engineering proceedings

MDPI

Proceeding Paper

A Study on Short-Term Water-Demand Forecasting Using Statistical Techniques †

Jungwon Yu *, Hyansu Bae, Mi-Seon Kang, Kwang-Ju Kim and In-Su Jang

Daegu-Gyeongbuk Research Division, Electronics and Telecommunications Research Institute, Daegu 42994, Republic of Korea; baehs@etri.re.kr (H.B.); tams37@etri.re.kr (M.-S.K.); kwangju@etri.re.kr (K.-J.K.); jef1015@etri.re.kr (I.-S.J.)

* Correspondence: gardenyoo@etri.re.kr

† Presented at the 3rd International Joint Conference on Water Distribution Systems Analysis & Computing and Control for the Water Industry (WDSA/CCWI 2024), Ferrara, Italy, 1–4 July 2024.

Abstract: This paper proposes a method for short-term weekly water-demand forecasting combining various statistical techniques. In the proposed method, training datasets are prepared through exploratory data analysis, several data preprocessing steps, and an input selection step; also, forecasting models are constructed by support vector regression. After this, weekly water-demand forecasts are calculated using iterated and direct strategies. To verify the performance, the proposed method is applied to urban hourly water-demand datasets provided by the Battle of Water Demand Forecasting organized in the 3rd WDSA-CCWI Joint Conference.

Keywords: water demand; short-term forecasting; support vector regression; iterated strategy; direct strategy

Citation: Yu, J.; Bae, H.; Kang, M.-S.; Kim, K.-J.; Jang, I.-S. A Study on Short-Term Water-Demand Forecasting Using Statistical Techniques. *Eng. Proc.* **2024**, *69*, 154. https://doi.org/10.3390/engproc2024069154

Academic Editors: Stefano Alvisi, Marco Franchini, Valentina Marsili and Filippo Mazzoni

Published: 20 September 2024

1. Introduction

This paper proposes a method for short-term weekly water-demand forecasting combining various statistical techniques. In the proposed method, firstly, we visually examine the effects of several meteorological factors on water demands through exploratory data analysis (EDA) and then select those that can be employed as inputs for forecasting models. Next, several data preprocessing steps such as missing value imputation, data normalization, data transformation, and data selection are performed. After this, among the initial input candidates composed of demand values measured from past to present, we only select those that are highly correlated with future demand values. Outlier detection and removal are also carried out by considering correlations between the selected inputs and output. After preparing training datasets, forecasting models are constructed using support vector regression (SVR), and finally, weekly water-demand forecasts are calculated using iterated and direct strategies that have been widely used for multi-step ahead forecasting. To verify the performance, the proposed method is applied to hourly urban water-demand datasets provided by the Battle of Water Demand Forecasting (BWDF) organized in the 3rd International WDSA-CCWI Joint Conference in Ferrara, Italy.

2. Proposed Method

Figure 1 describes the proposed short-term weekly water-demand forecasting procedure. Step 1 is to determine which of four weather factors, including rainfall depth (mm), air temperature (°C), air humidity (%), and windspeed (km/h), can be used as input variables for forecasting models; to do this, scatter plots between each of the four factors and net inflow (L/s) of a target area are examined. From these plots, we confirmed that there exist clear correlations between air temperature and net inflow; thus, it makes sense to employ air temperature as one of input variables. Step 2 is to generate one-hot encoded input vectors $\mathbf{h} \in \mathbb{R}^{24}$ and $\mathbf{w} \in \mathbb{R}^{7}$ to reflect daily and weekly periodicities in water demands

in forecasting models. Step 3 is to perform several data preprocessing steps such as missing value imputation, max-min normalization, data transformation, and data selection. Step 3-1 is to impute missing values. In Step 3-2, max-min normalization is carried out to convert air temperature and water-demand values to lie in [0, 1]. Step 3-3 transforms one-dimensional water-demand time series into multidimensional matrices. Step 3-4 is to select a portion of entire data vectors for model training to consider seasonality in water demands; the selected data vectors correspond to the days that are seasonally similar to evaluation weeks. In other words, selected data vectors correspond to the $7 \times w$ days immediately before the evaluation weeks, the days falling on the same dates as the evaluation weeks in past years, and the $7 \times w$ days immediately before and after these dates. The proper values of window size w can be determined by prior experiments; in this paper, the value of w is set to five.

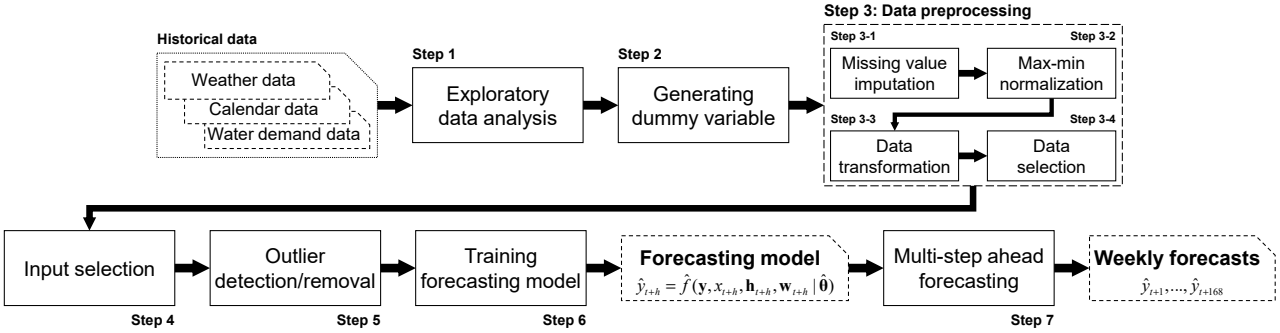

Figure 1. Proposed weekly water-demand forecasting procedure.

Step 4 is to select significant inputs for forecasting models among historical water-demand values. As initial input candidates for a model whose output is a demand value y_{t+h} at future time $t + h$ (where h = 1, 24, or 168), we consider 168 measured demand values $y_t, y_{t-1}, \ldots, y_{t-167}$ from past time $t - 167$ to present time t. After calculating absolute values of correlation coefficients between the 168 candidates and the output and sorting them in descending order, the first d candidates are selected. Since water-demand datasets can be contaminated by statistical outliers, robust covariance matrices [1] are used to calculate the correlation coefficients. The proper number of inputs d can be decided by cross validation (CV) technique; in this paper, the value of d is set to six using CV. Step 5 is to detect and remove statistical outliers based on correlations between the selected inputs and the output. To do this, the Mahalanobis distance parameterized by robust covariance matrices [1] is applied. In Step 6, the following forecasting model is constructed using SVR [2]:

$$\hat{y}_{t+h} = \hat{f}\left(\mathbf{y}, x_{t+h}, \mathbf{h}_{t+h}, \mathbf{w}_{t+h} | \hat{\theta}\right), \quad (1)$$

where $\hat{y}_{t+h}$ is the forecasted water-demand value at future time $t + h$, $\mathbf{y}$ is the vector composed of d = 6 relevant inputs selected in Step 4, x_{t+h} is the value of air temperature at future time $t + h$, $\mathbf{h}_{t+h}$ and $\mathbf{w}_{t+h}$ are one-hot encoded input vectors at future time $t + h$, and $\hat{\theta}$ is the model parameter vector estimated based on training datasets. Step 7 is to forecast weekly water-demand values based on Equation (1) using multi-step ahead forecasting strategies.

To perform multi-step ahead forecasting, iterated and direct strategies have been widely used [3]. An iterated strategy computes multi-step ahead forecasts by feeding the forecasted values back into input values recursively; in general, this strategy requires a forecasting model with the output y_{t+1} (i.e., in Equation (1), h = 1). A direct strategy calculates multi-step ahead forecasts directly based on single or multiple models. In this strategy, to calculate multiple forecasts, we only use observed water-demand values as inputs without reusing forecasts as inputs. For example, to compute 168 weekly forecasts of hourly water demands, 168 models with outputs $y_{t+1}, y_{t+2}, \ldots, y_{t+168}$ can be employed, or a single model with output y_{t+168} can be employed. In the former, training 168 forecasting models takes much time; in the latter, since only a single model is trained and used for

forecasting, its accuracy can be degraded. A compromise between using a single model and using 168 models can also be used. In other words, we can train seven models with outputs y_{t+24}, y_{t+48},. . ., y_{t+168} (here, h is set to a multiple of 24) and then calculate weekly demand forecasts; based on the model with output y_{t+24}, 24 demand values y_{t+1},. . ., y_{t+24} are predicted, based on the model with output y_{t+48}, 24 demand values y_{t+25},. . ., y_{t+48} are predicted, and so on.

In this paper, we use three multi-step ahead forecasting strategies, including an iterated strategy based on a single model with output y_{t+1}, a direct strategy based on a single model with output y_{t+168}, and a direct strategy based on seven models with outputs y_{t+24}, y_{t+48},. . ., y_{t+168}.

3. Experimental Results

This section provides the results of applying the proposed method in Figure 1 to the BWDF dataset [4]. This dataset consists of a weather dataset, a calendar dataset, and an urban hourly water-demand dataset (hourly net inflow dataset) collected from 10 DMAs. In [4], the readers can find more detailed information regarding the BWDF dataset. The aim of this battle is to forecast weekly water-demand values for four evaluation weeks (W1, W2, W3, and W4) as accurately as possible. To evaluate the accuracy of weekly forecasts, the following three performance indices and their sum PI (=PI_1 + PI_2 + PI_3) are used:

$$PI_1 = \frac{1}{24}\sum_{i=1}^{24}|y_i - \hat{y}_i|, \quad PI_2 = \max\{|y_i - \hat{y}_i||i = 1, \ldots, 24\}, \quad PI_3 = \frac{1}{144}\sum_{i=25}^{168}|y_i - \hat{y}_i|, \quad (2)$$

where $\hat{y}_i$ (i = 1,. . ., 168) is the 168 weekly forecasts and y_i are the 168 actual demand values. In this paper, it is assumed that the demand values in the 7 days immediately before the evaluation weeks are unknown, and the aim is to forecast these values. Only the results for the 7 days immediately before W1 are presented, and those for the others are omitted due to space constraints.

Table 1 lists the performance indices obtained by applying the three strategies described in Section 2 to weekly demand forecasting for the 7 days immediately before W1. As confirmed from this table, the average values of PI for the two direct strategies (7.05 and 7.29) are lower than that for the iterated strategy (7.71); the direct strategy with the single model can obtain a lower averaged value of PI than the direct strategy with the multiple models by 0.24. It is also worthwhile to emphasize that the types of direct strategies that can obtain better performance indices vary from DMA to DMA. This clearly indicates that the characteristics of water-demand time series differ from DMA to DMA; thus which of these strategies achieves lower performance indices can vary from DMA to DMA. Figure 2 shows the actual water-demand curves and forecasted water-demand curves by the two direct strategies; for DMAs A, E, H, and J, the direct strategy using the single model is applied, and for the remaining DMAs, the direct strategy using the multiple models is applied.

Table 1. Performance indices of three forecasting strategies for 7 days immediately before W1.

DMA ID	Iterated Strategy				Direct Strategy with Single Model				Direct Strategy with 7 Models			
	PI_1	PI_2	PI_3	PI	PI_1	PI_2	PI_3	PI	PI_1	PI_2	PI_3	PI
A	1.58	8.06	0.85	10.48	1.34	6.34	0.78	8.46	1.45	8.06	0.82	10.33
B	0.87	4.12	0.89	5.87	0.95	4.57	0.81	6.33	0.59	2.64	0.69	3.92
C	0.71	2.66	0.57	3.94	0.80	2.75	0.62	4.17	0.64	2.68	0.54	3.86
D	2.48	8.01	1.99	12.48	2.80	9.12	2.07	14.00	2.49	8.01	2.00	12.51
E	2.53	10.94	1.87	15.34	2.19	6.12	1.93	10.24	2.57	10.45	2.05	15.07
F	0.71	2.26	1.01	3.99	0.65	1.76	0.99	3.40	0.68	1.61	0.99	3.28

Table 1. *Cont.*

DMA ID	Iterated Strategy				Direct Strategy with Single Model				Direct Strategy with 7 Models			
	PI_1	PI_2	PI_3	PI	PI_1	PI_2	PI_3	PI	PI_1	PI_2	PI_3	PI
G	1.27	4.11	1.55	6.93	1.56	4.85	1.87	8.28	1.25	4.02	1.58	6.84
H	0.94	3.96	1.02	5.93	0.82	1.84	1.08	3.73	0.89	2.18	1.01	4.08
I	1.08	3.06	1.11	5.25	1.24	3.87	0.97	6.08	1.18	3.18	1.03	5.39
J	1.47	4.19	1.17	6.84	1.48	3.27	1.10	5.86	1.81	4.73	1.10	7.64
Average	1.37	5.14	1.20	7.71	1.38	4.45	1.22	7.05	1.35	4.76	1.18	7.29

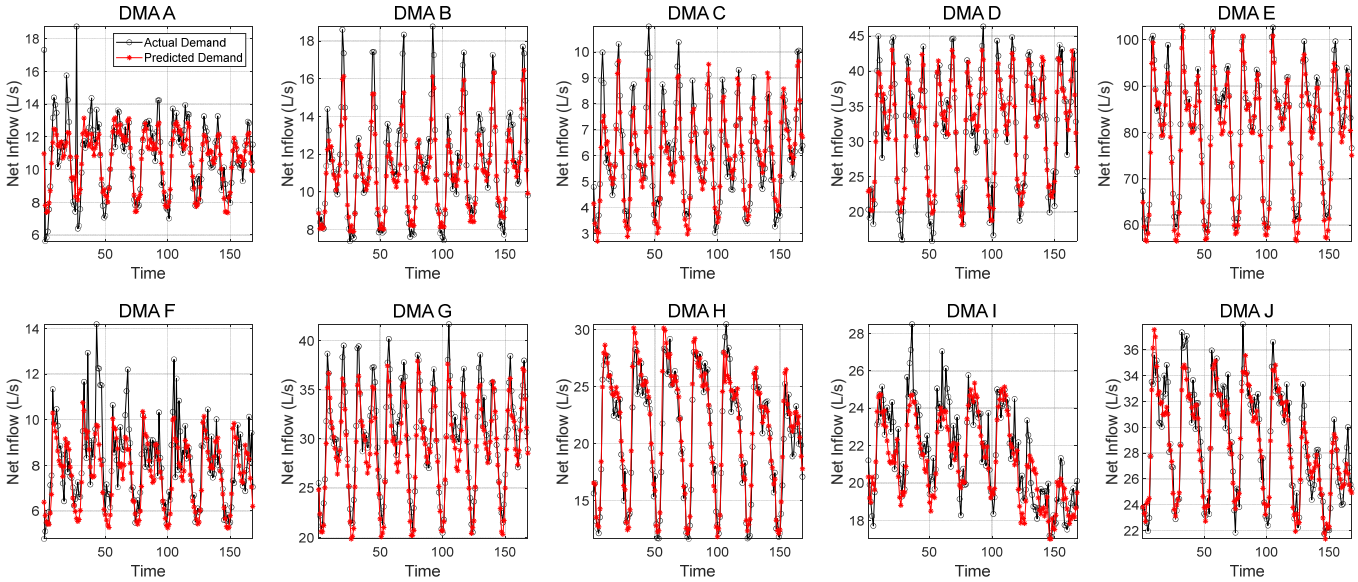

Figure 2. Actual and predicted water-demand curves for 7 days immediately before W1.

4. Conclusions

This paper proposed a weekly water-demand forecasting method combining various statistical techniques. In the proposed method, firstly, training datasets are prepared through EDA, several data preprocessing steps and input selection step, and then forecasting models are constructed by SVR. After this, weekly demand forecasts are calculated using three multi-step ahead forecasting strategies. To verify the performance, the proposed method was applied to urban hourly water-demand datasets provided by the BWDF in the 3rd International WDSA-CCWI Joint Conference in Ferrara, Italy. The experimental results showed that direct strategies can provide more accurate weekly forecasts than the iterated strategy. The results also showed that which of these strategies achieves lower performance indices can vary from DMA to DMA.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y.; software, J.Y.; validation, J.Y., H.B. and M.-S.K.; formal analysis, J.Y., H.B. and M.-S.K.; investigation, J.Y., H.B. and M.-S.K.; resources, K.-J.K. and I.-S.J.; data curation, J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y., H.B. and M.-S.K.; visualization, J.Y.; supervision, I.-S.J.; project administration, K.-J.K.; funding acquisition, K.-J.K. and I.-S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the DGIST R&D Program of the Ministry of Science and ICT (2024010409) and by ETRI grant funded by the Korean government (24ZD1120, Regional Industry IT Convergence Technology Development and Support Project).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in this study were openly available at https://wdsa-ccwi2024.it/battle-of-water-networks/ (accessed on 15 December 2023).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Rousseeuw, P.J.; Driessen, K.V. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **1999**, *41*, 212–223. [CrossRef]
2. Schölkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT press: Cambridge, MA, USA, 2002. [CrossRef]
3. Taieb, S.B.; Sorjamaa, A.; Bontempi, G. Multiple-output modeling for multi-step-ahead time series forecasting. *Neurocomputing* **2010**, *73*, 1950–1957. [CrossRef]
4. Website of the 3rd International Joint Conference on WDSA/CCWI 2024. Available online: https://wdsa-ccwi2024.it/wp-content/uploads/2024/01/BWDF_Instructions_rev4.pdf (accessed on 2 April 2024).