**Timeseries forecasting for water consumption using Collaborative Filtering**

Short title: Water Demand forecasting using collaborative Filtering

A. B. Fish\*^, C. D. Flesch\*, and E. Fowle\*\*

\* Department of Superficial Research, University of Life, 99 Crow Road, Addlestrop, Borsetshire BB3 1TC, UK

(E-mail: *fish@imaginaryaddresss.com*; *stag@bay.matrix.edu.uk)*

\*\* 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.]

The Athens Water Supply and Sewerage Company (EYDAP), is the largest active company in Greece in the water market. EYDAP's clientele in the field of water supply includes about 4,400,000 customers (2,160,000 unique connections/meters)

Due to the wide area of activity and the large number of customers, EYDAP is not able to measure all the installed water meters with the same frequency as water bills have to. From the total connections about 200,000 of them remain “un-measured” each quarter-year (800,000 in year). This paper describes and measures the performance of Statistical and Machine Learning models in forecasting the consumption of the customers that their meter did not measured properly. The statistical models that are studied are ARIMA, SARIMA and Gaussian Mixture Model while the Regression models are Feedforward Neural Networks, Long–Short–Term-Memory (LSTM) and Seasonal Naïve. The models in this research are trained and are tested with real data of 2,107,000 consumers from the city of Athens.

at the same rate as the water bills are

# Keywords

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

Clustering, Forecasting, Grid Losses, Machine Learning, Water Billing, Water Demand

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

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

Trying to understand and predict the water demand in the consumer level has been an active research topic since 1960(Reference). Nowadays, with the introduction new electronic water meters, which can monitor the consumption within very small time intervals (give examples of such meters), many 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)(Association, 1962). Replacing all the mechanic meters with electronic/smart ones is not only financially but also technically unobtainable, since the current infrastructure in most cities (especially historic ones) cannot support the installation of these meters as in order to be able to electronically measure the drawn volume because, special probes are needed with a dedicated power supply(Hauber-Davidson and Idris, n.d.).

The major drawback of the typical water meters is that they have to be physically read, a process that is time and labor consuming(Randall and Koech, 2019). Hence, the majority of water companies struggle and sometimes do not have adequate resources to actually read every single meter in a timely manner and under a feasible schedule.

Recently, a multitude of approaches in water demand forecasting have been proposed. The methods are varying depending on numerous factors such as the type quality (systematic data frame)(Mamo et al., 2013) of data available, as well as the time scale of the forecast(Kofinas et al., 2014). In terms of urban demand water forecasting the majority of research articles are focused more on short-term time frame (hourly and daily) but in our case the timeframe of research is mid-term(quarter-year). In order to overcame this problem studies from other fields were examined that behaved similarly in term with the seasonality of the timeseries, such as residential natural gas consumption(Liu and Lin, 1991).

The models are addressed into two categories stochastic and deterministic. Deterministic models take into consideration all the factors that inference the final result and are built finding the patterns between those factors. On the other hand, the majority of stochastic models are generated with the help of statistical models that are adapted on the previous data in the time-series(Box et al., 2015).

The most common stochastic models are: the autoregressive (AR), the moving average (MA), the combination of those two with an integration step 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. In order to overcome this problem and to be able to describe more complex behaviors long short-term memory neural networks (LSTM) have been proposed(Lim and Zohren, 2021). The introduction of nonlinear activation functions in the neural network architecture will overcome the problem that regular stochastic models have. Another deterministic method that is used 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) is similar to a matrix completion problem. The clustering models that are tested are k-nearest neighbors (KNN) and gaussian mixture models (GMM).

Generally, in water demand forecasting it is observed that neural network models and more specific LSTM had the best forecasting performance with statistical models (SARIMA) not failing a lot behind.(Kontopoulos et al., 2023). Also, another method that is proposed in the literature is forecasting with clustered customers which only one model is trained for a group of customers. In this paper such models were not examined because of its high complexity and bad overall forecasting performance.

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 could also be used to reduce the number of visits that are needed by the water metering crew. Also, the predictions will be compared with the most widespread method that the majority of distribution companies (water, electricity and gas) use the Naïve. The results will also be compared computationally, that mean that the time that is needed in order to run will be considered.

**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

With this method each forecast is set to be equal to the last value from the same season. For example, 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 its speed.

Β. ARIMA-SARIMA

ARIMA(p,d,q) is the combination of AR,I and MA models. AR(p) is Autoregressive, p is the number of regression terms, ΜΑ is the moving average, q is the number of moving average terms, and d is the difference time to make the data stationary series. 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 of the ARIMA(p,d,q) are estimated by least square method and moment estimate method. In order to express more seasonal phenomena, the SARIMA model have been developed which takes into consideration the seasonality of the problem. The additional parameters of the SARIMA model are (P,D,Q),m in which are the seasonal order terms while m is the seasonality. The models 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 it is chosen m=12 in order to follow the seasonality of a year.

C. Long Short Term Memory (LSTM)

Long Short Term Memory Neural Networks (LSTM) is a special kind of Recurrent Neural Networks which are designed to eliminate the standard problems RNN suffer(Hochreiter and Schmidhuber, 1997) . Each individual LSTM cell contains three different gates: input gate, forget gate and output gate. With this architecture we can increase the length of sequence without worrying about gradient vanishing or exploding problem. The internal architecture of a typical LSTM cell is shown below:

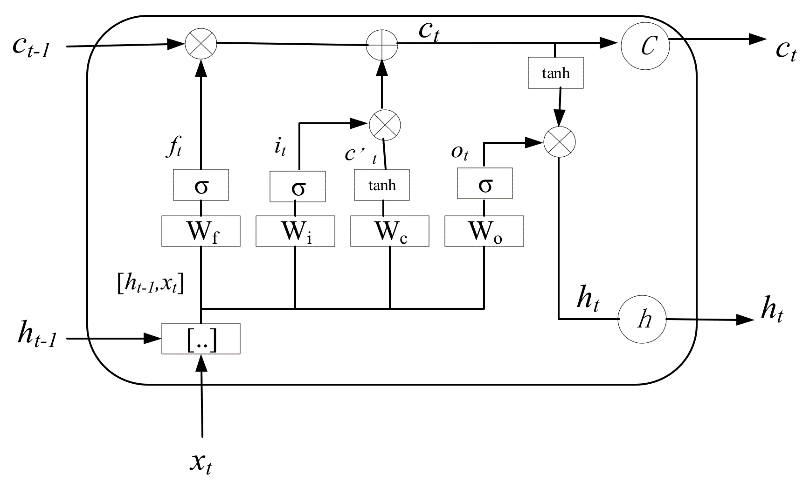


Figure 1: Architecture of a single LSTM block

|  |  |  |
| --- | --- | --- |
|  |  | (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. Throw out the training process the weighted matrices are modified accordingly in order to decode the important information from the previous values of the timeseries. Tuning LSTM models is not that easy, there are a lot of hyperparameters that must be tuned in order to achieve the optimal performance (Abbasimehr et al., 2020). For this study the only parameters that were tuned by trial and error was the lag (the number of previous data that it was feed to the LSTM, the number of units in the hidden layer and the number of epochs).

D. k-Nearest Neighbors

Trying to predict all the unmeasured meters it is possible for someone to thing that this is a problem of collaborative filtering in which we know the last consumption of the majority of the customers and we wish to forecast the consumption on those who did not. For our today needs they have been developed numerous methodologies in this field with the leading one being recommendation algorithms. In this research we will investigate if it is possible to make forecasts from the water meters that have already been measured. To be able to achieve this we must be able to represent each customer as a point in space. Because this analysis is in quarter-year scale we choose to work with one for each quarter. Each dimension is calculated as mean from the last years same quarter-year. Each dimension is calculated:

where N the number of years that were consider and the value of consumption. It is obvious that for the customers that we do not know the last value will be blank.

There are a lot of known algorithms for collaborative filtering in this research are tested the two most common ones k-Nearest Neighbors and Gaussian mixture models.

D1. k-Nearest Neighbors

k-Nearest Neighbors is the simpler supervised machine learning algorithm with only one hyper-perimeter k being the number of closest points that we taking information. kNN is primarily used as a clustering algorithm but it can be used as a forecasting one (Altman, 1992). In this research they are used to variation of the algorithm, in the first one the forecast is made by the mean of its neighbors and the other with a weighted factor that is proportional to his distance. The first equation describes the calculation process of the average values of the neighbors and the second with weighted neighbors.

where is the last value of the known customers, is the distance from the ith neighbor and k the number of nearest neighbors.

D2. Gaussian Mixture Models

Gaussian mixture model is the linear combination of multiple Gaussian distributions. The main advantage of this models is that it can fit any type of distribution even if the population is described as multiple different distributions (Yan and Tang, 2019). Gaussian mixture distribution is defined as:

where the weight, the mean value and the covariance matrix of each group. Because the point we are using are in the 4th dimension the normal distribution will be defined as:

In order the mixture models to be a probability density function the following condition must be followed

After the definition of the total number of groups based on the customers that we already measured. The probability of each customer that we have to forecast to be addressed in each group is calculated: (**bad use of English)**

the notion is to point that we did not consider the 4th dimension (the clustering was made from the 3 other that were knowned). After calculating the probability of each customer to all the groups the forecasting is calculated.

This method was inspired by a popular approach in recommendation systems(Yan and Tang, 2019)

Performace Measurement crireria

The metric functions that were used in this research was three. Mean absolute percentage error (MAPE), root mean square (RMSE) and Total Difference.

where and are the actual and predicted value respectively. MAPE and RMSE are the most common accuracy measuring functions (Zhang and Qi, 2005) while Total difference is there to measure the total deviation of the model for the total of the customers (shows if the model tends to over or under estimate the consumptions). It is worth mentioning that using MAPE may cause implications for our case because is not impossible to be equal to zero. (θελουμε site)

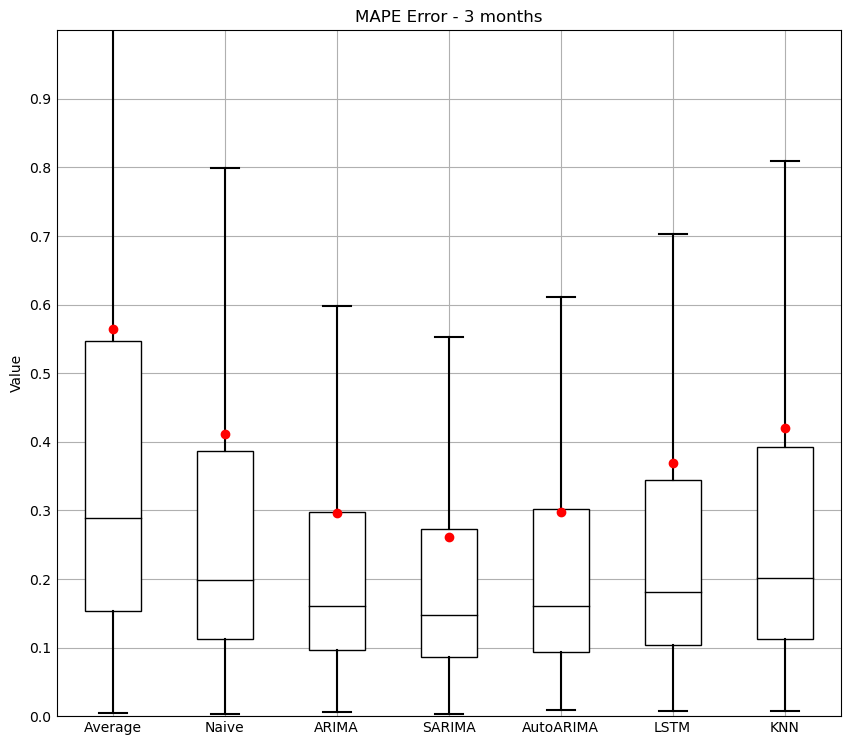
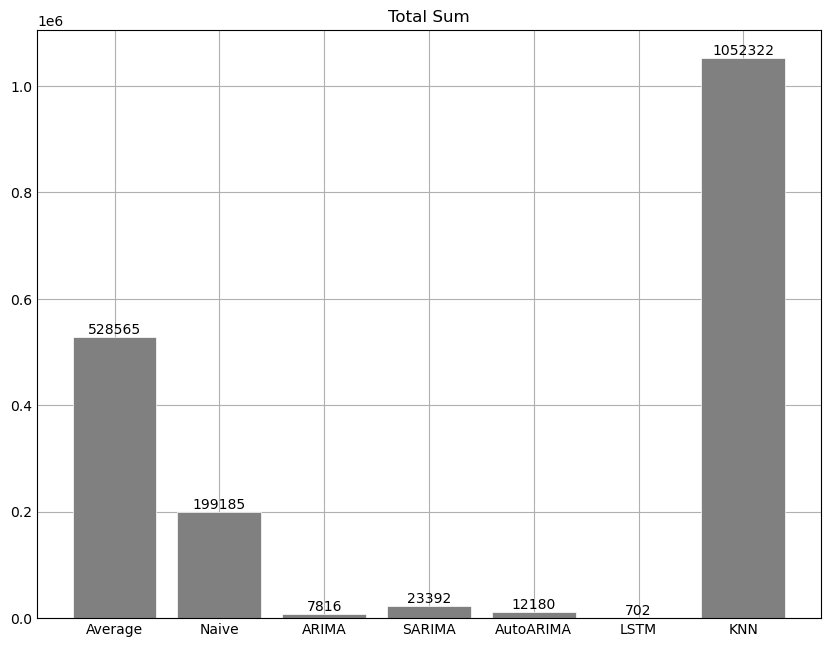
Dataset

In order to be able to evaluate the performance of each model and also to simulate the measured and unmeasured behavior of the dataset, the timeseries are split in two subsets, the measured whose the working crew has managed to record the consumption and into unmeasured were the crew did not record the usage. In reality the actual measured to unmeasured ration for the given case study is approximately 90-10 but in order to simulate more uncertain cases the chosen ration was picked to be at 80-20, to be more specific the measured was at 80% of the total timeseries while the rest 20% was the unmeasured. Furthermore, the unmeasured timeseries are broken down to train set and test set for measuring the performance for each model. The metrics that are used to evaluate the performance are:

1. Mean Absolute Percentage Error (MAPE) this metric was chosen instead of 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 is the metric that evaluates the tendency of the model to over or under estimate the forecasting and it is expressed as: . The closer to zero this metric is the more accurate this model is to predict for the total sum of the consumers.

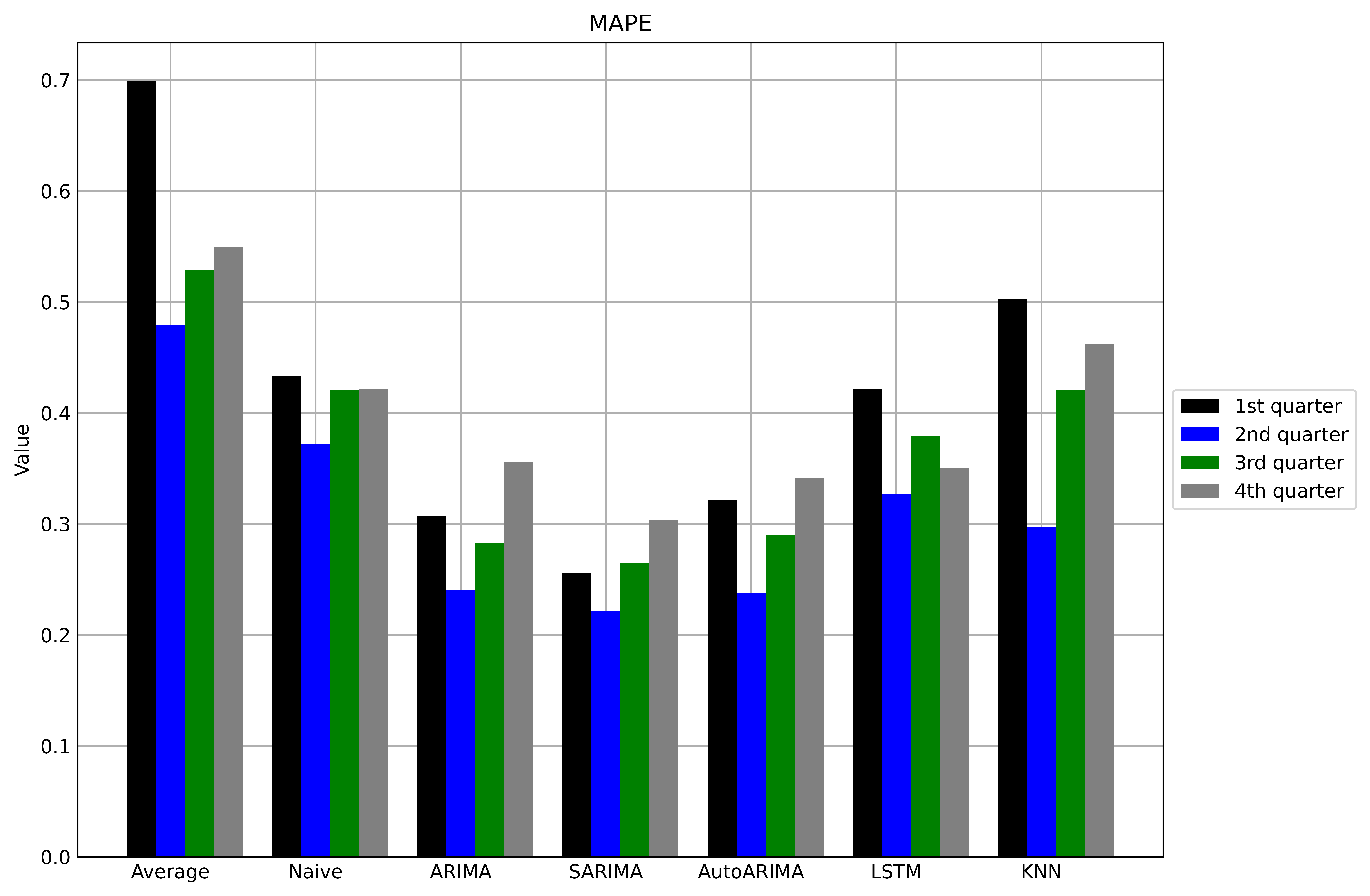
**Results and Discussion**: a clear presentation of experimental results obtained, highlighting any trends or points of interest.

In order to visualize better the distribution of the results of each model we used boxplots (box-and-whisker) diagrams for MAPE for quarter and monthly year bases. As for the total water balance a bar plot is used to represent the results.

From the results we can observe that the model with the best performance is SARIMA in terms of Mean Absolute Percentage Error (MAPE) as it is shown in figure (). Despite scoring the best performance in MAPE metric, in the Total balance metric it came 3rd with LSTM scoring only +702 m³ for the total of predictions. It must be addressed the large difference between the average (it is showed as a red dot in figure ()) from the median. This can be explained from the existence of customers with very high variance and unexpected consumption because of a leaks or because they just started a new business for example and the models does not have enough data in order to adapt to the sudden change. While the collaborate filtering algorithms use future data from other consumers did not manage to outperform the Naïve model scoring the worst score in Total Balance. One possible reason that this model failed to give accurate predictions is the large number of known timeseries and it faced overfitting phenomena.

Because of the large data that were available it was consider appropriate to investigate the performance of each model for each quarter year of demand. The results are showed in figure ().



As we can see all the models have a similar behavior in respect to season of forecast. Because of this finding we can assume that the models do not have a preference in the season that we want to forecast but we data itself are more consistent in this timespan and that explains the performance improvement in the k-NN method at the second quarter. The best in term of MAPE performance is still SARIMA scoring the lowest error for all quarter years of forecasting. Finally, the same models were tested for a monthly timestamp. In order to see if by decreasing the timestep thus, increasing the available information, we can achieve more trained models for our application.

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

**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, tif 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

Altman, N.S., 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. Am. Stat. 46, 175–185. https://doi.org/10.1080/00031305.1992.10475879

Association, A.W.W., 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.

Hauber-Davidson, G., Idris, E., n.d. 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

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.

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

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

Yan, H., Tang, Y., 2019. Collaborative Filtering Based on Gaussian Mixture Model and Improved Jaccard Similarity. IEEE Access PP, 1–1. https://doi.org/10.1109/ACCESS.2019.2936630

Zhang, G.P., Qi, M., 2005. Neural network forecasting for seasonal and trend time series. Eur. J. Oper. Res. 160, 501–514. https://doi.org/10.1016/j.ejor.2003.08.037