**Comparative Evaluation of Machine Learning Algorithms for Water Consumption Data Analysis and Forecasting: the case of Athens, Greece**

C. Michalopoulos\*, P. Dimas\*^, P. Kossieris\* and C. Makropoulos\*

\* Urban Water Management and Hydroinformatics Group, Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens (NTUA), Iroon Politechniou 5, 157 80 Zografou, Athens, Greece

# Abstract

This study explores the application of various machine learning algorithms for forecasting actual water consumption, a critical task for water utility companies. These algorithms, both stochastic and deterministic, are benchmarked against a baseline method to evaluate their performance. The utility of these methods extends beyond customer billing, particularly for those with inaccurate or untimely usage measurements. They also offer a means to estimate total water losses in the distribution grid by comparing consumed and exported water volumes from treatment plants. The data used for model training and testing were sourced from the Water and Sewage Company of Greece (EYDAP) and include water consumption timeseries for over two million consumers. Extensive preprocessing of raw data and timeseries analysis approaches are presented and discussed. The study underscores the potential of machine learning in enhancing the management of water systems and improving water services.

# Keywords

Demand forecasting, Machine Learning, Long-Short-Term Memory, Water consumption

**Introduction**

Accurate urban water demand forecasting is a cornerstone for operational, tactical, and strategic decisions within drinking water utilities (Gardiner and Herrington, 1986). These predictions enable utilities to effectively meet short-term objectives, such as determining the volume of water to be processed in treatment plants, and long-term goals, like quantifying distribution losses (Donkor et al., 2014) and setting medium- and long-term leakage reduction goals.

The advent of electronic water meters, which monitor consumption at granular time intervals, has spurred the development of numerous forecasting models and methods (Rahim et al., 2020). However, most installed meters remain mechanical (single-jet, multi-jet, volumetric) (American Water Works, 1962). The replacement of all mechanical meters with electronic or smart ones is financially and technically challenging, largely due to infrastructure limitations in many cities, particularly historical ones. The installation of electronic meters requires special probes and a dedicated power supply (Hauber-Davidson and Idris, 2006).

A significant drawback of traditional water meters is the necessity for physical reading, a labor-intensive and time-consuming process (Randall and Koech, 2019). This often results in water companies struggling to maintain a feasible meter reading schedule due to resource constraints.

Recent years have seen the development of numerous water demand forecasting approaches. These methods differ based on the type of available data and the forecast timescale (Kofinas et al., 2014). While most research on urban water demand forecasting focuses on short-term timescales (hourly and daily), our study concentrates on a mid-term timeframe, specifically monthly and quarterly, aligning with typical billing intervals for water companies. To this effect we drew insights from studies in other fields exhibiting similar seasonality patterns in time series, such as residential natural gas and electric power consumption (Liu and Lin, 1991; Tserkezos, 1992).

Water demand forecasting models are categorized into two main types: stochastic and deterministic. Deterministic models consider all factors that influence the outcome and aim to identify patterns among these factors. On the other hand, stochastic models are often developed based on statistical models adapted to previous data in the time series (Box et al., 2015). Common stochastic models include autoregressive (AR), the moving average (MA), the combination of those two with an integration step, the autoregressive integrated moving average (ARIMA), and the seasonal autoregressive integrated moving average (SARIMA) (Box et al., 2015; Hyndman and Athanasopoulos, 2018).  The forecasted values of those models are derived from a linear function of the previous observations. To capture more complex behaviors, long short-term memory neural networks (LSTM) have been proposed, which introduce nonlinear activation functions in the neural network architecture to overcome the limitations of regular stochastic models. Another deterministic method that is utilized is collaborative filtering. Collaborative filtering is primarily used in recommendation systems and matrix completion (Schafer et al., 2007);

In studies with shorter timeframe, it is shown that the precision of models fluctuates for different timesteps. As is described in (Bakker et al., 2013; Guo et al., 2018; Mu et al., 2020) the models achieved better overall performance when dealing with daily (24h) timestep rather than having hourly or 15 minute time frame. For a quarter year timesteps (Kofinas et al., 2014) achieves better performance in contrast to monthly analysis, while (Dikaios Tserkezos, 1992) achieves lower errors for monthly scale timeseries, with this in mind the timescale of the analysis has to be thoroughly be investigated. Because of the simplicity and the accessibility through programing modules, ARIMA model is widely used as a base line model for comparative purposes (Adamowski et al., 2012; Chen and Boccelli, 2018; Liu et al., 2023). Another very common model that has shown great potential in time series forecasting are Neural Networks (Chen and Boccelli, 2018; Herrera et al., 2010; Liu et al., 2023). The majority of them have shown neural network model to have the best forecasting performance for water demand timeseries (Chen and Boccelli, 2018; Kontopoulos et al., 2023; Liu et al., 2023). Another approach suggested in the literature is forecasting with clustered customers where a single model is trained for a group of customers (Huang et al., 2019; Kontopoulos et al., 2023). In this paper, such models were not examined because it has been recently demonstrated that they exhibit overall inferior performance compared to regular models with similar time-series data formats (Kontopoulos et al., 2023). The collaborative filtering method has not been used for straight time series forecasting but there has been some research in completing the missing values from time series (Ma et al., 2019). #TODO

The goal of this paper is to compare several Statistical and Machine Learning models in terms of their accuracy and computational efficiency in predicting water consumption. The predictions can be utilized by the water company to bill customers accurately while reducing the number of visits that are needed by water metering crews. A further complication that is identified and addressed in this work is the fact that in large service areas (like Athens) it is technically infeasible for water companies to conduct regular measurements with a constant time step. This means that available data are not in the appropriate format to be imported into statistical and machine learning models. While its usual practice in research to discard time series with missing values or with unusual timesteps, in our work we develop and test novel techniques to address these issues, ranging from simple interpolation to more complex kernel-based models (Rehfeld et al., 2011). Missing values ate filled in using an interpolation technique that accounts for seasonality, using as a proxy for seasonality the total exported water from the water treatment facilities at each month, as proposed by (Billings and Jones, 2011).

Predictions made in this study are compared against simpler approaches used by water companies to evaluate the added value of the new algorithms.

**Methods**

In this study, Naïve, ARIMA-SARIMA, LSTM, and k-NN time series forecasting methods were selected for further customization and comparison. These methods were chosen based on the belief that they can provide better solutions to the problem, as indicated by the available literature. Each of these models represents a distinct approach to making predictions: Probabilistic, Deterministic and a Clustering approach as well as a base line model for reference.

The models are represented as one benchmark, one stochastic, one deterministic and one that involves groups of customers.

*Seasonal Naïve Approach*

This method, known as seasonal naive forecasting, sets each forecast to be equal to the last observed value from the corresponding season. For example, the prediction for a missing winter quarter-year will be equal to the previous value of last year’s winter quarter-year. The equation describing the naïve approach is as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where m = the seasonal period, and k = , and denotes the integer part of u. (Hyndman and Athanasopoulos, 2018). This approach is the most used by water companies for billing the customers when there are not available measurements, due to its simplicity.

*ARIMA-SARIMA*

ARIMA(p,d,q) model is the combination of Autoregressive (AR), Integrated (I), and Moving Average (MA) models. The Autoregressive component AR(p) represents the regression terms, where p is the number of lagged observations used. The Moving Average component MA(q) represents the moving average terms, where q is the number of lagged error terms included. The Integrated component is denoted by d, which represents the differencing order required to make the data stationary. Firstly, a stable timeseries is developed by differencing by d the original non-stationary historical data. Then the ARMA(p,q) model is fitted to predict the consumption. The ARMA(p,q) model is expressed as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where and are constant, is the white noise sequence. The coefficients and the white noise variance of the ARIMA(p,d,q) model are estimated using the least squares method and moment estimate method. To capture seasonal patterns, the SARIMA model has been developed which takes into consideration the seasonality of the problem. The additional parameters of the SARIMA model are (P,D,Q),m which are the seasonal order terms while m is the seasonality. The model is expressed as SARIMA(p,d,q)(P,D,Q)m. For this research, the seasonal order was m=4 for quarter-year analysis while when performing monthly analysis, m=12 was selected to align with the yearly seasonality. Statistical models contain some uncertainty, this uncertainty can be a deterring factor in the evaluation of the effectiveness of the model. To solve this problem many statistical methods have been address with the most common being the Akaike information criterion (AIC) (Stoica and Selen, 2004). The AIC can be calculated as shown in the Equation 3.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

where, k is the number of model parameters and the maximized value of the likelihood function.

*Long Short-Term Memory (LSTM)*

Long Short-Term Memory Neural Networks (LSTM) are a special kind of Recurrent Neural Networks (RNN) specifically designed to address the challenges faced by standard RNNs (Hochreiter and Schmidhuber, 1997). Each LSTM cell contains three different gates: the input gate, forget gate, and output gate. This architecture enables the handling of longer sequences without encountering issues such as gradient vanishing or exploding. The internal structure of a typical LSTM cell is depicted in Figure 1.

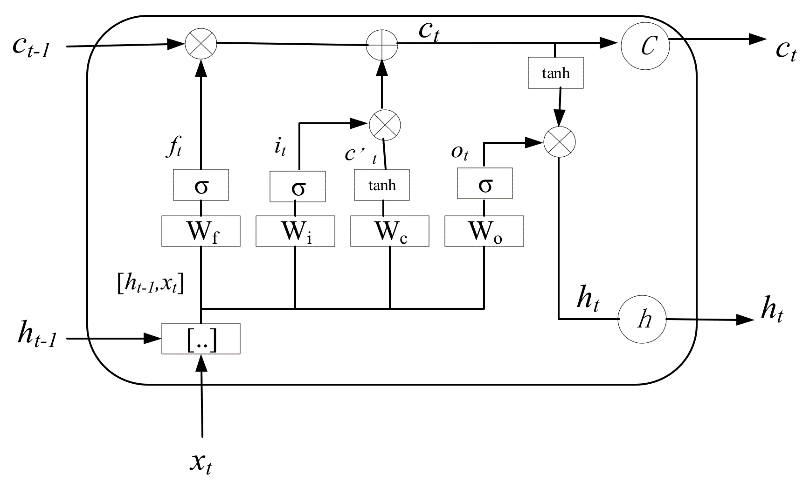


Figure 1: Architecture of a single LSTM block

Tuning LSTM models can be challenging due to the numerous hyperparameters that need to be optimized for optimal performance (Abbasimehr et al., 2020). In this study, the lag (number of previous data points fed to the LSTM) depends on the research timescale. For the monthly scale the input vector is consisted with 12 previous values to drain sufficient information from one year of data, while in quarter year scale the input vector contains 8 values which correspond to the last 2 years of data. For the other parameters there is a lot of available guidance from other researches (Song et al., 2020) suggesting that by increasing the number of stacked layers smaller errors can be achieved with the cost of execution time. Finally, it must be noted that each customers consumption is normalized to the interval [0,1] to avoid instability issues.

*Matrix completion (k-Nearest Neighbors)*

Collaborative filtering methods, including matrix completion algorithms, have been widely adopted by researchers to address missing data in various domains, such as recommendation systems and time series data correction. Because of the nature of the data that is available (measured and unmeasured customers), the unmeasured customers could be forecasted through matrix completion algorithms, as proposed by (Ma et al., 2019). There are a lot of matrix completion algorithms available, but in this research, only the simpler and most common k-Nearest neighbors (k-NN). For this algorithm, the forecasting is calculated as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

where is the value of the i nearest neighbor from the known customers and k is the number of nearest neighbors. The value of k, which is the only hyperparameter in this model, is selected through trial and error.

A flowchart of this process is described in Figure 2.

However, a limitation of this procedure is that it treats every neighbor the same, despite how similar they are. To address this limitation, it is wise to introduce some type of similarity coefficient in the equation. One popular example that is used in recommendation systems is Pearson correlation(Schafer et al., 2007), but for our application the existence of zeros in the timeseries will cause instability issues . For that reason here we adopt the cosine similarity metric (Cui, 2017) which is expressed as:

|  |  |  |
| --- | --- | --- |
|  |  | (5) |
|  |  | (6) |

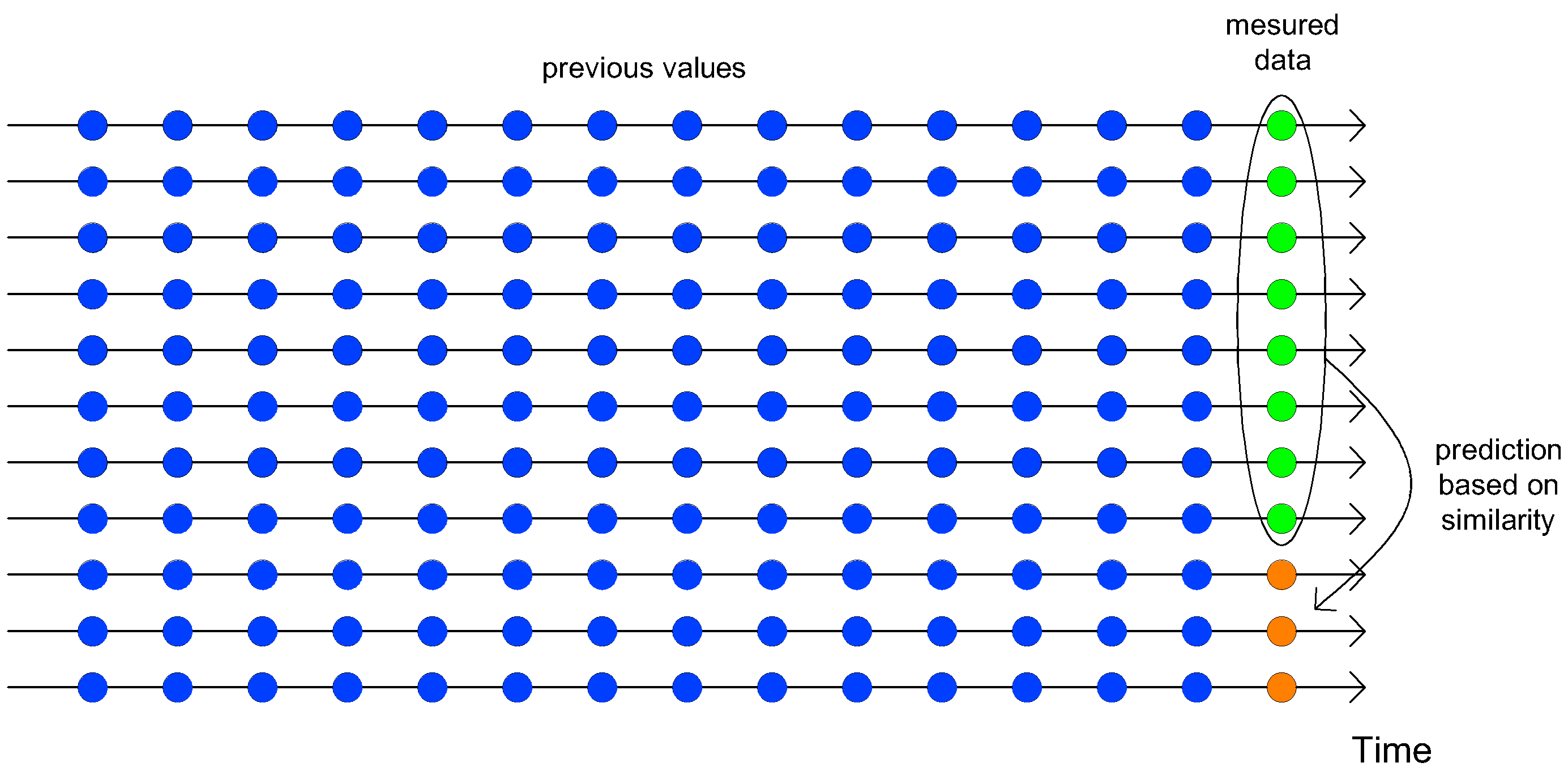


Figure 2: Flow diagram of collaborative filtering approach

Figure 3 illustrates the calculation procedure for the quarter and monthly time scale analysis. Since billing of each customer is done per quarter, the monthly analysis must make three consecutive estimations. As a result, for the monthly timestep, the models will have to take as inputs previous forecasted values, and the errors will be calculated for the sum of those three values. These two approaches are tested to determine if the larger training set can overcome the instability of using previous forecasts as inputs.

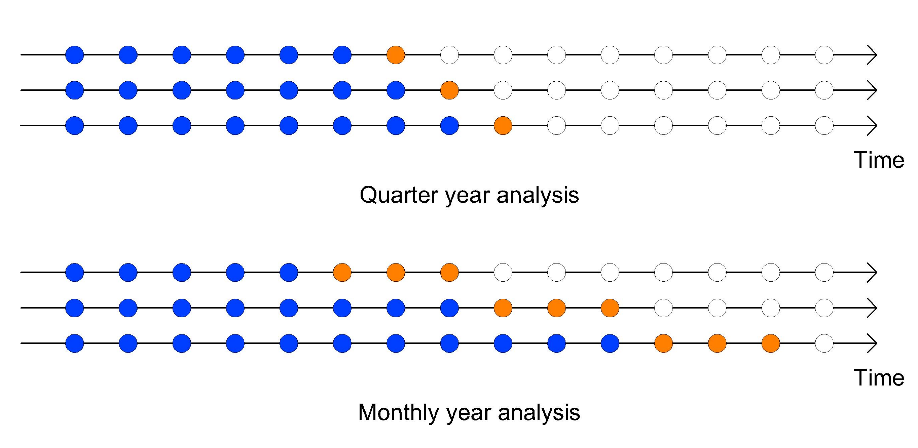


Figure 3:Train and Test data in a timeseries

There are two possible ways that customers are charged. The simplest way is if the company’s crew have managed to record the consumption and the bill promotional to the customer consumption. If the crew have not managed to visit the meter, the company has to bill the customer based on a prediction which in most of cases is equal as what it was the same period in the previous year. To simulate the problem that water companies have to deal, the data is splited into two subsets to subscribe the phenomenon: The measured, where the working crew has managed to record the consumption, and the unmeasured, where the crew did not perform a measurement. For the Water and Sewage Company of Greece (EYDAP) the measured to unmeasured ration is 90:10. In this paper the ratio that it is examined is 80:20 to simulate a more unfavorable situation. The data is spitted randomly to measured and unmeasured sets. The models are trained from the unmeasured subset. Furthermore, the unmeasured timeseries is further divided into a train set and test set for measuring the performance of each model. The metrics that are used to evaluate the performance are:

1. Mean Absolute Percentage Error (MAPE): This metric was chosen because it allows a fair comparison among consumers with relatively high and low consumption due to the property of being scale independent. MAPE is expressed as: , where is the actual value and is the forecasted value.
2. Total Balance: This metric assesses the model's tendency to overestimate or underestimate the forecasts in total. It is expressed as: . The closer this metric is to zero, the more accurate this model is in predicting the total sum for the water utility company.

**Data**

The original format of the consumption data provided by the Water and Sewage Company of Greece (EYDAP) did not serve the application of appropriate methodologies and statistical analysis, as it is characterized by irregular time intervals. For instance, the first measurement of a single timeseries took place in January while the first value of another timeseries was taken in February. For this reason, data preprocessing was performed to extract all the useful data and reform them into monthly timeseries. There are several different approaches when it comes to smoothening irregular timeseries with the most simple being interpolation (Rehfeld et al., 2011). Interpolation was used here for reasons of simplicity, with a simple variation to account for seasonality.

The calculation process is described below:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | (7) | |
|  |  | | (8) |

where and is the interpolation and seasonal coefficient respectively. The is calculated form the ratio of the days in the month to the total days between the measurement. The seasonality coefficient can be drawn from the total trend of the consumers. (Billings and Jones, 2011) has calculated the coefficients based on the water consumption in Alberta in Canada. For each system this can be easily done with the data of the total water exported from the water treatment plants. The average percentage of 10 years of each month for Athens is illustrated in Figure 4.

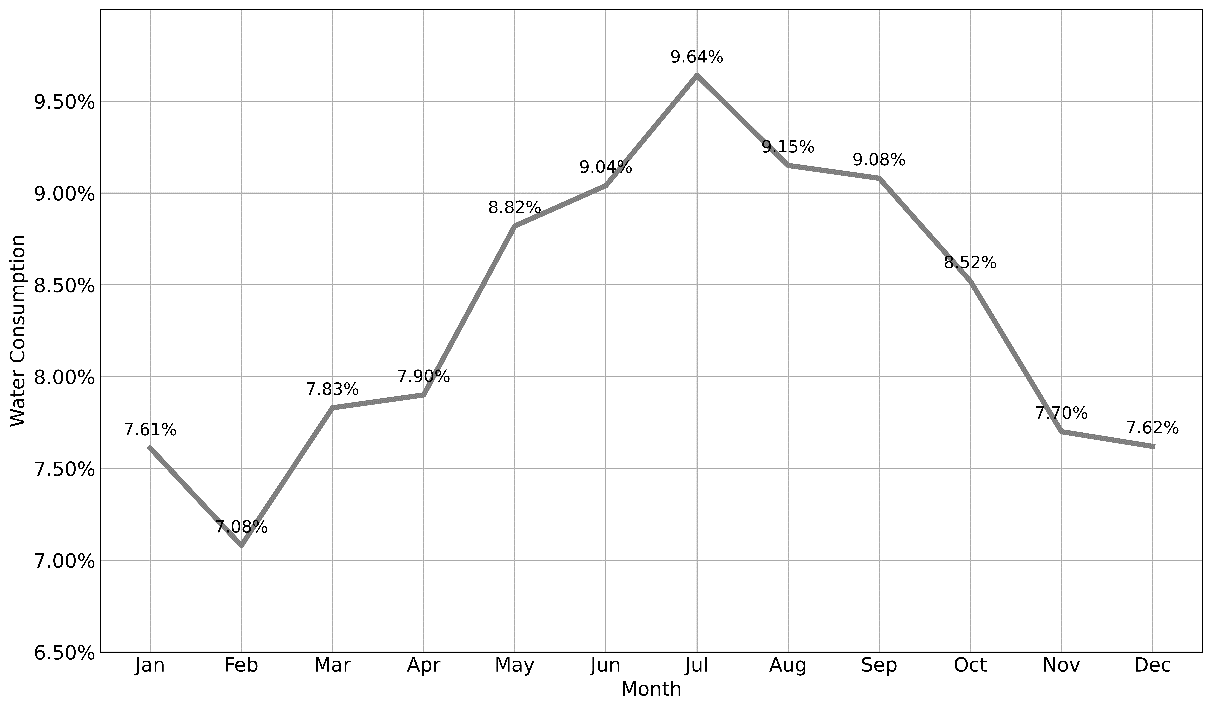


Figure 4: Total water consumption per month.

With this information the coefficient is calculated with the following formula:

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

where is the percentage of water consumption in the j month in respect to the total consumption of the year, while is the total months evolved. The values of for every month can be obtained from the Figure 4. A visual representation of the hole process can be seen in Figure 5. This graphical depiction offers a clear and concise overview of the various stages and components involved in the process.

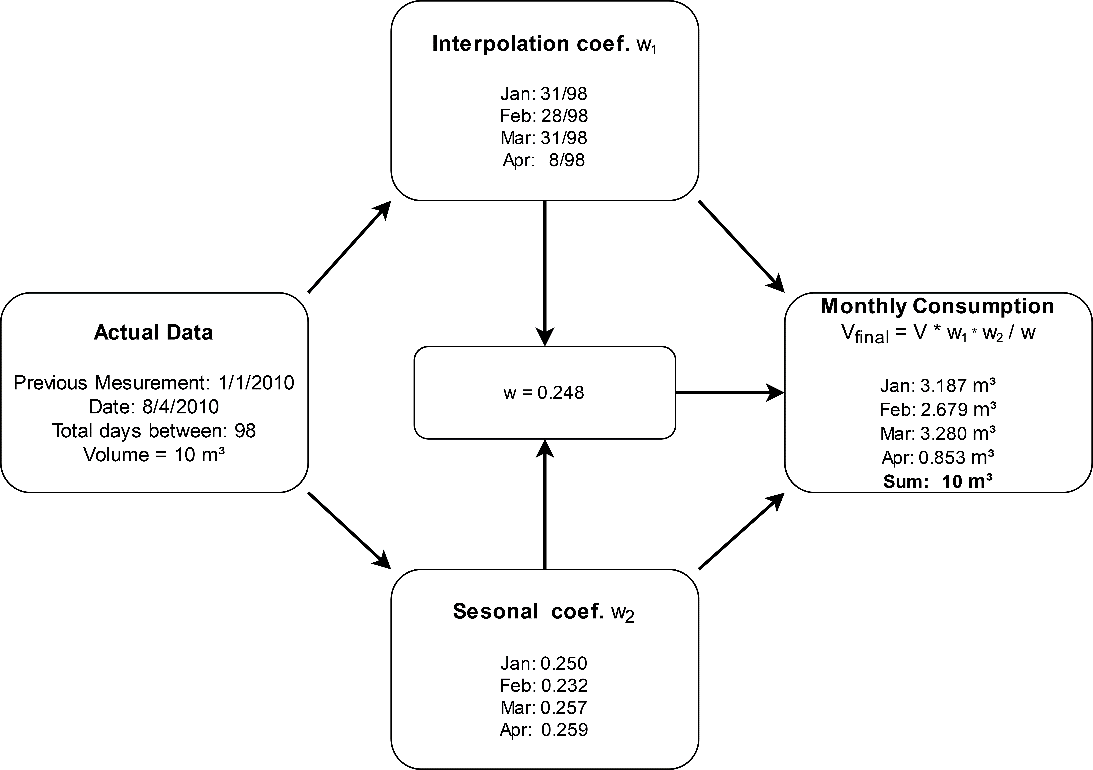


Figure 5: Data preprocessing flowchart

This process is done for every customer (2,107,637 in total) for every measurement that has been recorded for the last 10 years. After this procedure the total dataset is composed with 2,107,637 time series with monthly regular timestep.

**Results and Discussion**

The models that does not need any parameter tuning are the Naïve and the AutoARIMA. The optimal parameters of each model are calculated throu trial and error. For the case of the Arima and the SARIMA models each compination of parameters were tested as it is seen in Table-1. The parameters that were turnned for the LSTM model was the number of epochs as well as the number of neurons in the hidden layer, again the same exhaustive method was used as it seen at Table-2. The only parameter that k-NN method has is the number of nearest neihbors the value of error in respect to the number of neihbors are seen in Figure-6.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | ARIMA | | | | | | q  p | 0 | 1 | 2 | 3 | | 0 | 0.214 | 0.215 | 0.225 | 0.238 | | 1 | 0.214 | 0.219 | 0.230 | 0.238 | | 2 | 0.218 | 0.000 | 0.000 | 0.000 | | 3 | 0.000 | 0.000 | 0.000 | 0.000 | | |  |  |  |  |  | | --- | --- | --- | --- | --- | | SARIMA | | | | | | q  p | 0 | 1 | 2 | 3 | | 0 | 0.214 | 0.225 | 0.232 | 0.000 | | 1 | 0.000 | 0.000 | 0.000 | 0.000 | | 2 | 0.000 | 0.000 | 0.000 | 0.000 | | 3 | 0.000 | 0.000 | 0.000 | 0.000 | |

Table 1: MAPE score for every combination of parameters in ARIMA and SARIMA

As it is observed in Table-1 ARIMA and SARIMA models are insensitive to different parameters. The same pattern was noted from (Liu et al., 2023). The optimal combination to be …. #TODO

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LSTM | | | | | |
|  |  | Number of Neurons | | | |
|  |  | 0 | 1 | 2 | 3 |
| 50 | 0.000 | 0.000 | 0.000 | 0.000 |
| 100 | 0.000 | 0.000 | 0.000 | 0.000 |
| 200 | 0.000 | 0.000 | 0.000 | 0.000 |
| 300 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 2: MAPE score for each LSTM parameter

As it can be seen by increasing the number of epoches the performance of the model increases, that mean that the model is far from overfiting. The only downside of this is that the calculation time also increases as well making it unusabl

To better visualize the distribution of the results of each model, boxplot (box-and-whisker) diagrams are used for MAPE on a quarterly and monthly basis. As for the total water balance, a bar plot is used to represent the results.

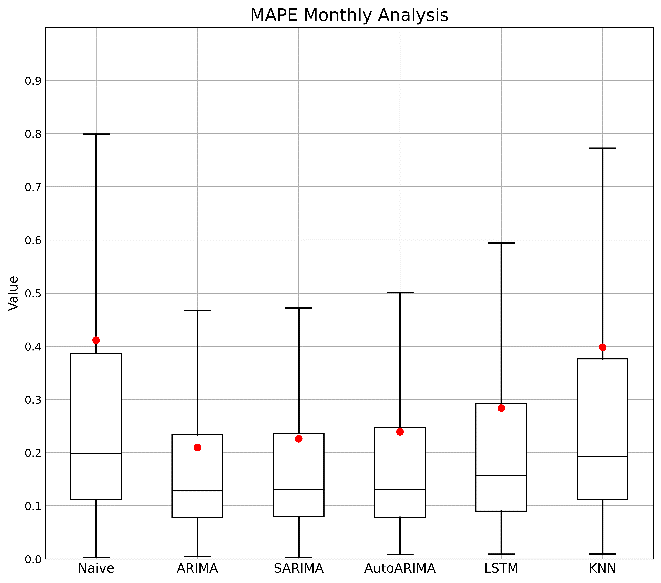
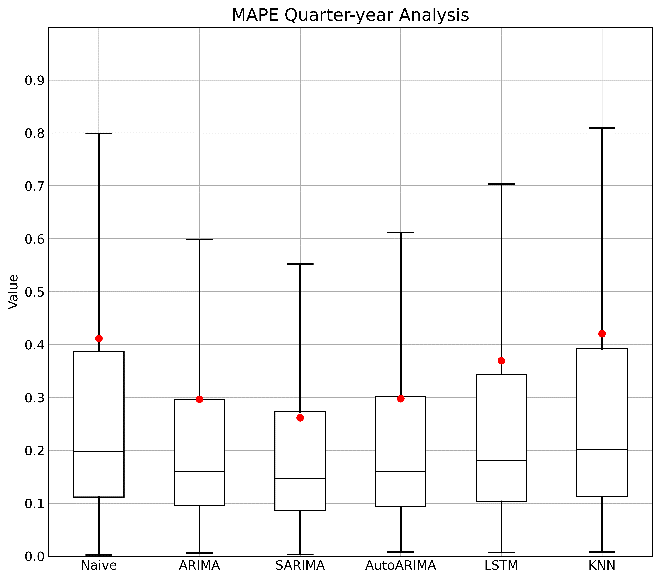


Figure 6: Performance of each model

As seen in Figure 6, all the models performed better on a monthly time scale in contrast to quarter year analysis. As it seems the models had easier time to get adapted to the monthly data rather, unlike the results illustrated by (Kofinas et al., 2014). The finer timescale contributes to larger timeseries in contrast to quarter yearly. The larger set of data gives the models the ability to form their parameters more optimal. (Liu and Lin, 1991) also suggest that a monthly timeframe results in more accurate predictions. The error of the naïve model remains the same because it is not influenced by the time step. The model with the best overall performance is ARIMA(1,1,0) exhibiting an average error close to 20%. The SARIMA(0,1,0) (1,0,1)12 exhibited a comparatively negligible deviation, maintaining similar error distribution as the ARIMA.

The reason for the lower percentage errors achieved by ARIMA and SARIMA models lies in the careful selection of parameters to minimize MAPE, whereas AutoARIMA focuses on minimizing the Akaike information criteria (AIC). LSTM and ARIMA showed the greater reduction in error in respect to the time scale. This was something to expect that LSTM would exhibit due to the substantial number of parameters it possesses. The larger data enables the effectiveness tuning of these parameters. The exact average performance of each model can be seen in Tables 1–2 below.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Models | | Naïve | ARIMA | SARIMA | AutoARIMA | LSTM | k-NN |
| MAPE | Quarter | 0.411 | 0.296 | 0.261 | 0.297 | 0.369 | 0.420 |
| Monthly | 0.411 | 0.209 | 0.226 | 0.239 | 0.283 | 0.398 |

Table 3: Average MAPE performance for Monthly and Quarter year analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Models | | Naïve | ARIMA | SARIMA | AutoARIMA | LSTM | k-NN |
| MAPE | Quarter | 0.197 | 0.160 | 0.146 | 0.160 | 0.180 | 0.201 |
| Monthly | 0.197 | 0.128 | 0.131 | 0.130 | 0.157 | 0.192 |

Table 4:Median MAPE performance for Monthly and Quarter year analysis

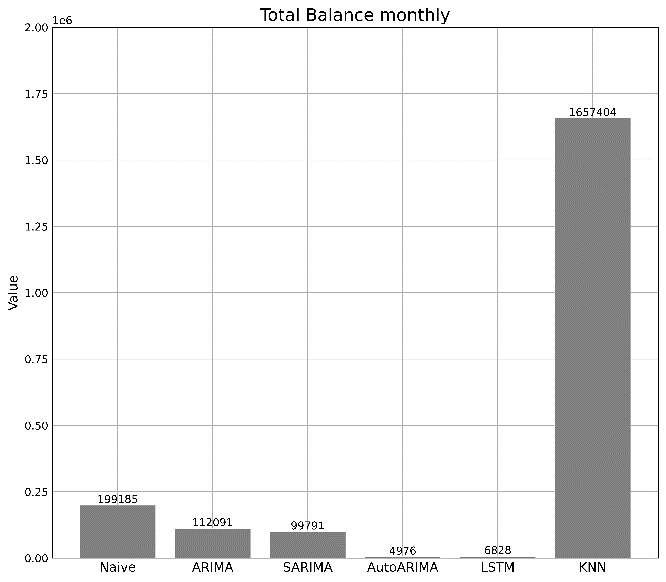
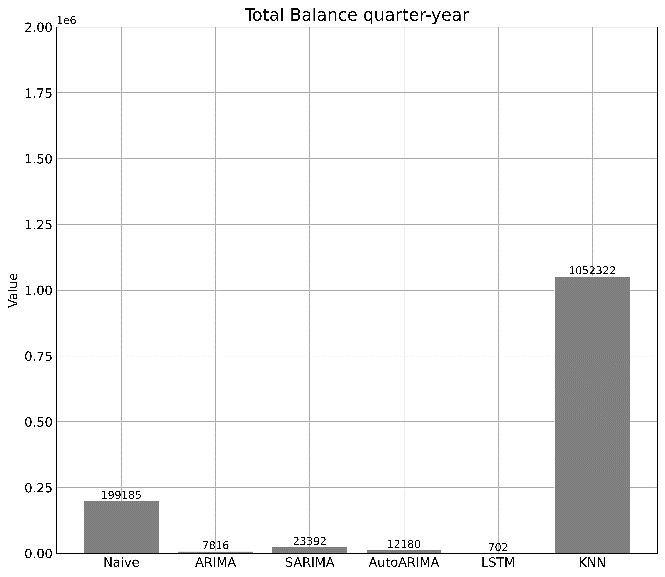


Figure 7: Total Balance performance

Although the MAPE metric favors the monthly time scale, the Total Balance metric tends to favor the quarterly scale, with all models achieving values closer to zero. It is observed that all models underestimate the consumption, with the best one being LSTM, as we can see in Figure 7. The collaborative filtering method achieved the poorest performance by far for those two cases. Finally, for the monthly scale, we measure the performance of each model by each quarter basis throughout the year. This allows us to see if the models exhibit any preference in a specific season or if certain seasons exhibit more predictable data patterns that others.

The results that are shown in Figure 8 confirm the results of Figure 6, illustrating that the best model of every season is the ARIMA. All the models exhibit the same pattern in terms of performance per quarter year. The most predictable period was the second quarter of the year with all models achieving the best performance. The collaborative filtering method did not outperform the base line naïve method. The same issue was addressed in (Kontopoulos et al., 2023) were the clustering approach did not outperform the regular models. One explanation for the poor performance of the collaborative filtering method is the large amount of measured data. The curse of dimensionality is s well documented problem that kNN models suffer when dealing with high dimensions of data (Kouiroukidis and Evangelidis, 2011).

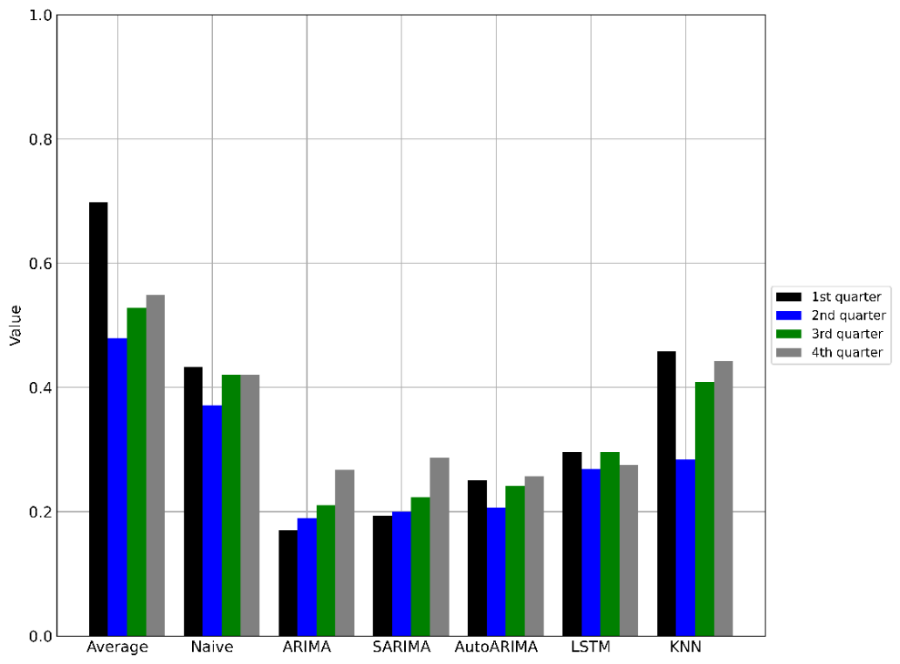


Figure 8: Model performance per season

**Conclusions**

This study presents a comparative analysis of various forecasting models, focusing on their ability to predict customer water consumption based on data from the Water Supply and Sewerage Company of Greece (EYDAP). Due to data inconsistencies, the time series were supplemented with data from water treatment plants, transforming irregular quarterly data into a consistent monthly timescale. These transformed timeseries served as inputs for machine learning models, with the objective of generating accurate future consumption predictions.

The models were evaluated on both monthly and quarterly timescales. The findings revealed that all models achieved lower average errors for the monthly timescale, with the ARIMA(1,1,0) model outperforming the others in terms of percentage error. However, the other statistical models were not significantly inferior. However, the other statistical models were not significantly inferior. Interestingly, the LSTM model, while not the most accurate for individual consumption forecasting, excelled in predicting total water usage, demonstrating its potential for water management applications. The poor performance can be attribute to the existence of small training set in respect to the parameters that they have to be trained. Moreover, the LSTM model exhibited consistent performance across all quarters of the year, indicating low variance between each quarter's error. In contrast, other models showed a preference for the first and second quarters. This consistency further underscores the LSTM model's robustness and reliability.

This work provides insights into the application of time series forecasting algorithms for water demand forecasting. It highlights the potential of these models, particularly LSTM, in enhancing operational efficiency providing key insights for water service providers and enhancing the value of demand timeseries collected for billing purposes. The findings serve as a foundation for future research and practical applications in big data analysis for water utilities.

# REFERENCES

Abbasimehr, H., Shabani, M., Yousefi, M., 2020. An optimized model using LSTM network for demand forecasting. Comput. Ind. Eng. 143, 106435. https://doi.org/10.1016/j.cie.2020.106435

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour. Res. 48. https://doi.org/10.1029/2010WR009945

Bakker, M., Vreeburg, J.H.G., van Schagen, K.M., Rietveld, L.C., 2013. A fully adaptive forecasting model for short-term drinking water demand. Environ. Model. Softw. 48, 141–151. https://doi.org/10.1016/j.envsoft.2013.06.012

Billings, R.B., Jones, C.V., 2011. Forecasting Urban Water Demand. American Water Works Association.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.

Chen, J., Boccelli, D.L., 2018. Forecasting Hourly Water Demands With Seasonal Autoregressive Models for Real-Time Application. Water Resour. Res. 54, 879–894. https://doi.org/10.1002/2017WR022007

Cui, B.-B., 2017. Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm. ITM Web Conf. 12, 04008. https://doi.org/10.1051/itmconf/20171204008

Dikaios Tserkezos, E., 1992. Forecasting residential electricity consumption in Greece using monthly and quarterly data. Energy Econ. 14, 226–232. https://doi.org/10.1016/0140-9883(92)90016-7

Guo, G., Liu, S., Wu, Y., Li, J., Zhou, R., Zhu, X., 2018. Short-Term Water Demand Forecast Based on Deep Learning Method. J. Water Resour. Plan. Manag. 144, 04018076. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000992

Herrera, M., Torgo, L., Izquierdo, J., Pérez-García, R., 2010. Predictive models for forecasting hourly urban water demand. J. Hydrol. 387, 141–150. https://doi.org/10.1016/j.jhydrol.2010.04.005

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Comput. 9, 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huang, M., Bao, Q., Zhang, Y., Feng, W., 2019. A Hybrid Algorithm for Forecasting Financial Time Series Data Based on DBSCAN and SVR. Information 10, 103. https://doi.org/10.3390/info10030103

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: principles and practice. OTexts.

Kofinas, D., Mellios, N., Papageorgiou, E., Laspidou, C., 2014. Urban Water Demand Forecasting for the Island of Skiathos. Procedia Eng., 16th Water Distribution System Analysis Conference, WDSA2014 89, 1023–1030. https://doi.org/10.1016/j.proeng.2014.11.220

Kontopoulos, I., Makris, A., Tserpes, K., Varvarigou, T., 2023. An evaluation of time series forecasting models on water consumption data: A case study of Greece.

Kouiroukidis, N., Evangelidis, G., 2011. The Effects of Dimensionality Curse in High Dimensional kNN Search, in: 2011 15th Panhellenic Conference on Informatics. Presented at the 2011 15th Panhellenic Conference on Informatics, pp. 41–45. https://doi.org/10.1109/PCI.2011.45

Liu, G., Savic, D., Fu, G., 2023. Short-term water demand forecasting using data-centric machine learning approaches. J. Hydroinformatics 25, 895–911. https://doi.org/10.2166/hydro.2023.163

Liu, L.-M., Lin, M.-W., 1991. Forecasting residential consumption of natural gas using monthly and quarterly time series. Int. J. Forecast. 7, 3–16. https://doi.org/10.1016/0169-2070(91)90028-T

Ma, W., Nowocin, K., Marathe, N., Chen, G.H., 2019. An interpretable produce price forecasting system for small and marginal farmers in India using collaborative filtering and adaptive nearest neighbors, in: Proceedings of the Tenth International Conference on Information and Communication Technologies and Development, ICTD ’19. Association for Computing Machinery, New York, NY, USA, pp. 1–11. https://doi.org/10.1145/3287098.3287100

Mu, L., Zheng, F., Tao, R., Zhang, Q., Kapelan, Z., 2020. Hourly and Daily Urban Water Demand Predictions Using a Long Short-Term Memory Based Model. J. Water Resour. Plan. Manag. 146, 05020017. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001276

Rehfeld, K., Marwan, N., Heitzig, J., Kurths, J., 2011. Comparison of correlation analysis techniques for irregularly sampled time series. Nonlinear Process. Geophys. 18, 389–404. https://doi.org/10.5194/npg-18-389-2011

Schafer, J.B., Frankowski, D., Herlocker, J., Sen, S., 2007. Collaborative Filtering Recommender Systems, in: Brusilovsky, P., Kobsa, A., Nejdl, W. (Eds.), The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 291–324. https://doi.org/10.1007/978-3-540-72079-9\_9

Song, X., Liu, Y., Xue, L., Wang, Jun, Zhang, J., Wang, Junqiang, Jiang, L., Cheng, Z., 2020. Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. J. Pet. Sci. Eng. 186, 106682. https://doi.org/10.1016/j.petrol.2019.106682

Stoica, P., Selen, Y., 2004. Model-order selection: a review of information criterion rules. IEEE Signal Process. Mag. 21, 36–47. https://doi.org/10.1109/MSP.2004.1311138