a margin to improve the model. Literature suggests that meteorological features like rainfall, relative humidity and cloud amount have an influence on the water demand that is not modeled in this paper.

For future work there are two main issues that have to be addressed. The first one is longer term prediction. This means for one the uncertainty of the prediction with rising prediction spans and the prediction of trends. An approach for Support Vector Machines is incremental learning. This allows the model to adapt to changing conditions and improve the demand forecast for changing conditions. The second issue is the prediction of the water demands for smaller areas in the city. One of the reasons for the good performance of the models presented in this paper is that we observe the consumption gathered for a relatively large area. This means that fluctuations in the individual behavior are not as important for the water demand. In concentrating on smaller areas the uncertainty of the prediction rises.

Table 1. Percentiles of relative errors for the prediction models.

Percentiles	67%	95%	99%	CC
SAR 1 Hour	±7.68%	±19.29%	±29.50%	0.95
SVR 1 Hour	±4.81%	±12.24%	±20.84%	0.97
SAR 24 Hour	±7.93%	±23.11%	±39.05%	0.93
SVR 24 Hour	±6.37%	±18.19%	±34.84%	0.95

Acknowledgements

The work presented in the paper is part of the French-German collaborative research project SMaRT-OnlineWDN that is funded by the French National Research Agency (ANR; project: ANR-11-SECU-006) and the German Federal Ministry of Education and Research (BMBF; project: 13N12180)

References

[1] S. L. Zhou, T. A. McMahon, A.Walton, J. Lewis. Forecasting daily urban water demand: a case study of Melbourne. J. of Hydrology, 236(3) (2000) 153-164.

[2] S. P. Zhang, H. Watanabe, R. Yamada. Prediction of daily water demands by neural networks. In Stochastic and statistical methods in hydrology and environmental engineering. Springer Netherlands (1994) 217-227.

[3] J. E. van Zyl, O. Piller, Y. le Gat. Sizing municipal storage tanks based on reliability criteria. J. of Water Resources Planning and Management, 134(6) (2008) 548-555.

[4] A. An, N. Shan, C. Chan, N. Cercone, W. Ziarko. Discovering rules for water demand prediction: an enhanced rough-set approach. Engineering Applications of Artificial Intelligence. 9(6) (1996) 645-653.

[5] SMaRT-OnlineWDN (2013). "http://www.smart-onlinewdn.eu/", December 12, 2013.

[6] O. Piller, I. Montalvo, J. Deuerlein, M. Braun, H. Ung, H. A gradient-type method for real-time state estimation of water distribution networks. International Conference on Hydroinformatics, 2014.

[7] A. Preis, A. J. Whittle, A. Ostfeld, L. Perelman. Efficient hydraulic state estimation technique using reduced models of urban water networks. J. of Water Resources Planning and Management, 137(4) (2010) 343-351.

[8] I. S. Msiza, F. V. Nelwamondo, T. Marwala,. Artificial neural networks and support vector machines for water demand time series forecasting. In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference, IEEE (2007, October) 638-643.

[9] H. R. Maier, G. C. Dandy. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software, 15(1) (2000) 101-124.

[10] K. Möller, J. Burgschweiger: Wasserversorgungskonzept für Berlin und das von der BWB versorgte Umland (Entwicklung bis 2040). 1. September 2008, S. 4, archiviert vom Original am 5. Mai 2011, abgerufen am 5. Mai 2011 (pdf; 2,8 MB, deutsch). http://www.stadtentwicklung.berlin.de/umwelt/wasser/download/wvk2040.pdf

[11] O. D. Anderson. Time series analysis and forecasting: the Box-Jenkins approach. London and Boston: Butterworths (1976) 182.

[12] R. H. Shumway, D. S. Stoffer. Time series analysis and its applications. New York: Springer (2000) (Vol. 3).

[13] libsvm (2014). " http://www.csie.ntu.edu.tw/~cjlin/libsvm/ ", Mai 1, 2014.

[14] A. J. Smola, & B. Schölkopf, (2004). A tutorial on support vector regression. Statistics and computing, 14(3), 199-222.