engineering proceedings

Proceeding Paper

Interpretable AI for Short-Term Water Demand Forecasting [†]

Aly-Joy Ulusoy *, Carlos Jara-Arriagada, Yuanyang Liu, Bradley Jenks and Ivan Stoianov

Department of Civil and Environmental Engineering, Imperial College London, London SW7 2BU, UK;
carlos.jara-arriagada18@imperial.ac.uk (C.J.-A.); yuanyang.liu16@imperial.ac.uk (Y.L.);
b.jenks21@imperial.ac.uk (B.J.); ivan.stoianov@imperial.ac.uk (I.S.)

* Correspondence: aly-joy.ulusoy15@imperial.ac.uk

[†] 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: Machine learning models such as artificial neural networks (ANNs) are becoming increasingly popular in short-term water demand forecasting. This is because, despite their lack of interpretability, ANNs are able to capture complex interactions between explanatory variables and water consumption better than a traditional time series analysis or simple linear regression. In this work, we forecast the hourly water demand of ten operational district metered areas using optimal trees, a machine learning model which has been shown to combine the interpretability of regression approaches and the accuracy of ANNs. We show that, compared to existing water demand forecasting models, optimal trees offer valuable insights without sacrificing predictive or computational performance.

Keywords: interpretable machine learning; optimal trees; water demand forecasting

1. Introduction

Accurate short-term water demand forecasts are essential for developing effective water distribution system operation and management strategies. While short-term water demand forecasting has traditionally relied on time series analysis and regression models, the past 15 years have seen a rise in more sophisticated machine learning methods, such as artificial neural networks (ANNs)—see [1] for a recent review. Compared to a traditional time series analysis, machine learning models can better represent complex interactions between explanatory variables and water consumption. They are, however, more difficult to interpret for water network operators. Even the large number of deep decision trees within random forest approaches [2] still limits the ability of humans to use these models to derive insights or a simple prediction logic. In this work, we propose to investigate recent machine learning algorithms which combine the interpretability of regression approaches and the accuracy of ANNs. In particular, we apply optimal regression trees to forecast the hourly water demand of ten operational district metered areas (DMAs) located in the northeast of Italy over a period of one week. We compare the results of the proposed approach with alternative state-of-the-art water demand forecasting models and show that optimal regression trees provide meaningful insights about short-term water demand without sacrificing predictive or computational performance.

2. Materials and Methods

2.1. Optimal Regression Trees

Regression trees are a type of classification method where a data set is recursively partitioned to yield a number of hierarchical, disjoint regions. A regression tree $\mathcal{T}$ is composed of a set of branch and leaf nodes, $\mathcal{T}_B$ and $\mathcal{T}_L$, where a split along a branch node $t \in \mathcal{T}_B$ is governed by parameters $\boldsymbol{a}_t \in \mathbb{R}^n$ and $b_t \in \mathbb{R}$, and each leaf node $l \in \mathcal{T}_L$ is associated with a label c_l. Consider, for instance, a data point $(\boldsymbol{x}, y) \in \mathbb{R}^n \times \mathbb{R}$ with features

check for updates

Citation: Ulusoy, A.-J.; Jara-Arriagada, C.; Liu, Y.; Jenks, B.; Stoianov, I. Interpretable AI for Short-Term Water Demand Forecasting. *Eng. Proc.* **2024**, *69*, 101. https://doi.org/10.3390/engproc2024069101

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

Published: 10 September 2024

$\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}$ and a label $y \in \mathbb{R}$. If $\boldsymbol{x}$ meets all previous conditions leading to a branch node $t \in \mathcal{T}_B$ and ${\boldsymbol{a}_t}^T\boldsymbol{x} < b_t$, the classification will follow the left branch down from node t (otherwise, if ${\boldsymbol{a}_t}^T\boldsymbol{x} \geq b_t$, it takes the right branch) and so on, until it reaches a leaf node $l \in \mathcal{T}_L$. We denote this by $c(\boldsymbol{x}, \mathcal{T}) = c_l$, the final prediction returned by $\mathcal{T}$ for $\boldsymbol{x}$ and the misclassification error for $(\boldsymbol{x}, y)$ is then given by $\left|\left|c(\boldsymbol{x}, \mathcal{T}) - y\right|\right|^2$.

Now consider a data set containing m observations $(\boldsymbol{x_i}, y_i)$, $i \in 1, \ldots, m$, each with n features $\boldsymbol{x_i} \in \mathbb{R}^{\boldsymbol{n}}$ and a label $y_i \in \mathbb{R}$. The regression tree that best represents this data set while maintaining low complexity corresponds to the solution of the problem:

$$\frac{\text{minimize } 1}{m} \sum_{i=1}^{m} \left|\left|c(x_i, \mathcal{T}) - y_i\right|\right|^2 + \alpha \frac{|\mathcal{T}|}{2}, \qquad (1)$$

where $|\mathcal{T}|$ represents the number of nodes in $\mathcal{T}$; ($|\mathcal{T}|/2$ is the number of branch nodes in $\mathcal{T}$); and α is a parameter penalizing tree complexity. Traditional heuristic methods (such as, e.g., CART) solve (1) through a top-down, greedy approach. Instead, we propose the implementation of interpretable AI's optimal regression tree module [3], which relies on a mixed-integer reformulation and global solution of (1) using the off-the-shelf mixed-integer optimization solver GUROBI [4].

2.2. Feature Selection

Our forecasting model incorporates three main types of features—see Table 1. Temporal features aim to capture the inherent diurnal and cyclic patterns of water usage, while weather variables account for the influence of environmental conditions. Our feature selection also includes previous (lagged) demand data, which represents the most important explanatory factor of short-term water consumption according to the literature [2].

Table 1. Feature selection.

Category	Description	Features
Time	Temporal (seasonal, monthly, weekly, and diurnal) characteristics of the forecast period	Quarter, month Day of week, day type (weekend/holiday) Time of day
Weather	Raw data corresponding to the forecast period collected from a local weather station	Air temperature Humidity Wind speed Rainfall depth
Previous waterdemand	Historical water consumption data corresponding to the week preceding the forecast period	1h lagged demand 24h lagged demand 168h lagged demand

2.3. Implementation Details

Our methodology leverages interpretable AI's software modules [3] to (i) impute missing data in the provided time series and (ii) solve Problem (1) with global optimality for each of the ten operational DMAs considered in this study. For data imputation, we implement the *OptImpute* module with the K-nearest neighbors objective function [5]. We then solve Problem (1) using the *OptimalTreeRegressor* module with automatic complexity parameter tuning and a maximum tree depth set to eight [6]. All interpretable AI software is implemented in Julia 1.9.3 [7].

Since past consumption is an important predictor of short-term demand [2], our approach combines three optimal tree models with prediction horizons of 1 h, 24 h, and 168 h using the previous demand features described in Table 1. This approach leverages the most recent inflow data to improve forecasting performance over the first hour and day of the overall (168 h) prediction horizon. We train our forecasting model using the latest inflow data with various window sizes, namely 1-week, 4-week, 26-week, and 52-week windows. The different window sizes aim to mitigate possible trend changes in DMA inflow (e.g. leakage, new development). The models are evaluated on three performance metrics: (i) mean absolute error over a 24 h forecast (MAE-24 h), (ii) maximum absolute error over a 24 h forecast (MaxAE-24 h), and (iii) mean absolute error over a 144 h forecast (MAE-144 h). For each DMA, we select the best model based on performance over a validation testing week and engineering judgement.

3. Results and Discussion

The proposed method is applied to forecast the demand of ten operational DMAs (DMAs A to J) over four validation weeks (W1 to 4) spanning from July 2022 to March 2023. The resulting forecasting models can be found at https://github.com/bradleywjenks/water_demand_forecasting.git (accessed on 9 September 2024). In most cases, we observe that predicted demands align well with the actual demand profile of the DMAs—see, e.g., Figure 1, which represents the forecast obtained for DMA E over W1. Table 2 summarizes the cumulative MAE-24 h, MaxAE-24 h and MAE-144 h performance of the proposed method over all ten DMAs. We observe that, except for W3, the performance of the method is consistent over the different validation weeks. We suspect the poor performance in W3 to be attributed to changes in demand behavior over the week preceding the forecast (winter holidays), which are not explicitly accounted for in the temporal features of W3—this will be the subject of future work. This observation is particularly pronounced for the peak morning demand predictions and reflected by the MaxAE-24 h.

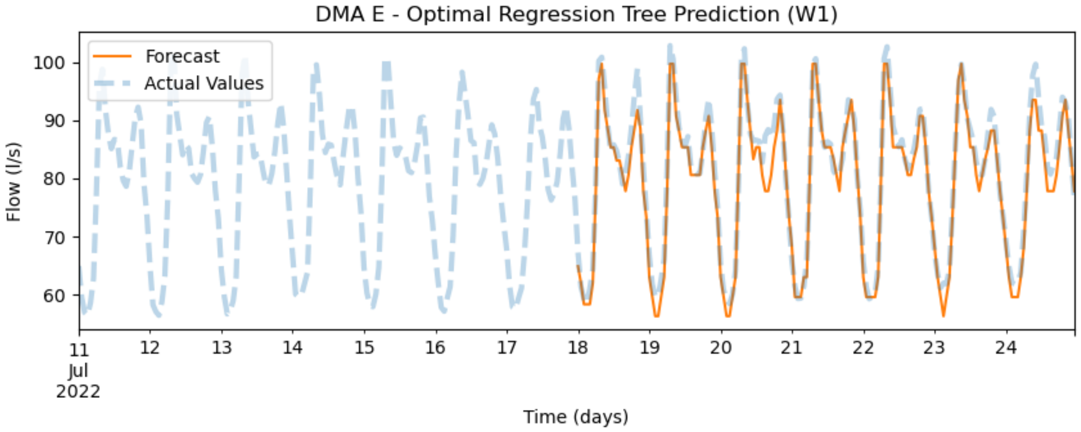

Figure 1. Example of demand prediction forecast for DMA E over validation week W1. The corresponding optimal tree is available at https://github.com/bradleywjenks/water_demand_forecasting/blob/master/results_practice1/plots/dma_e_opt_tree_168h_train_4.svg (accessed on 9 September 2024).

Table 2. Results comparison between optimal regression trees and SARIMAX for the forecast of 10 DMAs *.

Validation Week	Method	MAE-24 h	MaxAE-24 h	MAE-144 h	Combined Score
18 to 24 July 2022 (W1)	SARIMAX	11.52	36.93	11.49	59.94
	Optimal Trees	11.99	36.04	13.79	61.83
24 to 30 October 2022 (W2)	SARIMAX	11.29	30.13	14.49	55.90
	Optimal Trees	10.25	29.26	14.68	54.09
09 to 15 January 2023 (W3)	SARIMAX	10.09	33.11	11.70	54.90
	Optimal Trees	15.34	50.66	16.23	82.23

Table 2. *Cont.*

Validation Week	Method	MAE-24 h	MaxAE-24 h	MAE-144 h	Combined Score
25 February to 04 March 2023 (W4)	SARIMAX	7.99	24.96	8.65	41.60
	Optimal Trees	10.57	30.96	12.01	53.54

* The results presented here are the sum of the performance metrics for each DMA.

Moreover, we evaluate the performance of the proposed model against traditional SARIMAX(4,1,3)(0,1,1,168) models which are trained, for every validation week and DMA, on 26 weeks of historical demand data. (We use the SARIMAX function available in 'statsmodels.tsa.arima.model.ARIMA' from the package statsmodels 0.14.0 [8] in Python 3.9.18. and fit the models with the 'innovations_mle' method.) Table 2 shows that, except for W3, the computed optimal regression trees provide comparable predictive performance to the SARIMAX model. We also note that the optimal regression tree provides evident implementation benefits over the traditional statistical approach. First of all, the training time for the optimal regression tree is substantially shorter (approximately five minutes per DMA, compared to SARIMAX's thirty minutes to one hour training time per DMA). This faster training time makes the optimal regression tree a viable method for near-real-time online demand forecasting for water utilities. In addition to its efficiency, the optimal regression tree provides greater interpretability. Although the trained SARIMAX models return *p*-values and coefficients for the utilized features, these are still difficult to interpret in prediction results for specific times of the day. In contrast, the optimal regression tree presents a clear tree diagram with the selection of the features used and the decision split required for a water demand forecast.

These findings underscore the promising practical advantages of optimal regression trees over traditional statistical approaches, offering both efficiency and interpretability in demand forecasting for water utilities.

Author Contributions: Conceptualization, A.-J.U., C.J.-A., Y.L. and B.J.; methodology, B.J.; software, C.J.-A. and B.J.; validation, C.J.-A. and B.J.; formal analysis, C.J.-A., Y.L. and B.J.; investigation, A.-J.U., C.J.-A., Y.L. and B.J.; writing—original draft preparation, A.-J.U., C.J.-A., Y.L. and B.J.; writing—review and editing, A.-J.U.; visualization, C.J.-A. and B.J.; supervision, I.S.; funding acquisition, I.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2020–72210314, the Natural Sciences and Engineering Research Council of Canada (PGSD-577767-2023), the Royal Academy of Engineering Senior Research Fellow in Dynamically Adaptive Water Supply Networks (RCSRF2324-17-41), Bristol Water, Anglian Water Services, Cla-Val UK, and Analytical Technology.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code associated with this paper is openly available at https://github.com/bradleywjenks/water_demand_forecasting.git (accessed on 9 September 2024).

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

References

1. Niknam, A.; Zare, H.K.; Hosseininasab, H.; Mostafaeipour, A.; Herrera, M. A critical review of short-term water demand forecasting tools—What method should I use? *Sustainability* **2022**, *14*, 5412. [CrossRef]
2. Xenochristou, M.; Hutton, C.; Hofman, J.; Kapelan, Z. Short-term forecasting of household water demand in the UK using an interpretable machine learning approach. *J. Water Resour. Plan. Manag.* **2021**, *147*, 04021004. [CrossRef]
3. Interpretable AI Documentation. Available online: https://docs.interpretable.ai/stable/ (accessed on 12 January 2024).
4. Gurobi Optimizer Reference Manual. Available online: https://www.gurobi.com/documentation/current/refman/index.html (accessed on 26 March 2024).
5. Bertsimas, D.; Pawlowski, C.; Zhuo, Y.D. From predictive methods to missing data imputation: An optimization approach. *J. Mach. Learn. Res.* **2017**, *18*, 7133–7171.

6. Bertsimas, D.; Dunn, J. *Machine Learning under a Modern Optimization Lens*; Dynamic Ideas LLC: Charlestown, MA, USA, 2019.
7. Bezanson, J.; Edelman, A.; Karpinski, S.; Shah, V.B. Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* **2017**, *59*, 65–98. [CrossRef]
8. Seabold, S.; Perktold, J. Statsmodels: Econometric and statistical modeling with python. In Proceedings of the 9th Python in Science Conference, Austin, TX, USA, 28 June–3 July 2010. [CrossRef]

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.