engineering proceedings

MDPI

Proceeding Paper

Battle of Water Demand Forecasting: Integrating Machine Learning with a Heuristic Post-Process for Short-Term Prediction of Urban Water Demand †

Alexander Sinske *, Altus de Klerk and Adrian van Heerden on behalf of the GLS A3 Team

GLS Consulting, Stellenbosch 7600, South Africa; altus@gls.co.za (A.d.K.); adrian@gls.co.za (A.v.H.)

* Correspondence: alex@gls.co.za

† 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: The challenge in water demand forecasting within a Northeast Italy water distribution network (WDN) involves predicting demands across ten distinct District Metered Areas (DMAs) with varying characteristics and demand profiles. This is critical for optimizing system operation in the near future. The available data begins in January 2021, with unknown impacts of post-COVID socio-economic changes, like work-from-home policies. To address this, the team integrates heuristic and Machine Learning (ML) techniques to predict short-term demands and fill data gaps. A heuristic post-processing step, using weighted sums and historical trends, refines these predictions. This approach combines ML with traditional methods with a view to servicing developing nations.

Keywords: forecasting; statistical; water demand; weighted factors; SARIMAX

check for updates

Citation: Sinske, A.; de Klerk, A.; van Heerden, A., on behalf of the GLS A3 Team. Battle of Water Demand Forecasting: Integrating Machine Learning with a Heuristic Post-Process for Short-Term Prediction of Urban Water Demand. *Eng. Proc.* **2024**, *69*, 203. https://doi.org/10.3390/engproc2024069203

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

Published: 22 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Water scarcity is a global challenge exacerbated by climate change, population growth, and environmental degradation. Accurate forecasting of water demand is crucial for designing and maintaining water distribution networks. Short-term predictions with high granularity are useful for Water Distribution Network (WDN) operation and management to ensure efficient supply of water in a cost-effective manner.

The Battle for Water Demand Forecasting (BWDF) required forecasting for ten District Metered Areas (DMAs) of a case study WDN location in the North-East of Italy, supplying a number of areas that vary considerably in size, land use and water demand characteristics.

The available data, starting from 1 January 2021, are from the post-COVID period and, although the characteristics and population size of each DMA are not deemed to change in the battle period, the implications of socio-economic changes, such as work-from-home policies and possible reluctance to travel, cannot be reliably captured in the period of available data.

In the context of developing nations, advanced tools for demand predictions are not always available, or the knowledge to operate such tools is lacking. To this end, it was decided to attempt predictions as outlined by the battle without the use of advanced software tools that may not be easily accessible to developing nations.

Readily accessible functions in Python and Excel were used, while other more advanced statistical methods and Machine Learning (ML) were avoided, even though more detailed statistical methods, such as SARIMAX [1], could have been used.

It was also determined that that only a few points were available for each unique day, week, or month of the year from which to make a forecast.

2. Methodology

The method employed by the GLS A3 team explores the integration of techniques to perform an initial prediction for the short term as well as infilling of gaps in the demand data. A heuristic post-process, based on engineering judgement, is applied that considers selected representative dates and utilises a weighted sum function to further transform the prediction results.

The method used in this paper is simplistic by nature and the approach can be broken down into a few steps.

2.1. Preparing Data

The data preparation step is required to account for the daylight savings time switchovers that occur twice per year between normal and summertime periods to ensure homogenous timeseries data.

The month, day and hour data were converted to representative integer values and were the only exogenous variables used to help with gap filling. Preliminary analysis indicated that, whenever additional exogenous variables were added, such as temperature, the resulting predictions tended towards the average.

2.2. Removing Data

The second step is to remove the final 168 hours' worth of data from the supplied dataset. The remaining data will be used as training data to determine weights that provide the best estimated values to reduce the error scores compared to the actual values that were removed. These weights will then be used to predict the unknown week's values using the full dataset.

2.3. Imputed Data

The third step, using the training data, is to fill all the missing values in the data set, as well as predicting estimated values for the 168 hours' data that were removed from the given data.

Random Forest Regression [2] was used to fill any missing data gaps and produce the initial demand data for the forecast week.

Figure 1 shows the difference between the actual demand and the imputed demand for the final week of data. It can be seen that the imputed data fail to predict peaks and troughs well, but it does follow the general trend of demand.

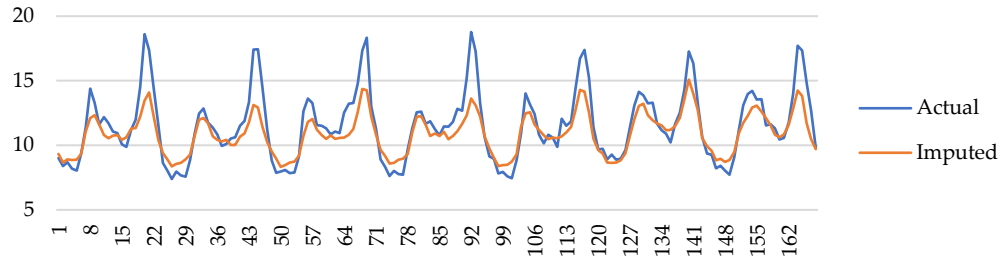

Figure 1. Difference between actual readings and initial imputed forecasts.

It is anticipated that the next step, where weights are calculated, will hopefully stretch the peaks and troughs to more closely align with the actual demand.

2.4. Creating Weights

An initial weight of 100% was given to the imputed values created and a score calculated using the formulas provided for the battle.

Weather and water use in the recent past is thought to provide a better indication of actual water usage compared to older water use data and therefore the temporal proximity of historic data were used to assign weights in the dataset used to calculate demand.

Assigning weights to the water demand for the previous week, and the week prior to that, provided better results, as these align with the weekly water usage patterns.

A Monte-Carlo type analysis was then performed by running through all the possible scenarios of weight combinations for the demands. The weight combination that provided the lowest error score for each DMA was determined for use in the next phase to calculate the estimated demand for the unknown week.

Table 1 shows the calculated weights as percentages.

Table 1. Weightings per DMA.

Desc	Hour Offset	DMA A	DMA B	DMA C	DMA D	DMA E	DMA F	DMA G	DMA H	DMA I	DMA J
2 Weeks	−336	69	57	100	26	100	70	75	62	31	45
1 Week	−168	1	43	0	15	0	9	25	38	11	1
Imputed	0	30	0	0	59	0	21	0	0	58	54

To determine the Weighted Demand value for the prediction, the following formula (Equation (1)) was used. If more weights are contemplated, the formula could be defined as:

$$Demand = \sum_{i=a}^{n} w_i * d_i \quad (1)$$

where:

w_i is the weight at a given time i as a ratio between 0 and 1.

d_i is the demand at a given time i.

Figure 2 below shows how the weighted prediction better aligns with the actual recorded data, as it is able to better stretch the model to match peaks and troughs. For this DMA (DMA B), the imputed score was 7.21, while the weighted score was 5.68 using the formulas provided. In total, all the imputed DMAs scored 92.32 while the total score for the weighted DMAs was 59.78. Though not perfect, numerous peaks are better predicted.

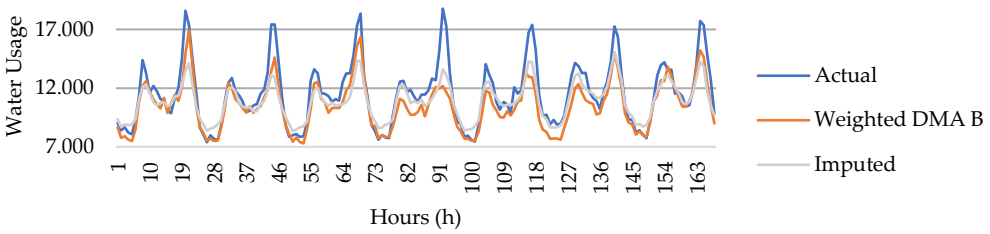

Figure 2. Difference between actual, imputed and weighted forecasts.

A SARIMAX forecast was performed on a subset of the data. Comparing this with the results (Figure 3) of the simplistic weighted solution described in this paper, it can be seen that some days are predicted better, while others are further away.

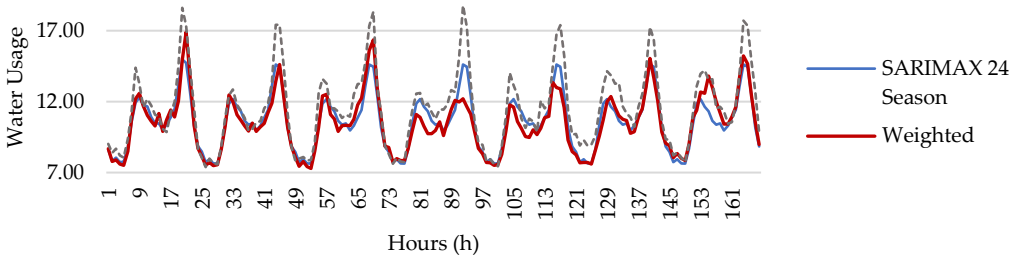

Figure 3. Difference between SARIMAX and the weighted forecasts.

2.5. Forecasting

The 168 hours' worth of data that were removed were placed back into the dataset. The same process was then followed to determine imputed values for the full dataset. The weights determined with the test set, using the training data, were then applied to predict the unknown week's demand using the imputed values of the unknown week, and the demands of the two weeks before that.

3. Conclusions

In conclusion, this study presents a simplified yet seemingly effective methodology for water demand forecasting within a diverse WDN context, specifically focusing on ten distinct DMAs in Northeast Italy. The approach integrates accessible tools, like Python and Excel, avoiding complex statistical or machine learning techniques due to data limitations and resource considerations common in certain settings. By leveraging heuristic post-processing and weighted sum functions based on historical trends and engineering judgment, the method refines initial predictions derived from basic data preparation and imputation steps. While not perfect, this holistic approach demonstrates adaptability in addressing real-world challenges of water demand prediction.

Author Contributions: Conceptualization, A.d.K.; methodology, A.d.K. and A.v.H.; software, A.v.H.; validation, A.d.K. and A.v.H.; formal analysis, A.d.K. and A.v.H.; investigation, A.v.H.; data curation, A.d.K.; writing—original draft preparation, A.v.H.; writing—review and editing, A.S. and A.d.K.; visualization, A.v.H.; supervision, A.S.; project administration, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Please contact any of the authors for access to the data compiled.

Acknowledgments: Nicholas Moult: Review and Proof-reading.

Conflicts of Interest: All authors were employed by the company, GLS Consulting (Pty) Ltd. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Autoregressive Integrated Moving Average. In Wikipedia, The Free Encyclopedia. Available online: https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average (accessed on 12 April 2024).
2. RandomForestRegressor. In Scikit-Learn: Machine Learning in Python. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 12 April 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.