Prediction of drinking water needs : the case of Bobo-Dioulasso

Pierre KONATE[1], Abdoulaye SERE[2], mamadou DIARRA[3]
{prrkonat15@gmail.com[1], abdoulaye.sere@recifaso.org[2], diarra.md21@gmail.com[2] }

Université Nazi BONI, Bobo Dioulasso 01 BP 1091, Burkina Faso[1,2,3]

Abstract. The objective of this study is to develop a solution to predict daily water consumption in order to optimize the water management system in the city of Bobo-Dioulasso. To achieve this, neural networks are used to predict consumption using historical consumption data and daily temperature in the city as the parameters. Four neural network algorithms were implemented for this study: Multi-Layer Perceptron (MLP), simple recurrent neural network, Long Short-Term Memory (LSTM) recurrent neural network, and Gated Recurrent Unit (GRU) recurrent neural network. The study focused on the eight distribution zones of the National Water and Sanitation Authority in the city. In view of the results, certain algorithms stood out from others in terms of prediction. The GRU network algorithm performed better on nearly half of the training data, followed by the MLP algorithm. The resulting models allow for the prediction of daily consumption on D-day, given the consumption and temperature of D-day-1. This work was carried out using tools such as the Jupyter NoteBook environment and the Python language.

Keywords: Machine learning, data analysis, prediction.

1 Introduction

The management of drinking water becomes increasingly complex over the years in view of the considerable increase in demand for domestic, industrial. uses. This strong demand is the result of the growth in the number of populations in urban and rural areas, resulting in limited access to water resources. This complexity leads us to have other management methods and techniques that can facilitate the management of the system.

To face this complexity, predicting water demand is an effective means of managing the system. For that, knowledge of information on water use (i.e. data) in the past is necessary for prediction. Indeed, from the modeling of coagulant dosage in treatment stations [3], to water monitoring through sensors [2], to solving water management problems [1] specifically in consumption, it is necessary to have a sufficient amount of historical data to facilitate prediction. The availability of prediction tools is also necessary in search of a solution.

JRI 2023, December 18-20, Ouagadougou, Burkina Faso

DOI 10.4108/eai.18-12-2023.2348136

Managing the distribution system is not an easy task because the methods used by managers of the National Water and Sanitation Office (ONEA) are traditional and do not meet the needs expressed following new developments. The classic method used is a statistical estimation (ONEA, 2019) to have the daily volume of water necessary for each distribution zone. These methods are therefore outdated and less reliable, and new methods must be proposed that can meet current needs. Among these new available methods, we have neural networks.

Compared to classical statistical methods, neural networks have a great capacity for solving prediction problems. Thanks to their complex structure composed of artificial neurons, neural networks have a great capacity for analyzing data in order to make predictions. They have a structure composed of several layers, which in turn are composed of a certain number of neurons. The neurons in each layer are connected in such a way as to work based on the information received. The more neurons increase, the more neural networks gain in solving complex problems. Each neuron performs a simple processing or calculation but this complex power results from their interaction [7].

Depending on the desired objective, there are different models of neural networks whose data processing requires the interaction of the different neurons that make up the model. The models used in our study are those whose processing is based on quantitative data. This choice of model is therefore due to the fact that the data collected from ONEA are quantitative data.

The remainder of this document will include a description of the methods used, data collection, data processing and presentation of results followed by a discussion.

2 Method Description

The neural network models used in the search of solutions are four (04) in number, including multi-layer perceptrons, simple recurrent networks, LSTM networks, and GRU networks.

2.1 The Multi-Layer Perceptron (MLP)

The MLP is a neural network with layers whose activation functions of the same layer are identical [7]. The MLP network consists of three layers; the input layer, the hidden layer, and the output layer (see figure 1). Each layer has an activation function, which helps process the received information. This information processing is carried out using the backpropagation method ([4]), which allows for the processing of information from the input to the output layer, as well as from the output to the input layer.

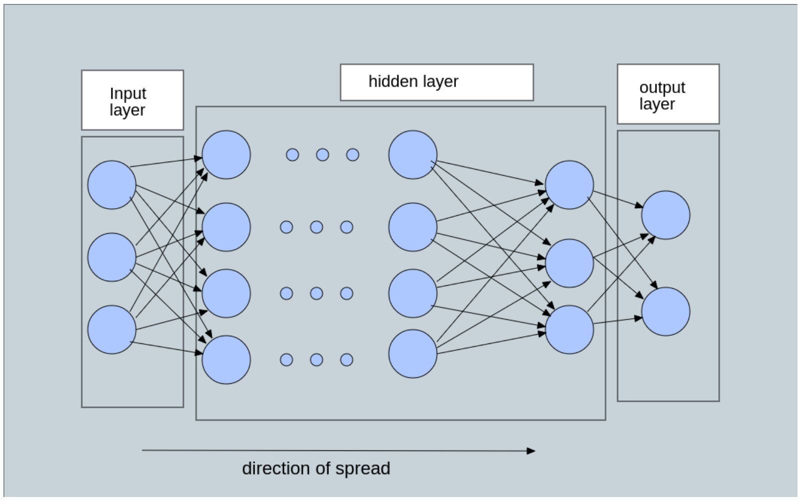


Fig. 1. The architecture of a multi-layer perceptron.

2.2 Simple recurrent neural networks

Recurrent neural networks are a type of cyclic neural network that perform exceptionally well in processing sequential data (such as time series data, video and audio streams).

Simple recurrent neural networks are a basic version of recurrent networks. Similar to multi-layer perceptrons, recurrent network algorithms use the backpropagation algorithm ([5]) to process sequential data. Their operation is very similar to that of MLP algorithms by associating the time steps all running in a recurrent way.

Fully recurrent neural networks (FRNNs) establish connections between all neurons' outputs and inputs. This topology represents the most general neural network structure, as it can encompass all other network configurations by selectively nullifying connection weights to simulate the absence of specific inter-neuron connections. The accompanying illustration on the right may lead to potential misinterpretations, as practical neural network structures are often visually organized into 'layers,' giving the impression of distinct, separate layers. However, what may initially appear as layers actually signifies different time steps within the same fully recurrent neural network. The illustration's leftmost component reveals the recurrent connections, denoted as the 'v' labeled arc, which are temporally 'unfolded' to create the layer-like appearance.

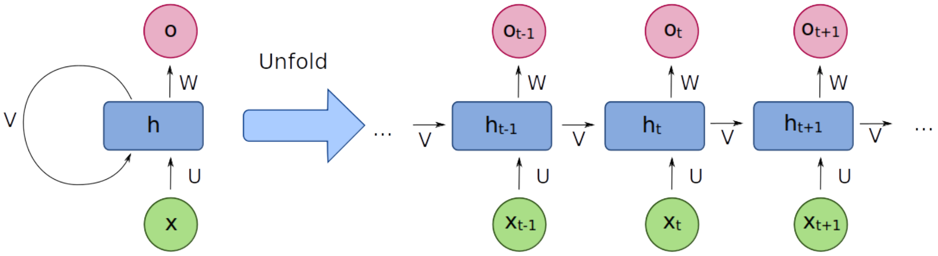


Fig. 2. Schema of a RNN. [8]

2.3 Long Short-Term Memory (LSTM)

The LSTM is a highly complex type of neural network with two memories, one long-term and one short-term. This network structure consists of four interacting elements that allow for the addition and removal of information. These elements are gates that represent the network's layers, including the Input Gate, Forget Gate, Memory, and Output Gate. The key feature of LSTMs is their ability to manage long-term dependencies.

The equations of the LSTM model are defined by the set of the following equations ([6]):

$$\begin{aligned}
F_t &= \sigma\left(b^F + x_t U^F + h_{t-1} W^F\right) \\
I_t &= \sigma\left(b^I + x_t U^I + h_{t-1} W^I\right) \\
O_t &= \sigma\left(b^O + x_t U^O + h_{t-1} W^O\right) \\
F_t &= \sigma\left(b^F + x_t U^F + h_{t-1} W^F\right) \\
c_t &= F_t c_{t-1} + I_t tanh\left(b + x_t U + h_{t-1} W\right) \\
ht &= tanh\left(c_t\right) O_t \\
\hat{y} &= g\left(b^{\hat{y}} + h_t W^{\hat{y}}\right)
\end{aligned}$$

Where F corresponds to the ”forget gate,” the gate responsible for updating the memory cell c. I and O correspond to the ”input” and ”output” gates, allowing the input and output of information in the LSTM based on the input data x and ht-1, which represents the hidden state, the output vector of the LSTM at time step t-1. The vectors U and W correspond to the weights associated with the input data and recurrent weights associated with the hidden layer at time step t-1. g represents an activation function used to obtain ŷt, the LSTM's prediction or classification at time step t.

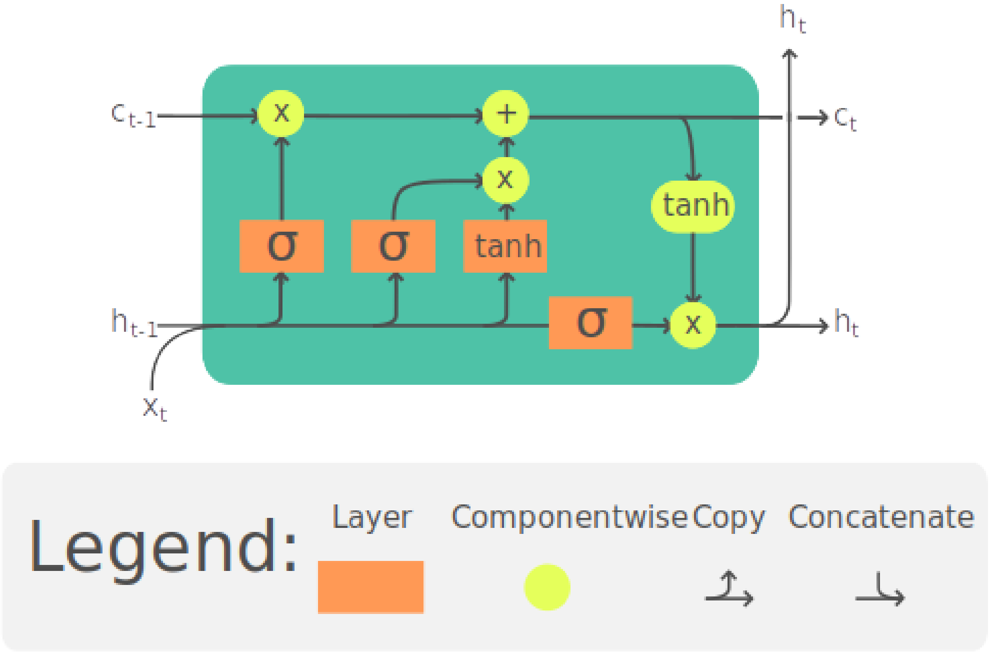


Fig. 3. Schema of a LSTM. [8]

2.4 Gated Recurrent Unit (GRU)

The GRU model is a simplified version of the LSTM network model. It also has the capability to manage long-term temporal dependencies, but its structure includes two gates; the Gate reset and the Gate update. Its uniqueness lies in its ability to reduce the learning time of a model, making it faster in computations. The figure 4 shows the diagram of a GRU.

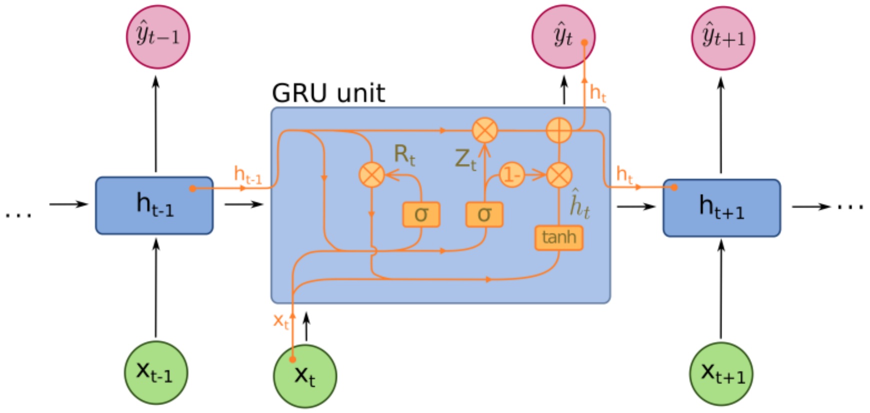


Fig. 4. Schema of a GRU. [6]

3 Data collection

The data used are daily water consumption of data that were collected from the National Water and Sanitation Office (ONEA) of Bobo-Dioulasso. In fact, the Bobo-Dioulasso area and its surroundings are served with drinking water by ONEA. For effective water management, the Office has divided the city and its environs into eight (8) water distribution zones. The data on the amount of water distributed for each zone constitute the study data for the period from November 2018 to November 2021. In addition to consumption data, temperature data were of great interest in this study. To do this, temperature data for the Bobo-Dioulasso area for the same period from November 2018 to November 2021 were collected from the official platform of the agency responsible for meteorology in Burkina Faso.

The data made available to us has undergone a cleaning process in order to obtain reliable and consistent data. At the end of this cleaning, nine variables were retained for the remainder of the study, namely data on:

- The distribution of KUA

- The distribution of LOW SARFALAO,

- The distribution of HIGH SARFALAO,

- The distribution of BAMA,

- The distribution of BELLE VILLE,
- The distribution of LAFIABOUGOU,
- The distribution of BOLO VERS VILLE,
- The distribution of BELLE VILLE QUILTING,
- Maximum temperature.

After statistical analysis, we obtain the results (summaries) in the table 1 :

Table 1: Statistics on the parameters used.

	Mean	Standard deviation	Min	Max
KUA	8044.157	1320.111	360.0	10000.0
LOW SARFALAO	5326.553	744.506	528.0	6000.0
HIGH SARFALAO	1804.199	553.291	8.0	2500.0
BAMA	18543.119	2903.644	225.0	21000.0
BELLE VILLE	1730.155	899.705	70.0	3000.0
LAFIABOUGOU	6313.376	1195.902	2.0	7000.0
BOLO VERS VILLE	5065.460	1059.768	195.0	6000.0
BELLE VILLE QUILTING	2445.469	230.131	135.0	3000.0

4 Data processing

Data processing is a process of model learning that involves cleaning and normalization to make data of good quality. Cleaning is the stage of the process that takes up more than half of the data processing time. It makes the data reliable, consistent, and valuable. Normalization, on the other hand, allows data to be placed on the same scale without the differences in value ranges being distorted and without any loss of information. In this case the Z-score normalization is the normalization method used, it consists of subtracting the average of the data from the raw value and dividing it by the standard deviation (equation 1).

Before the normalization phase, data is separated into training data and test data. However, the proportion of data does not have a defined rule, as researchers have suggested for similar studies ratios of 75% and 25% or 80% and 20%, respectively, for learning and testing [1]. In this study, data from each zone and temperature underwent the processing process, from cleaning to normalization, passing through the separation of data into proportions.

$$X_{norm} = \frac{X_i - \bar{X}}{\sigma_x} \tag{1}$$

where:

- X_{norm} : normalized value
- X_i : gross value
- $\bar{X}$: average of the data
- σ_x : standard deviation

5 Results and discussions

Upon completion of data processing, we obtain data ready for model training. Thus, models were trained on the various data obtained using the four neural network algorithms mentioned earlier. The data from each zone was passed through each of the four algorithms, and models were obtained after these trainings.

5.1 Prediction results of Belle ville

The data on consumption in the Belle ville area and the temperature in the locality are the two input parameters for each of the algorithms. Of all the models obtained, the GRU algorithm achieves better results and Figure 5 shows the variations in the predicted data and those of the observations.

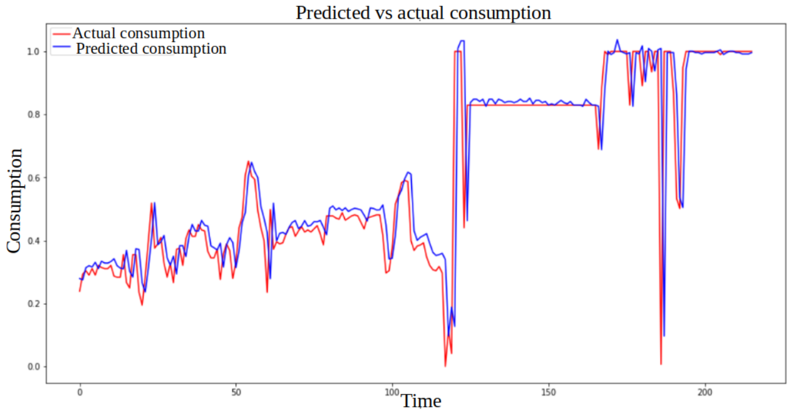


Fig. 5. Variation of observations and predicted values of Belle ville

5.2 Prediction results of Lafiabougou

The forecast of consumption on D-day required the use of the temperature and consumption of the previous day (D-day-1) as input parameters for each of the models obtained after learning the different algorithms.

The figure 6 illustrates the variation of observed values and predicted values of the same model

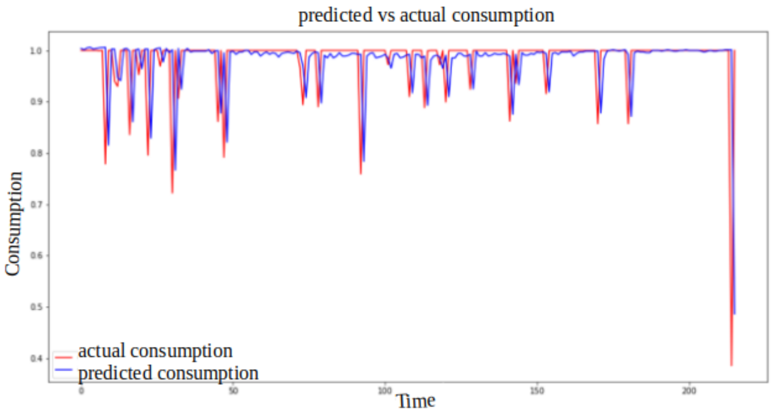


Fig. 6. Variation of observations and predicted values of LAFIABOUGOU

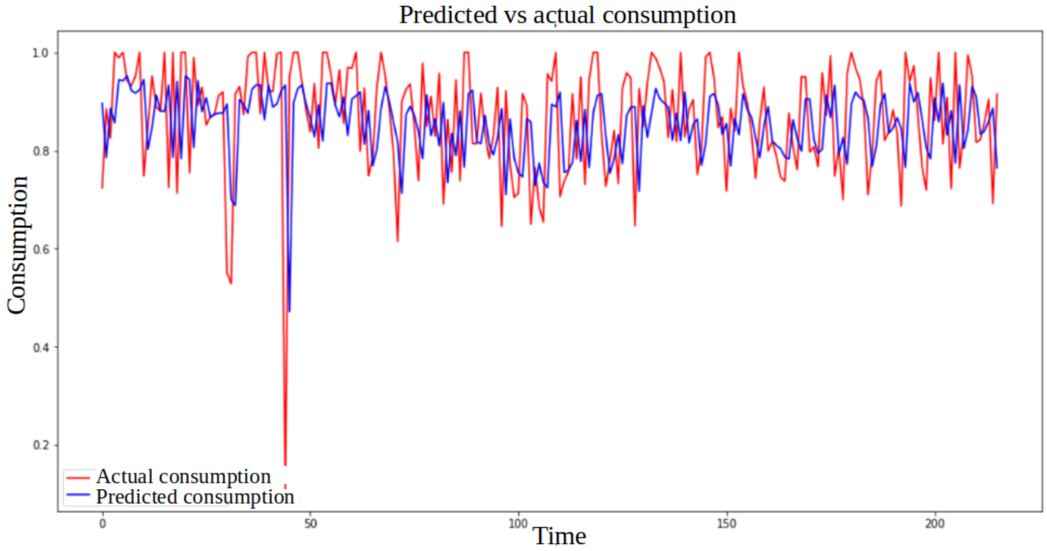


Fig. 7. Variation of observations and predicted values of Bolo vers ville

5.3 Prediction results of bolo vers ville

Among all the algorithms employed for modeling on Bolo vers ville data, the best model is attained using the GRU algorithm with input parameters consisting of consumption data from the area and temperature data. Figure 7 displays the variation between the predicted and observed data.

5.4 Prediction results of HIGH SARFALAO

The data used being the temperature and daily consumption in the HIGH SARFALAO area, the forecast model obtained takes as input parameters the temperature and consumption of the previous day (D-day-1) and the output of the model is the consumption of the D-day.

The variations of the actual data and the predicted data are also represented in the figure 8 :

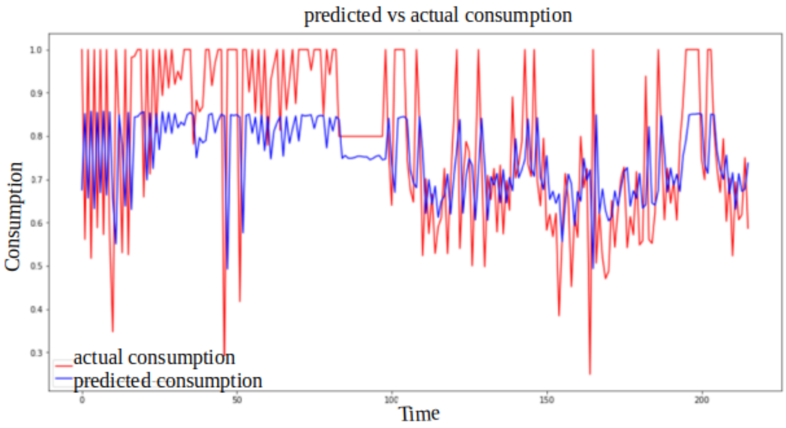


Fig. 8. Variation of observations and predicted values of high Sarfalao

5.5 Discussions

The entire training process revolved around eight (8) datasets, each containing 1091 rows, using four (4) neural network algorithms, including MLP, RNN, GRU, and LSTM. Multiple rounds of experiments were conducted on each algorithm, during which the evaluation criteria consisted of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE).

At the end of validation, the models are ready to make predictions on the data. The relevance of these models is based on the difference between the predicted and observed data, which is determined by the RMSE and MAE errors. Some algorithms outperformed the others, such as the GRU, MLP, and LSTM algorithms. In the case of the GRU algorithm, it achieved superior results on most of the data. The average MAE errors obtained during validation are the lowest, with values of 0.1389 for high Sarfalao, 0.0708 for Belle ville, and 0.0341 for Lafiabougou data. As for the MLP algorithm, it yielded the lowest average RMSE errors of 0.0804 for BAMA data and 0.0884 for Bolo vers ville. These two algorithms are ranked as the best in terms of MAE errors. Overall, the GRU algorithm provides better predictions for most of the data used.

The table 2 presents the various best values obtained for Mean Absolute Errors (MAE) after training with different algorithms.

Table 2: Table recapitulating the MAE error values of the different algorithms after evaluation.

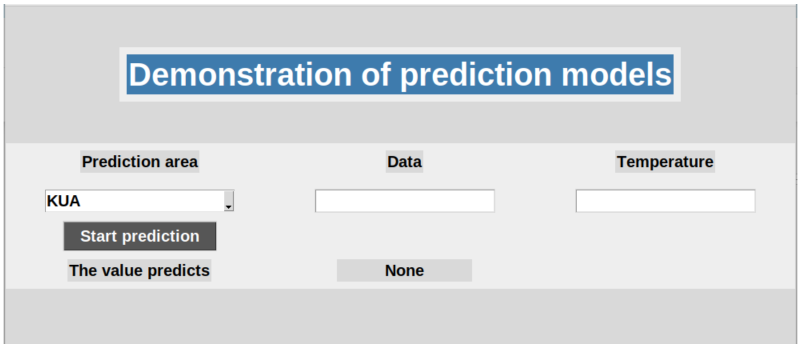
Model / Data	MAE			
	MLP	Simple RNN	LSTM	GRU
KUA	0,0856	0,0847	0,0849	0,0861
LAFIABOUGOU	0,0838	0,0957	0,0801	0,0341
LOW SARFALAO	0,0906	0,0841	0,0834	0,0841
HIGH SARFALAO	0,1515	0,1659	0,1584	0,1389
BAMA	0,0804	0,0855	0,0512	0,0870
BELLE VILLE	0,2272	0,2265	0,2260	0,0708
BOLO VERS VILLE	0,0884	0,0934	0,0908	0,0908
BELLE VILLE QUILTING	0,0978	0,4493	0,4271	0,0173

6 Application

At the end of the training of the various models, we exported the models in files which allowed us to carry out a phase of simulation starting from the random data of consumption and temperature. During this phase, we implemented a small interface (figure 9 et 10) allowing us to facilitate the simulation. This figure is composed as follows:

- a drop-down list to select a given model,
- a consumption data entry area,
- a temperature entry zone,
- a button to press to ”start prediction”,
- an area that displays the result of the prediction,

This phase ended with the setting up of a table (confer table 3) containing the prediction of the consumption of D-day from the random data of the D-day-1



Demonstration of prediction models

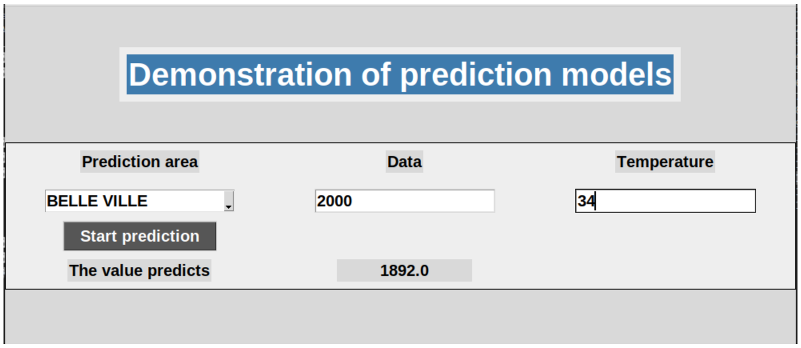
Prediction area: KUA

Data:

Temperature:

The value predicts: None

Fig. 9. Prediction interface



Demonstration of prediction models

Prediction area: BELLE VILLE

Data: 2000

Temperature: 34

The value predicts: 1892.0

Fig. 10. BELLE VILLE prediction interface

Table 3: Simulation result of the different models.

	Day J-1		Day J
	consumption	Temperature	consumption
KUA	8000	34	7914
LOW SARFALAO	4500	34	5320
HIGH SARFALAO	2500	34	2191
BAMA	19000	34	18701
BELLE VILLE	2000	34	1892
LAFIABOUGOU	5500	34	5688
BOLO VERS VILLE	4500	34	4711
BELLE VILLE QUILTING	1500	34	1644

7 Conclusion and perspectives

Conclusion

This task was the place for us to apply algorithms in the field of water. The results obtained allowed us to compare the performance of the different algorithms on each of the eight (8) data used. At the end of the work, the models allow predicting daily water consumption, a solution that shows the importance of neural networks in the hydraulic field. The solution thus makes it possible to optimize water management and save resources for a better future.

Through this study, we have covered the essentials of the technologies of the machine learning which are used more and more nowadays. These technologies have importance in the human's life because they help to improve these living conditions. Several of them are increasingly essential in changes in certain areas of life. This is the case of our field of study which see the intervention of these technologies in the improvement and optimization of its management. The tools used in the implementation allowed us to learn of advantages on their importance and to know their capacity in the implementation place forecasting models.

Perspectives

Although we conducted a study to find solutions for improving water management through technology, it is worth noting that these solutions are proposed with the aim of optimizing the management of drinking water. Indeed, during the data collection process from ONEA, we were able to identify the management issues and offer implementable solutions. Thus, the implementation of a data storage system is proposed to gather consumption data from distribution areas. This system will facilitate data input for forecasting models, providing readily usable data. Furthermore, a study can be conducted to predict the quantity of water consumed while taking into account demographic factors, which can be a contributing factor to limited access to drinking water. Regarding long-term solutions, once the distribution system is optimized, it would be prudent to establish a water leak control system, as water leaks represent a considerable loss of a resource with limited access.

References

[1] Dr Pezon , Christelle and Nansi, Juste and Bassono, Richard. De l'accès aux systèmes de distribution d'eau potable à l'accès aux services d'eau potable : méthode et outils, IRC (Avril 2012).

[2] Mohamed, Bouamarand and Mohamed, Ladjal. Système multicapteur utilisant les réseaux de neurones artificiels pour la surveillance des eaux potables, LASS Laboratoire d'Analyse des Signaux et Systèmes Université de M'sila Algérie (2007).

[3] BELOURGHI, Bariza and HOUICHI, Larbi and HEDDAM, Salim. Reseaux De Neurones Artificiels Pour La Modelisation Du Dosage Du Coagulant Dans Les Stations De Traitements Des Eaux De Surface a Faible Turbidite, Université de Batna, Algérie (septembre 2012).

[4] Nutini, Julie. Feedforward Neural Nets and Backpropagation, University of British Columbia (septembre 2016).

[5] Gelly, Gregory. Reseaux de neurones recurrents pour le traitement automatique de la parole, Université Paris-Saclay préparée à l'Université Paris-Sud (septembre 2017).

[6] Toque, Florian. Prévision et visualisation de l'affluence dans les transports en commun à l'aide de méthodes d'apprentissage automatique, UNIVERSITÉ PARIS-EST, Ecole doctorale MSTIC (décembre 2019).

[7] Zaki, Sabit Fawzi Philippe. Classification par réseaux de neurones dans le cadre de la scattérométrie ellipsométrique, These de doctorat de l'universite de lyon opérée au sein de l'université jean monnet de saint étienne, Ecole Doctorale N° 488 Science-Ingénierie-Santé, p. 68, (Décembre 2016).

[8] Valentin NOËL. Séries temporelles et réseaux de neurones récurrents, Ecole normale superieure Paris-saclay (2022).