environmental sciences proceedings

MDPI

Proceeding Paper

Evaluation of the Performance Gains in Short-Term Water Consumption Forecasting by Feature Engineering via a Fuzzy Clustering Algorithm in the Context of Data Scarcity †

Georgios Tzanes [1,2,*], Christiana Papapostolou [2], Miltiadis Gymnopoulos [1], John Kaldellis [2] and Anastasios Stamou [1]

[1] Laboratory of Applied Hydraulics, Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechniou 5, 15780 Athens, Greece; mgymnopoulos@mail.ntua.gr (M.G.); stamou@mail.ntua.gr (A.S.)
[2] Soft Energy Applications & Environmental Protection Laboratory, Mechanical Engineering Department, School of Engineering, University of West Attica, 250 Thivon & Petrou Ralli, 12244 Athens, Greece; chrispap@uniwa.gr (C.P.); jkald@uniwa.gr (J.K.)
* Correspondence: g.t.tzanes@uniwa.gr or gtzanes@mail.ntua.gr
† Presented at the 16th International Conference on Meteorology, Climatology and Atmospheric Physics—COMECAP 2023, Athens, Greece, 25–29 September 2023.

Abstract: Accurate short-term water consumption forecasting is a crucial function of modern water supply systems, which, in turn, play a crucial role in the sustainable management of water resources, particularly in regions with limited access to water supplies. This study presents an evaluation of the performance gains in short-term water consumption forecasting by the exploitation of a fuzzy clustering algorithm to engineer new features corresponding to water consumption clusters. The evaluation takes place under data scarcity, meaning both a small dataset and only in situ water consumption measurements. To evaluate the gains, data registered to consumers on the remote island of Tilos are processed to produce two datasets which differ in terms of the addition of clusters. The datasets are consumed by deep neural networks to produce hour-ahead predictions. The inclusion of the clusters in the dataset results in a decreased mean absolute error and root-mean-square error, reduced by 29% and 17% on average, respectively.

Keywords: deep neural networks; short-term forecasting; water consumption

Citation: Tzanes, G.; Papapostolou, C.; Gymnopoulos, M.; Kaldellis, J.; Stamou, A. Evaluation of the Performance Gains in Short-Term Water Consumption Forecasting by Feature Engineering via a Fuzzy Clustering Algorithm in the Context of Data Scarcity. *Environ. Sci. Proc.* **2023**, *26*, 105. https://doi.org/10.3390/environsciproc2023026105

Academic Editors: Konstantinos Moustris and Panagiotis Nastos

Published: 28 August 2023

1. Introduction

Accurate short-term water consumption forecasting is a crucial function of modern water supply systems, which, in turn, play a crucial role in the sustainable management of water resources. The forecasts are exploited for a series of operational decisions. Among other things, forecasts are required for the day-ahead scheduling of water pumping [1], which, in turn, affects electricity consumption, or for delivering short-term schedules [2] (e.g., hour-ahead schedules, minute-head schedules, etc.), which are often employed for scheduling the operation of pressure-reducing valves, hence, reducing water losses in distribution systems.

This paper presents an evaluation of the forecasting performance gains of deep neural networks (DNNs) by the addition of new features. The developed DNNs predict the hour-ahead aggregated water consumption (AWG) of 48 consumers. The additional features are engineered via a fuzzy clustering algorithm using only water consumption as input data and refer to water consumption clusters (WCC).

The DNNs are built and tested in a data-scarcity context, given that only in situ water consumption measurements are available and that the test dataset is limited, extending to 205 h. Using a fuzzy instead of a conventional clustering approach becomes significant

in such a context since it allows for the generation of an arbitrary number of additional features, each corresponding to the membership in an arbitrary number of clusters, enabling more accurate predictions. At the same time, to enhance result robustness and to mitigate effects that might arise due to the small size of the dataset, the performance gains are evaluated twice using two pairs of different DNNs for the same technology.

2. Materials and Methods

2.1. Data Resources

The data are collected from consumers located in Livadia, a community that lies on the east coast of Tilos, a remote island in the South Aegean Sea. Currently, 100 smart water meters (SWMs) are installed, capable of transmitting data every minute via the wireless M-Bus protocol into a local server in the town hall. At the time of writing, the installations are ongoing; hence, the data availability is limited. More precisely, the available data are registered to 76 SWMs between 1 August 2022 00:00 and 11 November 2022 14:00 and refer to measurements of the consumed cumulative water volume (m^3) with a 10 min frequency. Following the data collection, the data are validated using the following steps:

1. Data filtration. The data during the testing period of the installations are discarded. These include data from between 1 August 2022 00:00 and 15 September 2022 00:00 and the readings of 18 SWMs. In addition, the data registered between 11 November 2022 00:10 and 11 November 2022 14:00 are discarded to maintain an exact number of days in the dataset. As an example, the collected measurements of an SWM with id = 24895800 are shown in Figure 1a;
2. Time series resample. The 76 time series, one for each SWM, are resampled with a 10 min frequency. The process does not change the frequency of the time series but adds the missing timestamps. The missing values are filled with NaNs;
3. Outlier replacement. For each SWM, the outliers are detected and replaced with NaNs using the Tukey method [3]. The result is shown in Figure 1b;
4. Data imputation or removal. For each SWM, the ratio of NaNs to the sample size is computed. If the ratio is larger than 6%, then the SWM is discarded from the dataset. Otherwise, the data gaps are filled using linear interpolation, with the result for SWM 24895800 shown in Figure 1c. This step leads to the omission of 10 SWMs;
5. Discharge computation. Each time series left is resampled with an hourly frequency, assigning the max. of every six measurements of 10 min duration to the start of every hourly period. Subsequently, the dataset is differentiated by one timestep. By doing so, the water consumption in terms of water discharge (m^3/h) is obtained for each SWM;
6. Aggregated water consumption computation. The AWC is computed as the sum of the remaining SWMs. The dataset contains 49 columns, 48 for each of the SWMs and one for the AWC. The row count is 1369 (timesteps). The dataset has a time index ranging between 15 September 2022 00:00 and 10 November 2022 23:00.

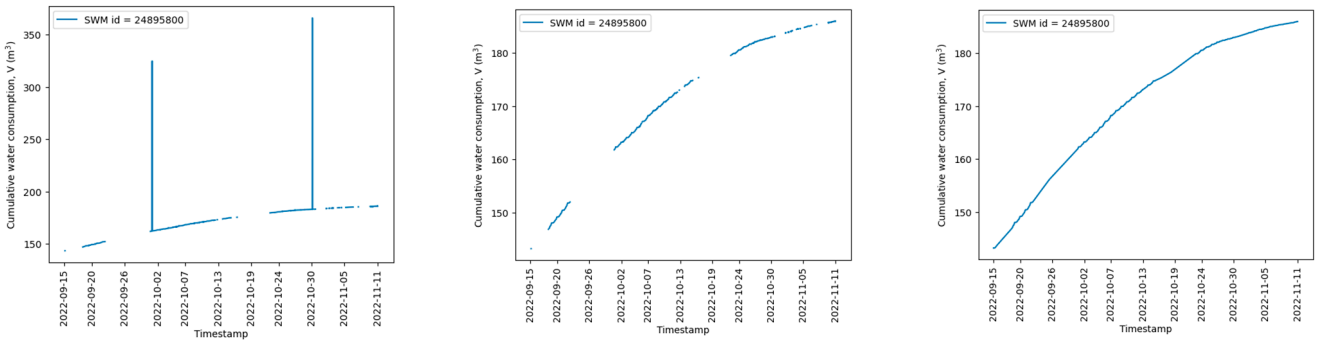

Figure 1. (**a**) The cumulative water consumption measurements of the 24895800 SWM; (**b**) the time series after removing outliers; (**c**) the time series after filling the gaps using linear interpolation.

2.2. Machine Learning Pipeline

The dataset is further processed in order for it to become consumable for a machine learning (ML) pipeline which (a) prepares the dataset, (b) trains an ML model and (c) tests its performance. The exact steps are as follows:

1. The dataset is split into the train, validation and test subsets;
2. The subsets are transformed to achieve stationarity using the train subset to compute the necessary transformation characteristics, thus avoiding data leakage;
3. New features are engineered, aiming to reduce prediction errors;
4. Data are reshaped in order to become consumable for the model;
5. The model is built and trained using the train and validation subsets;
6. The performance of the model is evaluated by comparing the ground truth (measurements) with the respective predictions that are produced using the test subset.

Since the best hyperparameters of an ML model are subject to experimental analysis, the pipeline is repeated 256 times using the ASHA scheduler [4]. The outcome is 256 models, which are tested against their prediction error using the popular mean absolute error (MAE) and the root-mean-square error (RMSE) indices. The ML pipeline is also repeated for two datasets (with and without the WCC), thus, capping at 512 times.

2.2.1. Data Split

Datasets are split into the train, validation and test subsets. The train subset includes ~70% of the total rows, equivalent to 958 rows or hours. The validation subset includes the next 15% of the total rows, equivalent to 205 rows or hours. The validation subset includes the last 15% of the total rows, equivalent to 205 rows or hours.

2.2.2. Data Transformation

The datasets are transformed to make the time series stationary, facilitating the training algorithm convergence. To do so, the following steps are taken:

1. The datasets are log transformed. The transformation reduces the distribution skewness while stabilizing variance over time;
2. The log-transformed datasets are subsequently detrended using linear regression. The linear model is fitted (a = -0.000411, b = -0.444404) using the train subset;
3. The datasets are standardized, meaning the subtraction of the average and the subsequent division by the standard deviation. The standard deviation and the mean are computed using the corresponding train subset.

2.2.3. Feature Engineering

Both datasets are enriched with features as follows:

1. The SWMs are aggregated into groups using bins of 5 m^3/h. The resulting groups have a larger correlation coefficient with the AWC;
2. Seven new features are engineered using the statistical properties of the AWC. These include a comparison of the AWC with the previous daily mean, sum, max, 3rd and 1st quartile, as well as with the ramp during the last hour and the sum of the last 3 h.

One of the two datasets is enriched with the WCC, which are built using a fuzzy clustering algorithm, namely the skfuzzy [5]. The approach is based on unsupervised learning, meaning that the algorithm discovers associations or groups in a dataset (patterns) by itself without knowing whether they exist, how many there are and which ones they are. The problem the algorithm must solve, also known as the learning task, concerns the clustering of the AWC values as high, medium or low. The algorithmic steps are as follows:

1. Initialization of the fuzzy partition, i.e., a matrix U(0), which contains the degree of membership μ in a predetermined number of clusters;
2. Calculation of the corresponding vectors at the center of the clusters;
3. Calculation of the Euclidean distance d between the data points (water consumption values) and the centers of the clusters;

4. Calculation of the new membership degree μ and fuzzy partition matrix U(1);
5. Convergence check, comparing pairwise each μ between U(0) and U(1). If the maximum absolute difference is greater than the predefined threshold ε = 0.001, steps (2) to (5) are repeated with new cluster centers; otherwise, the algorithm stops.

Upon completion of these steps, an unsupervised learning model has been built where a water consumption value is entered, and it outputs the degree of membership in each of the predefined clusters. The predetermined number of clusters is an input parameter of the model and can be selected arbitrarily or using the fuzzy partition coefficient [6].

2.2.4. Deep Neural Networks Architecture

The DNNs are built based on the Seq2Seq model architecture [7], featuring an attention mechanism. Additionally, a final gated recurrent unit's layer, with one unit and a linear activation function, is stacked. Seq2Seq models comprise two neural networks in an encoder–decoder configuration. In this paper, these neural networks are long short-term memory (LSTM) networks. The networks are arranged and operate as follows:

1. A three-dimensional matrix is inserted into the encoder network. The results of the calculations of this network are: (a) the final memory content for the last timestep, which is called the context vector (CV), and (b) the outputs of the output layer, which are rejected. The CV is the input to the decoding network;
2. In the decoding network, the CV from step (1) is entered for the first timestep and then iteratively continues the calculations using the network's units.

The attention mechanism is applied to Seq2Seq models. The attention mechanism was proposed in 2015, and several variants of it have already been developed [8]. With the addition of the attention mechanism, the following differences arise:

1. A CV is calculated for each h_t of the encoding network. Without the mechanism, only the last timestep t to generate a single CV is considered;
2. An alignment score is generated. The score is calculated based on the h_t of the decoder (Luong attention mechanism) by a stacked neural network. This score may be interpreted as weights that give "attention" to the most "important" input data.

2.2.5. Tuning of the Hyperparameters

The hyperparameters to be tuned and the search field of their optimal values are shown in Table 1. For each combination of these hyperparameters, a new DNN is built, trained and tested. The chosen training algorithm is the Amsgrad [9]. The training process aims to minimize the RMSE (cost function).

Table 1. The hyperparameters of the Seq2Seq-Attention model to be tuned and their search space.

Nr.	Hyperparameter	Search Space
1	Units of the encoder and the decoder	Integer in [8, 768]
2	Encoder activation function	Choice from relu, sigmoid, softplus, softsign, tanh, selu, elu, exp and LeakyReLU
3	Decoder activation function	
4	Learning rate of the training algorithm	Choice from [0.0008, 0.01]

3. Results

3.1. Water Consumption Clustering

The selected number of clusters to proceed with is three. The fuzzy partitioning coefficient is high (~88%), meaning there is a crisp partitioning of the data. The transformed water consumption values lower than −1 are labeled as low (Cluster 1). The transformed water consumption values ranging between −0.926 and 0.616 are labeled as medium (Cluster 2), and those exceeding 0.619 are labeled as high (Cluster 3). In Figure 2, the transformed water consumption values are shown on the secondary *y*-axis over the degree of membership in each of the three clusters (primary *y*-axis) for a total of 1 week. As an

example, during the 108th timestep, the consumption peaks, and the membership is 100% allocated in Cluster 3 (brown color) and is thus labeled by the model as high.

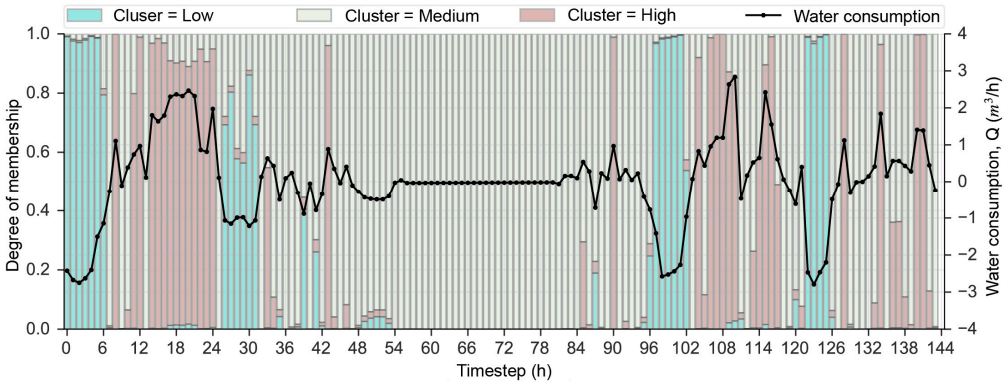

Figure 2. The transformed water consumption for a week against their membership degree in each of the three clusters.

3.2. DNNs Performance

In the present paper, two datasets are examined with the inclusion of three more features, i.e., the WCC, as the sole difference. For each dataset, an individual hyperparameter tuning process is conducted, examining 256 DNN configurations. The configurations with the lowest MAE and RMSE, respectively, are shown in Table 2. The DNN with the lowest MAE achieves a reduction of 31.4% and 18.9% in MAE and RMSE, respectively. The DNN with the lowest RMSE achieves a reduction of 26.9% in MAE and 14.2% in RMSE.

Table 2. The forecasting error metrics of the selected DNNs with and without the addition of WCC.

Nr.	Dataset	Model	MAE	RMSE
1	Without WCC	Seq2Seq-Attention-MAE	0.118	0.164
2		Seq2Seq-Attention-RMSE	0.134	0.169
3	With WCC	Seq2Seq-Attention-MAE	0.081	0.133
4		Seq2Seq-Attention-RMSE	0.098	0.145

The predictions of the AWC for one hour ahead are shown in Figure 3 against the ground truth (AWC), as measured by the SWMs. The predictions are registered to the DNN that has the minimum MAE among the 254.

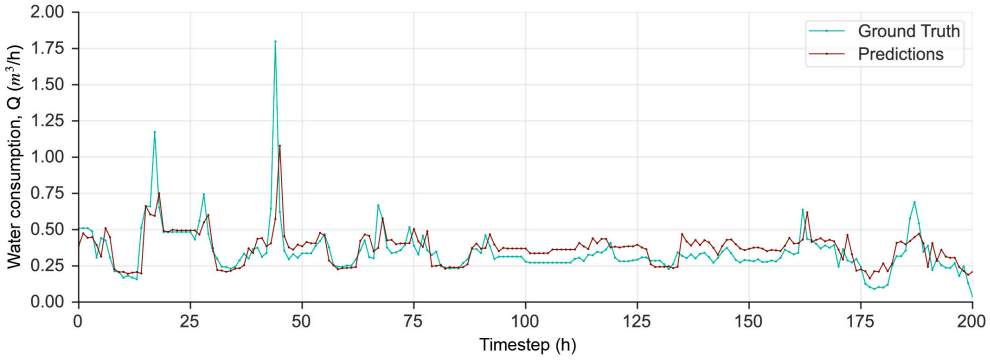

Figure 3. The hourly predictions of the selected model against the ground truth for the test subset.

4. Discussion

The reduction in the error metrics is large, proving the significance of the additional features, i.e., the three WCC. The test subset, based on which the error metrics are computed, is small, containing 205 values, corresponding to hours. To increase the results' robustness,

the experiments are duplicated, assessing the performance of two DNNs, which are selected among the trained models based on their MAE and RMSE. As shown in the results, in both cases, the addition of the WCC achieves a reduction in MAE and RMSE, i.e., a reduction of 29% and 17% on average, respectively.

Given that the MAE is larger than the RMSE and that the RMSE penalizes large errors more than MAE, the additional features do not favor the models' performance near consumption peaks, but more over valleys and time periods where consumption changes smoothly.

Using the fuzzy clustering algorithm to engineer additional features provides flexibility given that the number of clusters is arbitrary, and each cluster is mapped to a new feature. That is also the main difference of more conventional approaches where one feature is built for characterizing the water consumption levels.

5. Conclusions

In the present paper, a fuzzy clustering algorithm was applied in order to engineer new features which correspond to three water clusters of low, medium and high water consumption. Using this approach, two datasets were built, with these additional features as the sole difference. The datasets were exploited to train deep neural networks to produce hour-ahead predictions of the aggregated water consumption of 48 consumers. The addition of the clusters resulted in reduced MAE (29%, on average) and RMSE (17%, on average), with the forecast errors being reduced the most during off-peak periods of water consumption.

Author Contributions: Conceptualization, G.T.; methodology, G.T.; software, G.T.; validation, C.P. and M.G.; data curation, G.T. and J.K.; writing—original draft preparation, G.T.; writing—review and editing, C.P., M.G., J.K., A.S.; visualization, G.T.; supervision, J.K. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship, and Innovation under the call RESEARCH—CREATE—INNOVATE (project code: T2EDK-01578).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sarmas, E.; Spiliotis, E.; Marinakis, V.; Tzanes, G.; Kaldellis, J.K.; Doukas, H. ML-based energy management of water pumping systems for the application of peak shaving in small-scale islands. *Sustain. Cities Soc.* **2022**, *82*, 103873. [CrossRef]
2. Kavya, M.; Mathew, A.; Shekar, P.R.; Sarwesh, P. Short term water demand forecast modelling using artificial intelligence for smart water management. *Sustain. Cities Soc.* **2023**, *95*, 104610. [CrossRef]
3. Tukey, J.W. Exploratory Data Analysis. In *The Concise Encyclopedia of Statistics*; Springer: New York, NY, USA, 2008; pp. 192–194. [CrossRef]
4. Li, L.; Jamieson, K.; Rostamizadeh, A.; Gonina, E.; Hardt, M.; Recht, B.; Talwalkar, A. A System for Massively Parallel Hyperparameter Tuning. *arXiv* **2018**, arXiv:1810.05934.
5. JDWarner/scikit-fuzzy: Scikit-Fuzzy Version 0.4.2. Available online: https://zenodo.org/record/3541386 (accessed on 1 June 2023).
6. Tzanes, G.; Zafirakis, D.; Makropoulos, C.; Kaldellis, J.K.; Stamou, A.I. Energy vulnerability and the exercise of a data-driven analysis protocol: A comparative assessment on power generation aspects for the non-interconnected islands of Greece. *Energy Policy* **2023**, *177*, 113515. [CrossRef]

7. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. *arXiv* **2014**, arXiv:1409.3215.
8. Luong, M.-T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
9. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of ADAM and Beyond. In Proceedings of the 6th International Conference on Learning Representations, Vancouver Convention Center, Vancouver, BC, Canada, 30 April 2018–3 May 2018. [CrossRef]