**Comparative Evaluation of Machine Learning Algorithms for Water Consumption Data Analysis and Forecasting**

Short title: Water Consumption Analysis and forecasting

C. Michalopoulos\*^, insert names here\*, 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

\*\* Bodgett & Billham plc, Hamilton House, Hangover Square, London NC1 4TS, UK (E-mail: [*birdie@walford.cotton.co.uk*](mailto:birdie@walford.cotton.co.uk))

^Corresponding author

[Please note all co-authors email addresses need to be included during the submission process, and you should indicate who the corresponding author is]

# Abstract

[The abstract should be no more than 200 words briefly specifying the aims of the work, the main results obtained, and the conclusions drawn. Citations must not be included in the Abstract.]

To handle the variety of time-series datasets seen in many areas, many machine learning algorithms have been designed. In this study, various machine learning algorithms are tested in terms of their ability and reliability to forecast actual water consumption. These methods can be utilized by water utility/distribution companies not only for billing customers who haven't measured their usage accurately and timely but also to approximate the total water losses in the distribution grid by considering the balance between consumed and exported water. The models that are benchmarked are both stochastic and deterministic also a baseline method is covered to compare its performance with the more complex ones. The data used to train and test the models were provided by the Water and Sewage Company of Greece (EYDAP) from over 2 million consumers. Due to data inconsistencies, the original time series had to be reformulated based on the recorded volume of water exported from the water treatment plant for the same period.

# Keywords

[Please include six keywords in alphabetical order. These should indicate the main subject matter of your paper.]

Timeseries forecasting, Machine Learning, Long-Short-Term Memory, collaborate filtering, Timeseries data preprocessing, Water consumption

**Main text:** [for clarity this should be subdivided into:]

**Introduction:** describing the background of the work and its aims.

Trying to understand and predict water demand at the consumer level has been an active research topic since 1960(Reference). With the advent of new electronic water meters capable of monitoring consumption at small time intervals, researchers have developed numerous models and methods for consumption forecasting (Rahim et al., 2020). The majority of the installed meters are still the typical mechanic ones (i.e., single-jet, multi-jet, volumetric)(American Water Works, 1962). Replacing all the mechanical meters with electronic or smart ones is not only financially but also technically unfeasible since the existing infrastructure limitations in most cities (especially historic ones) cannot support the installation of these meters to be able to electronically measure the drawn volume because special probes are needed with a dedicated power supply (Hauber-Davidson and Idris, 2006).

The main disadvantage of the typical water meters is that they require physical reading, which is a time-consuming and labor-intensive process (Randall and Koech, 2019).  Consequently, many water companies face difficulties and lack sufficient resources to read each meter promptly and on a feasible schedule.

Recently, a multitude of approaches in water demand forecasting have been proposed. These methods vary based on factors, such as the type (systematic data frame)  of data available as well as the time scale of the forecast (Kofinas et al., 2014). While most research articles on urban water demand forecasting focus on short-term time frames (hourly and daily), our study focuses on a mid-term timeframe (quarter-year). To address this challenge, we examined studies from other fields that exhibit similar seasonality patterns in time series, such as residential natural gas consumption (Liu and Lin, 1991).

The 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 forecasting of those models is derived from a linear function of the previous observations. These models derive forecasts from linear functions of past 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); the structure of the forecast and the data available (measured and unmeasured customers) resemble a matrix completion problem. The clustering model that is tested is k-nearest neighbors (k-NN), with the variation of incorporating a similarity coefficient as described in (Cui, 2017)

Generally, in water demand forecasting, it is observed that neural network models, and more specifically, LSTM had the best forecasting performance, with statistical models (SARIMA) not falling a much behind (Kontopoulos 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. In this paper, such models were not examined because of their high complexity but also because it is shown inferior performance compared to regular models with the same of time-series data format (Kontopoulos et al., 2023).

Due to data inconsistency, it is common in other research studies to discard time series with missing or differently timed values. To overcome this problem, some techniques have been proposed, including simple interpolation and more complex ones such as kernel-based models (Rehfeld et al., 2011). For this study, the data were adjusted to account for the total water volume exported from the water treatment facility. ("Θα μπει κι άλλο για το data preprocessing")

The goal of this paper is to compare numerous Statistical and Machine Learning models in terms of their accuracy and computational efficiency in predicting water consumption. The predictions are utilized by the water company to bill the customers accurately, but they could also be used to reduce the number of visits that are needed by the water metering crew. Additionally, the predictions will be compared against the commonly used Naïve approach employed by most distribution companies (water, electricity, and gas). Moreover, the models will be assessed for their performance in estimating total water losses during the distribution phase, providing valuable insights for water distribution companies (Mutikanga et al., 2013).

**Methods**: [a brief description of the methods/techniques used (the principles of these methods should not be described if readers can be directed to easily accessible references or standard texts).]

1. 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. In other words, the prediction for this winter quarter-year will be equal to the previous value of last year’s winter quarter-year.

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

where m = the seasonal period, and k = , and denotes the integer part of u. (Hyndman and Athanasopoulos, 2018). This approach is used for reference because of its simplicity and speed.

Β. 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 regression, moving average coefficient, and the white noise variance. 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 studying for monthly analysis, m=12 was selected to align with the yearly seasonality.

C. Long Short Term Memory (LSTM)

Long Short-Term Memory Neural Networks (LSTM) is a special kind of Recurrent Neural Networks 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.

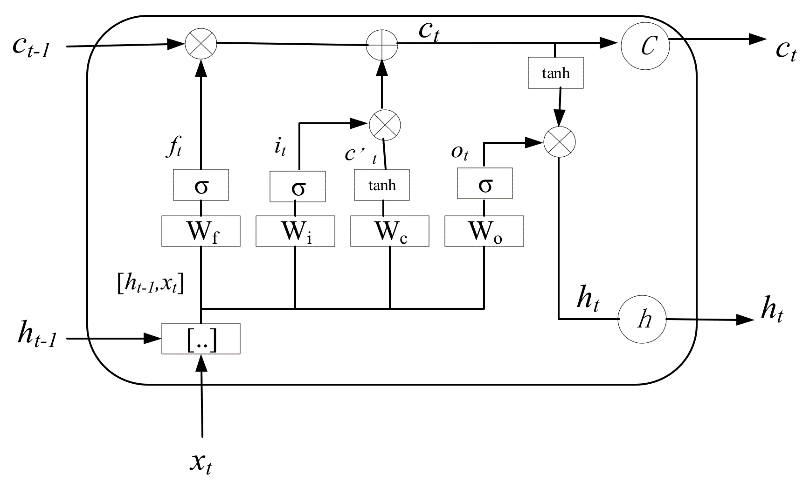


Figure 1: Architecture of a single LSTM block

|  |  |  |
| --- | --- | --- |
|  |  | (3) |
|  |  | (4) |
|  |  | (5) |
|  |  | (6) |
|  |  | (7) |
|  |  | (8) |

where and the weight matrices for the inputs of the network, , are the bias factors, is the input vector, and the hidden state which is also the output of the cell. The calculation process is illustrated in Figure 1. During the training process, the weighted matrices are modified accordingly to capture the important information from the previous values of the time series. 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), the number of units in the hidden layer, and the number of epochs were the parameters tuned through trial and error.

D. 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:

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

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, one limitation of this is that it treats every neighbor the same, despite how close they are or having other similarities. 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). In this research cosine similarity (Cui, 2017) is used and it is expressed as:

|  |  |  |
| --- | --- | --- |
|  |  | (10) |
|  |  | (11) |

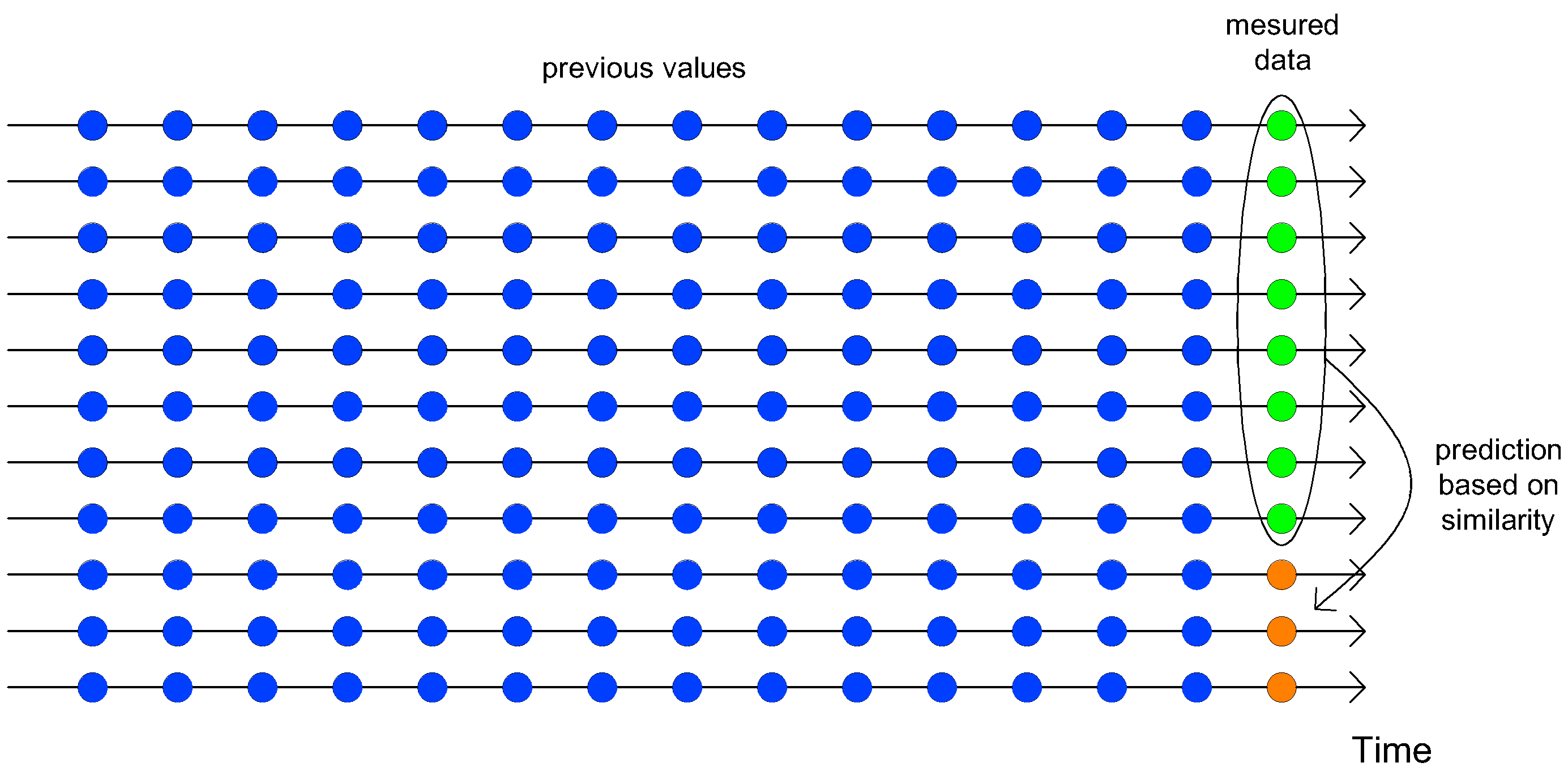


Figure 2: Flow diagram of collaborative filtering approach

Dataset # TODO Fixed

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 guesses. That means 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.

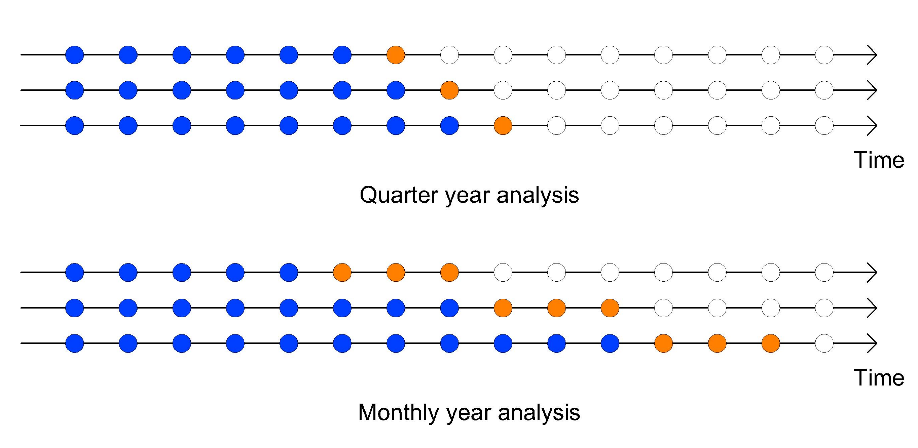


Figure 3:Train and Test data in a timeseries

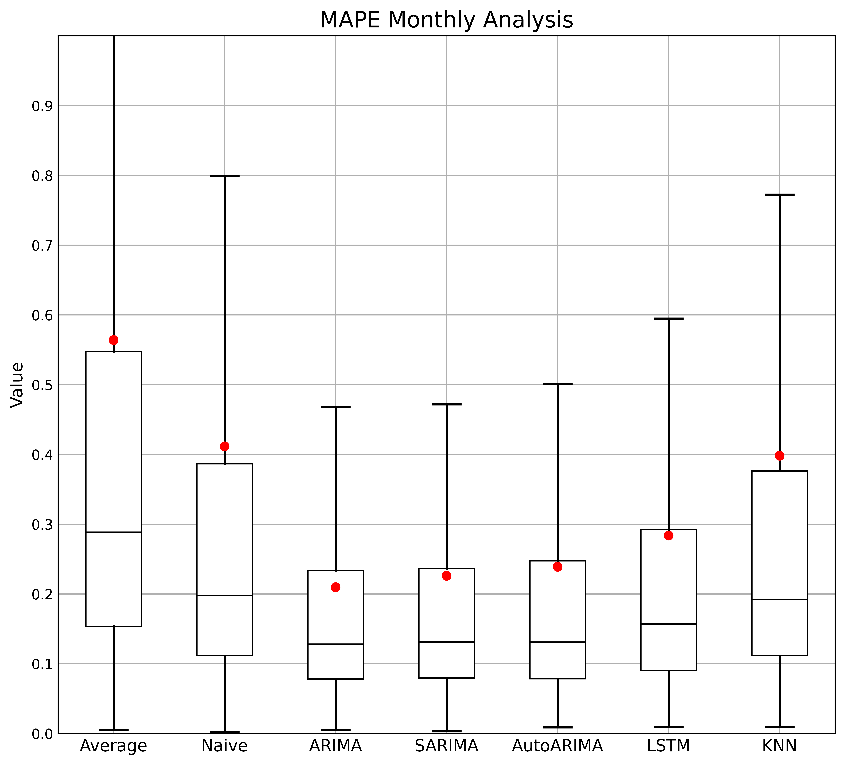
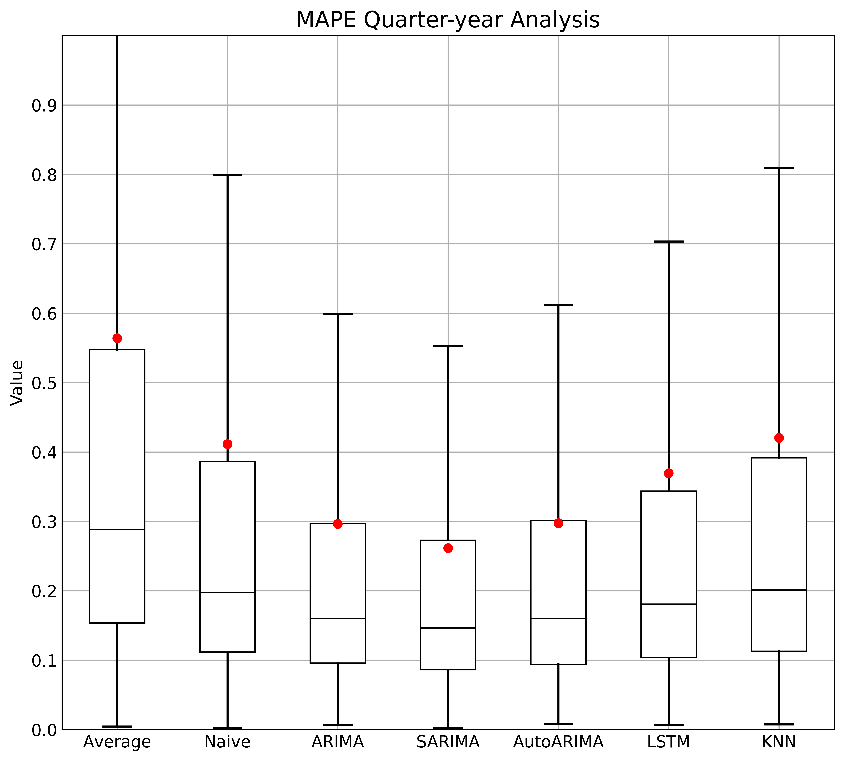
To evaluate the performance of each model and also to simulate the measured and unmeasured behavior of the dataset, the timeseries is divided into two subsets: the measured, where the working crew has managed to record the consumption, and the unmeasured, where the crew did not record the usage. In the actual case study, the measured to unmeasured ratio for the given case study is approximately 90:10. However, to simulate more uncertain cases, the chosen ratio was picked to be 80-20. Specifically, measured was at 80% of the total time series, while the rest was unmeasured. 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 over Mean Absolute Error (MAE) because of the ability to be a 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. 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.

**Results and Discussion**: a clear presentation of experimental results obtained, highlighting any trends or points of interest.

In order 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 4: Performance of each model

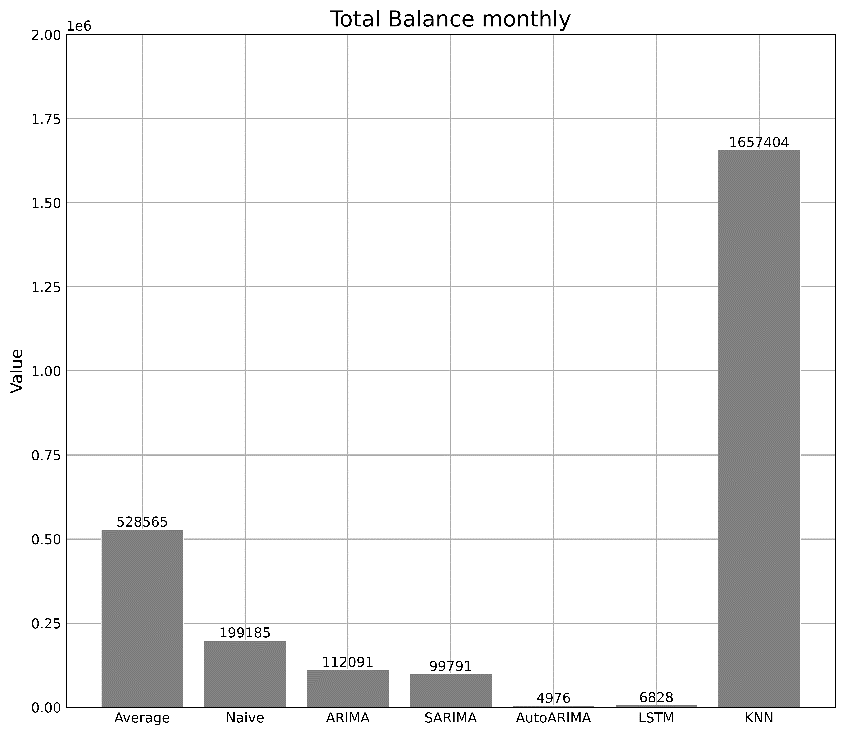
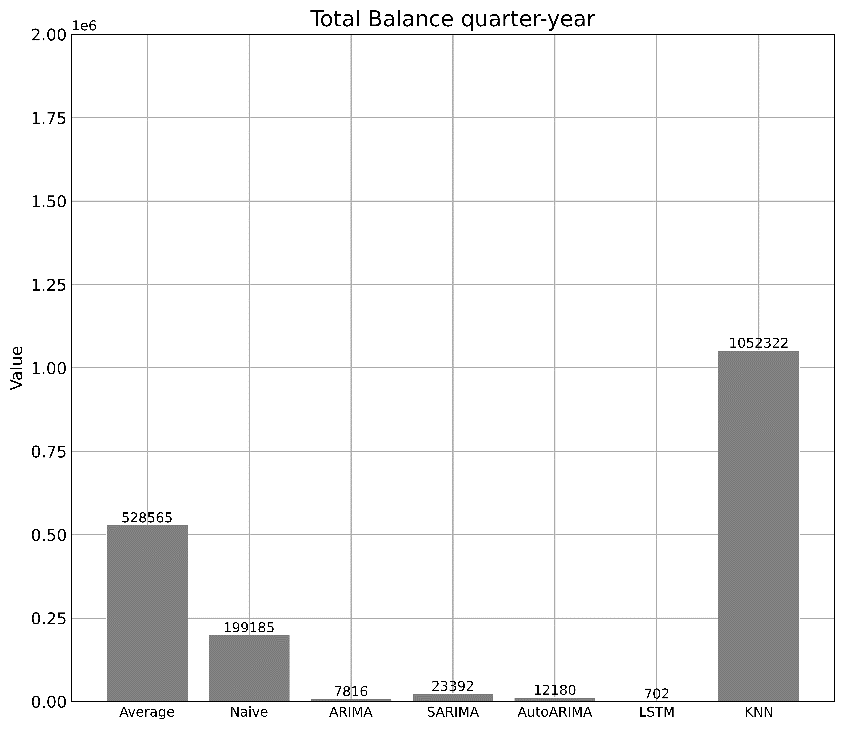


As we can observe in Figure 4, the models performed better on a monthly time scale, despite generating two values based on predictions, unlike the results illustrated by (Kofinas et al., 2014) while (Liu and Lin, 1991) suggest that a monthly timeframe result in more accurate predictions. The reason behind this behavior must be the existence of more data to train the models because, of the finer timescale there is more data for a given time. Obviously, 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 did not fall too much behind, having the same 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 Bayesian and Akaike information criteria. The exact average performance of each model can be seen in Tables 1–2 below.

Table 1: Average MAPE performance for Monthly and Quarter year analysis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 : Median 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 |

Figure 5: 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 3. The collaborative filtering method achieved the poorest performance by far for the two cases. Finally, for the monthly scale, we measure the performance of each model by quarter to see if the models exhibit any time preferences. The results that are shown in Figure 6 confirm the results of Figure 4, and it is illustrated that the best model of every season is the ARIMA. All the models behaved in the same pattern, so it can be said that the models can be fitted more easily with data from the 2nd quarter. One possible explanation for the poor performance of the collaborative filtering method is the large amount of measured data, leading to overfitting issues.

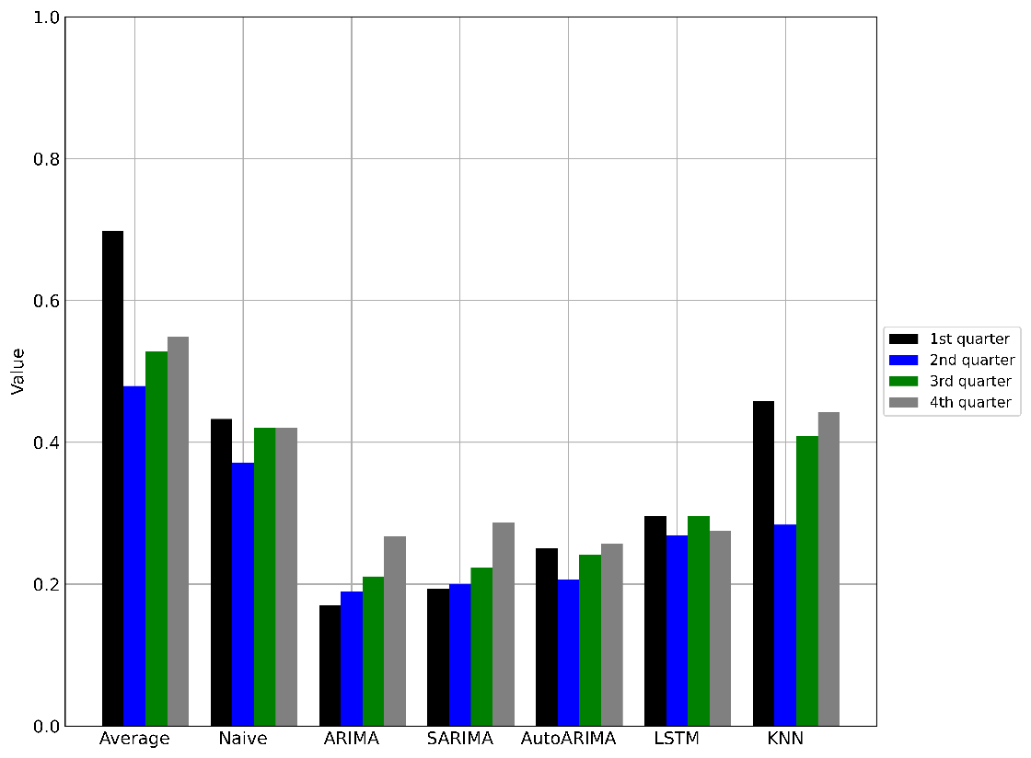


Figure 6: Model performance per season

**Conclusions**: a brief explanation of the significance and implications of the work reported.

In this paper, we compare the effectiveness of the models in forecasting the water consumption of the customers based on the data provided by the Water Supply and Sewerage Company of Greece (EYDAP). Due to inconsistent data, the time series were filled with data from the main refineries, transforming inconsistent quarter-year, to a regular monthly scaled timeseries. These timeseries were fed into the models to mesure the forecasting performance, the same models were examined for quarterly time scale. The results indicated that all models are achieving lower average error for monthly time scale time series, with the best model being ARIMA(1,1,0). The other statistical models were not fall further behind. The LSTM model might not be the most accurate model for forecasting individual consumption, but it performed phenomenal in the total Balance metrics, being the best model for predicting water usage as a total. Furthermore, the model that did not show any preference for a specific quarter of the year was LSTM showing low variance between each quarter error, while the statistical models show some preference for the first and second quarter of the year. Overall, this study aimed to evaluate the performance of various time series forecasting algorithms for essential tasks that water utility companies have to deal with, such as customer billing and leak detection in the distribution system. The results have provided valuable insights and information on improving and advancing the overall performance of water management.

**References**: these should be to accessible sources. Please ensure that all work cited in the text is included in the reference list, and that the dates and authors given in the text match those in the reference list. References must always be given in sufficient detail for the reader to locate the work cited (see below for formats). Note that your paper is at risk of **rejection if there are too few (<10) or too many (>25) references,** or if a disproportionate share of the references cited are your own.

Please make sure to cite all references in the paper

**Supplementary Material**: Appendices and other Supplementary Material are permitted, and if the paper is accepted they will be published online only. A link to the supplementary material will be provided in the print version.

**Figures** should be embedded in the paper, as well as being supplied as a gif, jpeg, tie file at the end of the paper. They should be a minimum of 300dpi for readability.

**Tables** should be included in an editable format and not as images. Number tables consecutively in accordance with their appearance in the text and place any table notes below the table body. Be sparing in the use of tables and ensure that the data presented in them do not duplicate results described elsewhere in the article. Please avoid using vertical rules.

# REFERENCES

The reference should be arranged according to the alphabetical order by the lead author’s last name. Please make sure to include all authors of references.

# Examples of Journal References

Andrews, J.F. 1993 Modeling and simulation of wastewater treatment processes. *Wat. Sci. Tech.* **28** (11/12), 141–150.

Casey, T.G., Ekama, G.A., Wentzel, M.C. and Marais, G.v.R. 1993 An hypothesis for the causes and control of low F/M filamentous organism bulking in nitrogen (N) and nutrient (N & P) removal activated sludge systems. In *Proc. of the IAWQ First Int. Conf. on Microorganisms in Activated Sludge and Biofilm Processes*, Paris, 27–28 September.

Dold, P.L., Ekama, G.A. and Marais, G.v.R. (1980) A general model for the activated sludge process. *Prog. Wat. Tech.* **12**, 47–77.

# Examples of Book References

Bell J. 2002 *Treatment of Dye Wastewaters in the Anaerobic Baffled Reactor and Characterisation of the Associated Microbial Populations*. PhD thesis, Pollution Research Group, University of Natal, Durban, South Africa.

Henze M., Harremoës P., LaCour Jansen J. & Arvin E. 1995 *Wastewater Treatment: Biological and Chemical Processes*. Springer, Heidelberg.

McInerney M. J. 1999 Anaerobic metabolism and its regulation. In: *Biotechnology*, J. Winter (ed.), 2nd edn, Wiley-VCH Verlag, Weinheim, Germany, pp. 455-478.

Sobsey M. D. & Pfaender F. K. 2002 *Evaluation of the H2S method for Detection of Fecal Contamination of Drinking Water*, Report WHO/SDE/WSH/02.08, Water Sanitation and Health Programme, WHO, Geneva, Switzerland.

*Standard Methods for the Examination of Water and Wastewater* 1998 20th edn, American Public Health Association/American Water Works Association/Water Environment Federation, Washington DC, USA.

# Example of an Online Reference

Alcock S. J. & Branston L. 2000 SENSPOL: Sensors for Monitoring Water Pollution from Contaminated Land, Landfills and Sediment. <http://www.cranfield.ac.uk/biotech/senspol/> (accessed 22 July 2005)

A more detailed description can be found on each journal’s online Instructions to Authors page: <https://iwaponline.com/pages/Instructions_for_authors>

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

American Water Works, 1962. Water Meters: Selection, Installation, Testing, and Maintenance. 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.

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

Hauber-Davidson, G., Idris, E., 2006. SMART WATER METERING.

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

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.

Lim, B., Zohren, S., 2021. Time-series forecasting with deep learning: a survey. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 379, 20200209. https://doi.org/10.1098/rsta.2020.0209

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

Mamo, T.G., Juran, I., Shahrour, I., 2013. Urban Water Demand Forecasting Using the Stochastic Nature of Short Term Historical Water Demand and supply Pattern. J. Water Resour. Hydraul. Eng. 2.

Mutikanga, H.E., Sharma, S.K., Vairavamoorthy, K., 2013. Methods and Tools for Managing Losses in Water Distribution Systems. J. Water Resour. Plan. Manag. 139, 166–174. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000245

Rahim, M.S., Nguyen, K.A., Stewart, R.A., Giurco, D., Blumenstein, M., 2020. Machine Learning and Data Analytic Techniques in Digital Water Metering: A Review. Water 12, 294. https://doi.org/10.3390/w12010294

Randall, T., Koech, R., 2019. SMART WATER METERING TECHNOLOGY FOR WATER MANAGEMENT IN URBAN AREAS. Water E-J. 4, 1–14. https://doi.org/10.21139/wej.2019.001

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