MLRA2021
MACHINE LEARNING & RISK ASSESSMENT
Wrocław, Poland, 15 – 17 March 2021

Groundwater time-series forecasting model based on real data

Xin Yin*[1], Zhen Jiang[1] and Qingyu Zhang[2]

[1] School of Civil and Transportation Engineering, Hohai University, Nanjing 210098, China.
E-mail: yxin@hhu.edu.cn

[2] School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China.
E-mail: bzzqyfy@126.com

Keywords: Time-series forecasting; LSTM; Groundwater;

1 Introduction

The environmental issues of engineering construction have gradually attracted the attention of geotechnical engineers. One of these issues is to preserve the natural ground water level. A decreasing ground water level due to constructing activities will have detrimental effects on nature and buildings in a broad zone along the project.

In order to carry out the correct actions to mitigate a decreasing ground water table during construction, it is vital to have information about what will be the natural ground water level, un-affected by construction activities, which is a time-series forecasting problem. The actual measured values in the construction period can then be compared with the forecasted values in order to highlight construction influence on the local ground water table.

Machine learning technique is playing an important role in time-series forecasting problem over the years. An LSTM model is presented here for the realistic forecasting problem of the combined railway and road project called FRE16, located in the southeastern part of Norway. We utilize the dataset collected from installed sensors along the megaproject between 2018-8-10 and 2020-6-29 as training set and validation set to train the model, and to make the prediction of six target sensors' data trends in the next three months. To improve the accuracy of the LSTM model, the position information and historical data are taken into account. Consequently, we obtained an LSTM model to forecast the un-affected ground water information.

2 Methodology

2.1 Data preprocess

The training dataset contains sensor measurements from 240 sensors, located along the planned line for the combined railway/highway project. Six types of data are provided, namely the pore pressure, the water level in lake, the ground water in core drilled hole, the air pressure, the temperature and the precipitation. The task is to perform time-series forecasting on two groundwater sensors of the first three types, in total 6 sensors, for a total period of 90 days (2020-6-30 to 2020-9-27). Data of 25 sensors in this period are provided as features. So we chose these 25 sensors' contemporaneous data and 6 sensors' historical data as inputs to train our model. Besides, due to the fact that few data between 2017-03-08 to 2018-8-9 are available, we leave this part of dataset out of account. Other missing values

are filled by linear interpolation. Then a 31-column dataset is obtained as input feature and a 6-column dataset is obtained as label.

2.2 LSTM model

In this study, a groundwater time series forecasting model was built based on the Long short-term memory neural network (LSTM) and Pytorch deep learning framework. The structure of proposed network consists of 2 LSTM layers and 2 fully connected layers sequentially, and the ReLU function is adopted as activation function between the connected layers. Each LSTM layer contains 100 neurons. With regard to the loss function, we chose the root mean squared logarithmic error (RMSLE) as an evaluation metric since the RMSLE will be less affected by deviations in error-terms compared to the traditional RMSE. To ensure the effect of historical data onto the prediction result, the LSTM prediction models are established for all characteristic series from contemporaneous data and historical data of sensors. We applied the sliding window to select training and validation dataset in model training process. Length of the sliding window is temporarily set to 90 and the sliding step size is 30, which means using the data of the past 3 months to predict the data of the next 3 months. The reason why we adopt this strategy is trying to predict enough data at one time. Theoretically, this strategy can predict the sequence data for a long time in the future, but the accuracy and credibility are issues that must be paid attention to. Considering that the amount of historical data is quite insufficient, the optimizer Adam is adopted in the training process. In the process of validation, 10-fold cross validation is used to estimate the skill of the model on new data.

3 Research Outcomes

As the LSTM model is set up, the prepared training and validation datasets are fed into the program for training. In order to achieve the best performance of the model, hyperparameters of the model are tuned according to RMSLE. For now, the learning rate is set to 0.0001, the batch size is set to 5, the epoch is set to 100 and the decay rate is set to 0.00001.

After sufficient training and proper validation, the average train loss turns out to be 2.29e-4 and the average valid loss is 6.41e-4. The training result in last 90 days and the prediction result are shown in Fig 1-6.

As is shown in Fig 1 and 2, the predictions of ground water accord well with the actual measured value in the previous two months, while the difference between the prediction and measurement in the 3rd month appears to be more obvious. Such trends might be indicative of the accumulating error of long-term dependencies, even though the LSTM layers are adopted. From Fig 3 and Fig 4, it can be found that the error between the predicted value and the measured value is small but the model did not capture the trend of the pore pressure data well. This situation could be attributed to the small variance of the dataset. As for the water level, Fig 5 shows the result of VS_ABBOR sensor, which is relatively similar in data trends. Similarly, the error gradually increases with time in Fig 6. The error can be further reduced by increasing the epoch, the batch size and the number of LSTM layers during the training process, but this will also increase the computational cost.

The evidence from this study suggests that it is reasonable to carry out the time-series forecasting for two months at a time. Smaller sliding step may improve the accuracy of the model, since it shortens the time period for a single prediction. By doing this, you need to input the predicted value obtained last time as the measured value into the model for the next prediction. Moreover, weighting the feature data based on the distance between the feature sensors and the target sensors could be another method to improve the performance of the model. In general, the model presented in this study is acceptable as a reference.

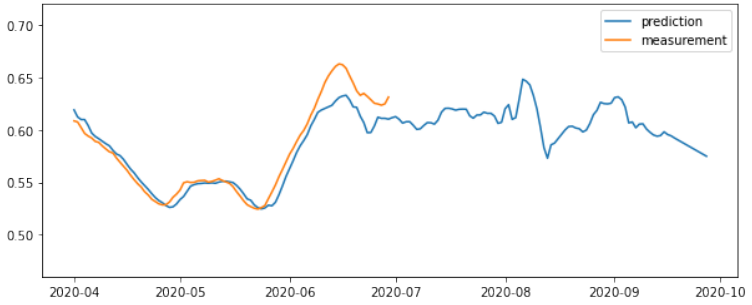

Figure 1. 4C09006 (ground water)

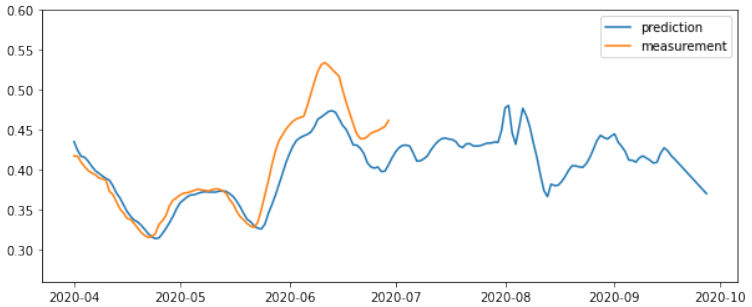

Figure 2. 4C11009 (ground water)

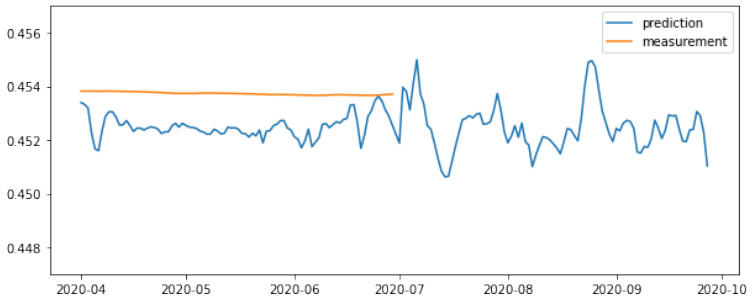

Figure 3. 5G09018Z_A (pore pressure)

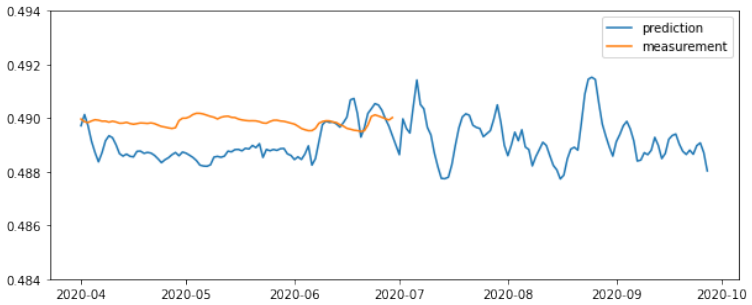

Figure 4. K02_A (pore pressure)

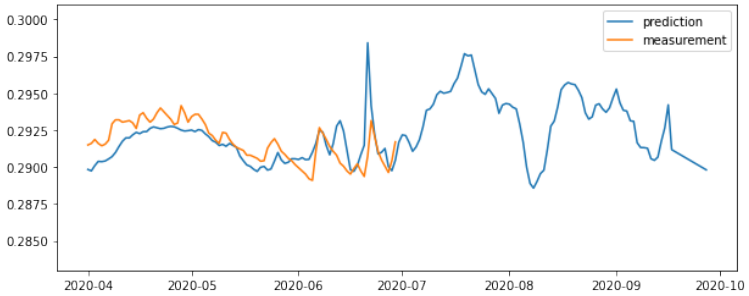

Figure 5. VS_ABBOR (water level)

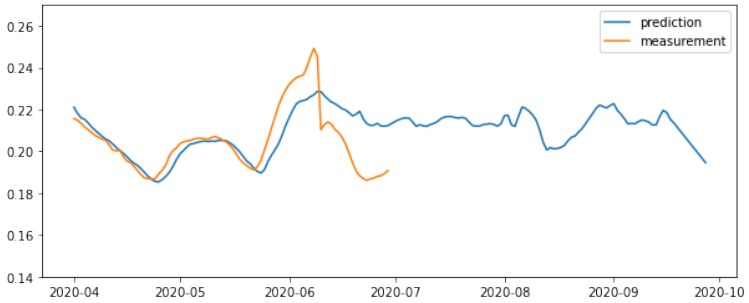

Figure 6. VS_Kroksund (water level)

References

Vu, M. T., et al. "Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network." Journal of Hydrology 597 (2021): 125776.

Gers, Felix A., Douglas Eck, and Jürgen Schmidhuber. "Applying LSTM to time series predictable through time-window approaches." Neural Nets WIRN Vietri-01. Springer, London, 2002. 193-200.

Manav Sehgal. Titanic Data Science Solutions from https://www.kaggle.com/startupsci/titanic-data-science-solutions