Utilize temperature and precipitation as features to predict water quality of drinking fountains at NYC parks

DS-UA 301

Team: Drinking Quality

Luca Wang & Siqi Wang

Abstract:

This project predicts water quality in New York City's public drinking fountains located in parks to promote outdoor activities and reduce greenhouse gas emissions. We built and trained Random Forest, DNN, RNN, and LSTM models and found that the LSTM and RNN models outperformed others. The study demonstrates the potential of machine learning methods for water quality prediction and emphasizes the importance of public health aspects of climate change mitigation.

Background:

Accurate prediction of water quality can benefit the public in three key ways: protecting public health, encouraging outdoor activities, and reducing greenhouse gas emissions. Predicting the water quality of fountains in public parks is especially important, as park visitors of all ages, including children and the elderly, rely on these fountains for drinking water. Poor water quality can lead to health issues such as gastrointestinal illness, skin rashes, and infections. By predicting the water quality in these fountains, park visitors can be alerted to potential risks and take appropriate precautions, such as bringing their own water or using a different fountain. In addition to promoting public health, predicting fountain water quality can also encourage people to spend more time outdoors, which has numerous health benefits. Having access to clean and safe drinking water in parks can make it easier for people to enjoy outdoor activities such as hiking, biking, and jogging. Spending time outdoors can help improve physical fitness, reduce stress, and boost mood, as well as lead to a greater appreciation of nature and promote

environmental awareness. Predicting fountain water quality can help mitigate climate change by reducing greenhouse gas emissions. The production and transportation of bottled water have a significant carbon footprint, and reducing the demand for bottled water can help reduce emissions. By encouraging outdoor activities and reducing the need for bottled water, predicting fountain water quality can contribute to a healthier and more sustainable future.

Historically, the primary approach to monitor water quality has been sampling and testing the water at treatment plants. (Benoit-Bird and Gagnon ,2015) While this approach can be effective, it has several limitations. First, it may not provide timely information about changes in water quality, since testing can be infrequent or delayed. Second, it requires significant human labor to collect and analyze samples, which can be costly and time-consuming. Machine learning methods have the potential to address these limitations by providing early detection of changes in water quality, improving accuracy, and reducing costs. Machine learning models can be trained to analyze data in real-time. We are also detecting trends in the water quality with feature importance analysis, providing early detection of changes in water quality and allowing for timely intervention. Additionally, machine learning can improve the accuracy of water quality predictions by taking into account a wider range of variables and more complex relationships than traditional methods. Finally, machine learning can be cost-effective by automating data collection and analysis, reducing the need for human labor.

Data Collection:

The dataset for this project includes four main files. The DrinkingFountains.csv file contains the locations of all drinking fountains at New York City parks from 2015 to 2022 (NYC

Parks, n.d.). The longitude and latitude of each fountain are provided in the format of a POINT (-73.98659181365889 40.60753207315604). The WaterQualityMonitoringData.csv file includes information on the water quality at various sampling sites in NYC (New York City Department of Environmental Protection, 2022). The water quality is assessed based on Residual Free Chlorine (mg/L), Turbidity (NTU), and Fluoride (mg/L) measurements. Although this file includes the sample site numbers and dates, it does not provide the specific locations of the sampling sites. The WaterQualitySamplingSites.xlsx file provides the locations of all water sampling sites in New York City, represented by X and Y coordinates under the State Plane Coordinate System (SPCS) (New York City Department of Environmental Protection, 2022). Lastly, the Df_prcp.csv and df_tavg.csv files include daily average temperature and precipitation data for New York City, collected from the NOAA website using an API (National Oceanic and Atmospheric Administration, 2022). This data was collected at LaGuardia Airport observing station and will be used to investigate any correlations between weather patterns and water quality (please refer to DataCollection.ipynb for detailed steps).

Data Pre-processing: (please refer to Data Pre-Processing.py)

As we collected data from different sources, an important step of the entire project was data pre-processing, which helped us generate our own dataset that would be suitable to feed to the machine learning models. We took the 5 following steps in the data pre-processing:

STEP 1: in WaterQualitySamplingSites.xlsx, we transformed the coordinates to longitude and latitude for convenience. This was done with the help of the NGS Coordinate Conversion and Transformation Tool (NCAT) (https://geodesy.noaa.gov/NCAT/);

STEP 2: We started with WaterQualityMonitoringData.csv and joined it with WaterQualitySamplingSites.xlsx so that we get the location of every water sampling site in longitude and latitude;

STEP 3: We looped over every pair of drinking fountains and sampling sites to calculate their distances. By doing this, we are able to find the nearest water sampling site for each drinking fountain. Thus, the sampling sites and drinking fountains were paired together;

STEP 4: We pivoted the data we have so that each row represents a day, and each column represents a water quality assessment at a specific sampling site (e.g. FLU at Alley Park, RFC at Forest Park);

STEP 5: We joined df_prcp.csv and df_tavg.csv to the data above by date.

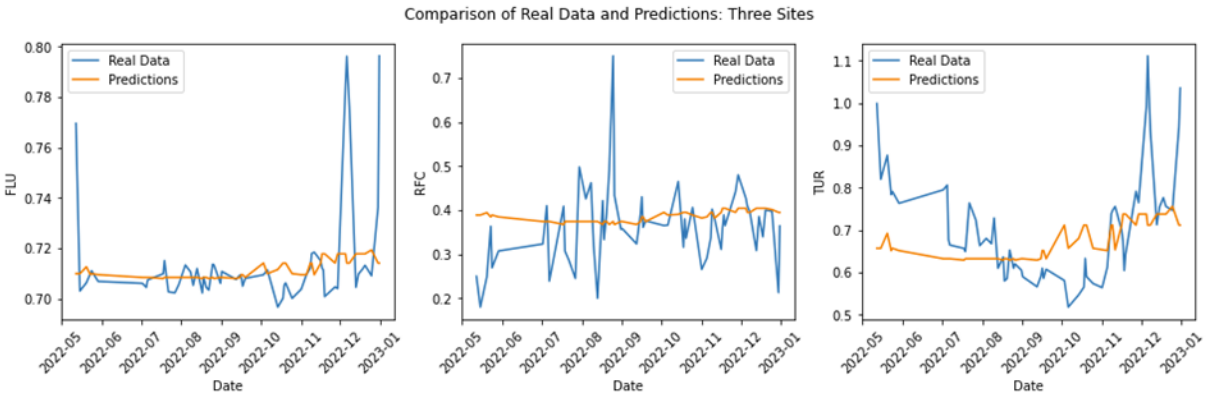
Based on the steps above, we generated our dataset, which is named as DrinkingDataFinal.csv.

Methods (please refer to Report.ipynb):

1. *RandomForest*

We first tried to use RandomForest as our model. We selected 3 columns, which are the FLU, RFC and TUR of one specific drinking fountain (Alley Pond Park) as data for our model. We did a train-test split with a test size of 0.3.

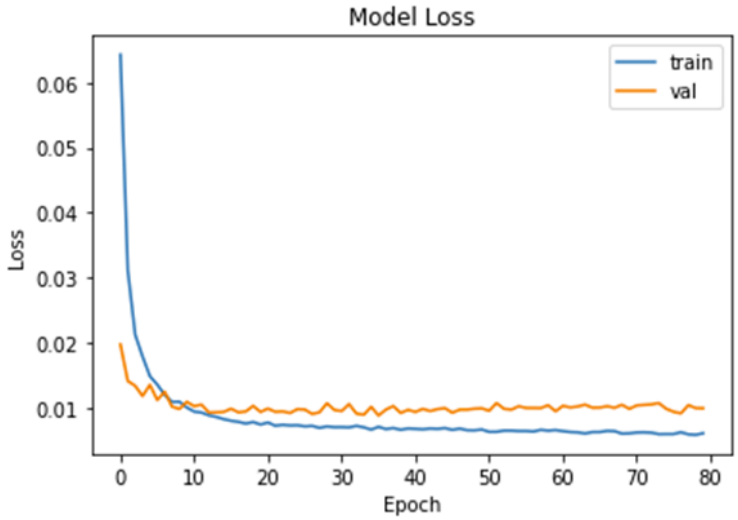
The testing R-squared of the model is 0.04, the testing MSE is 0.007, and the testing MAE is 0.055. This is not a very good result in terms of the testing R-squared value. We also plotted the predicted results and the true results of the last 60 days in the testing dataset to see the trend. As you can see from the visualization below, the prediction is poor, and it fails to catch the trend of real data over the period of time.



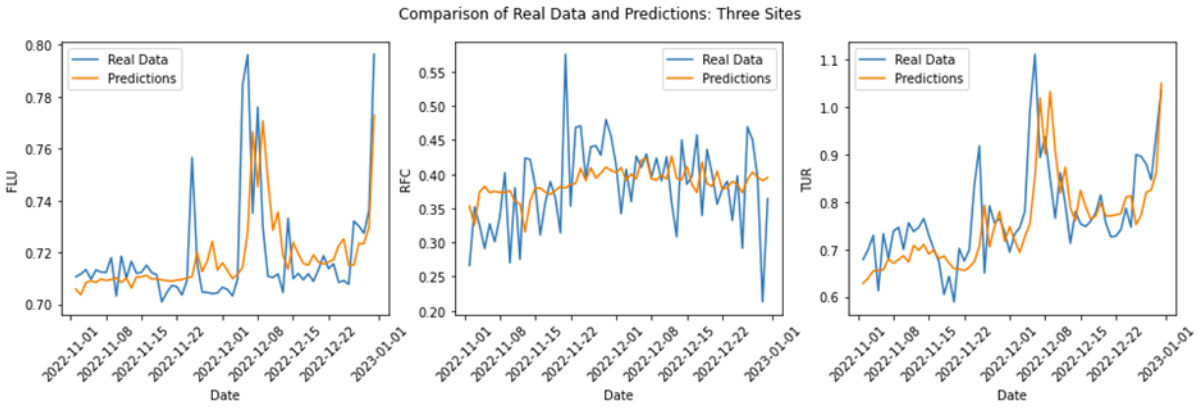
2. *DNN*

Our second trial is to use DNN. This time, we selected 6 columns, which are the FLU, RFC and TUR of two specific drinking fountains (Alley Park, and Alley Pond Park) as data for our model. Before feeding data to the model, we decided to add data of the previous 14 days as new columns. In other words, we are using the FLU, RFC, TUR, temperature and precipitation data from the previous 2 weeks as extra predictors for the water quality of a given day. We also normalized the data.

We built a 9-layer DNN with unit sizes of (512, 256, 128, 64, 32, 16, 8). We used ReLU as our activation function and MSE as the loss function. We also tried different epochs and plotted the trend of training and testing loss over the epochs. This helps us find the best epoch number without underfitting or overfitting. Based on our trials, the best epoch is 80.



The testing R-squared is 0.17, the testing MSE is 0.004, and the testing MAE is 0.041. The DNN model has a much better performance than the Random Forest model. We also plotted the predicted results and the true results of the last 60 days in the testing dataset to see the trend. As you can see, the prediction is much better, and it successfully catches the trend of real data.

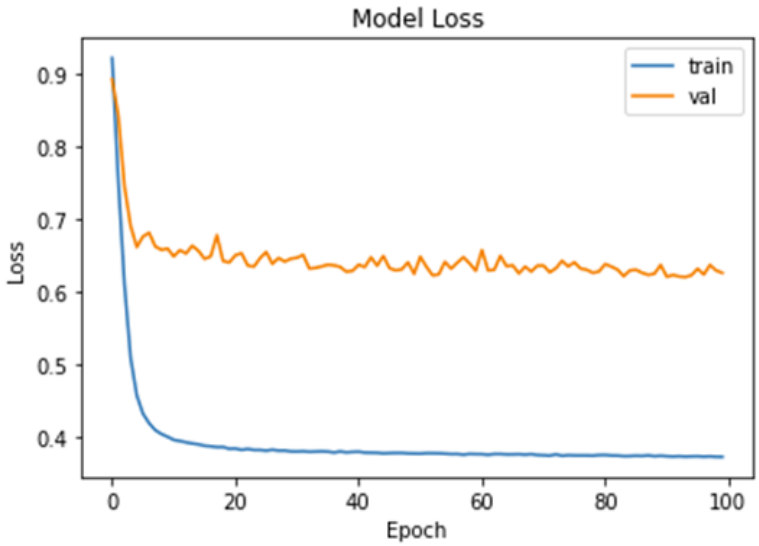


We also found the feature importance of the DNN model. The top ten features of importance are mostly water quality data from 1 week. This indicates that temperature and precipitation are not playing a large role in this model.

3. *RNN and LSTM*

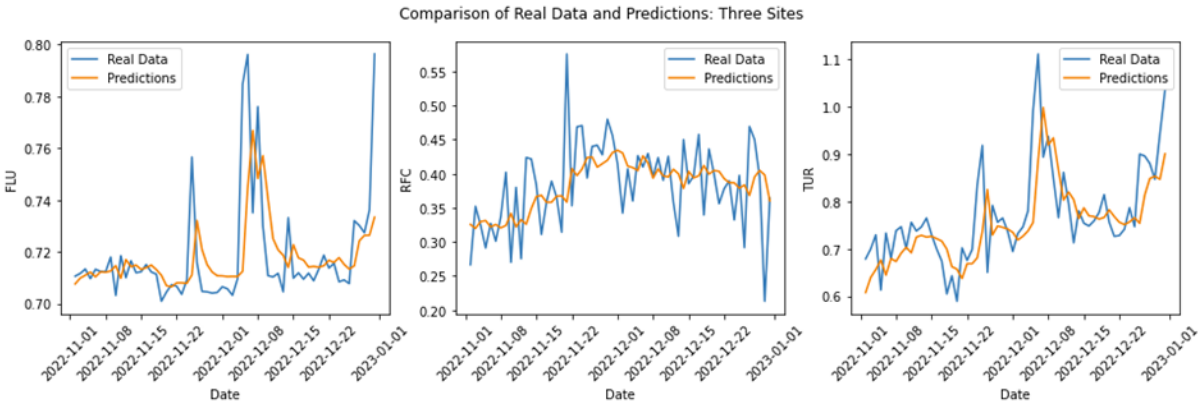
Because the data we have is based on time series, we think that RNN and LSTM may also be helpful here. We started with an LSTM model. Again, we selected two drinking fountains (Alley Park, and Alley Pond Park) as data for our model. We normalized the data, did a train-test split, and then generated sequential data with sequence length (SEQLEN) of 14 days and prediction length (pre_len) of 60 days.

In the LSTM model, we put 256 units and use Sigmoid as the activation function. The loss function is MSE. The learning rate is set to 0.0001 because faster rates would make it hard for the model to converge. We also tried different epochs and plotted the trend of training and testing loss over the epochs. This helps us find the best epoch number without underfitting or overfitting. Based on our trials, the best epoch is 100.



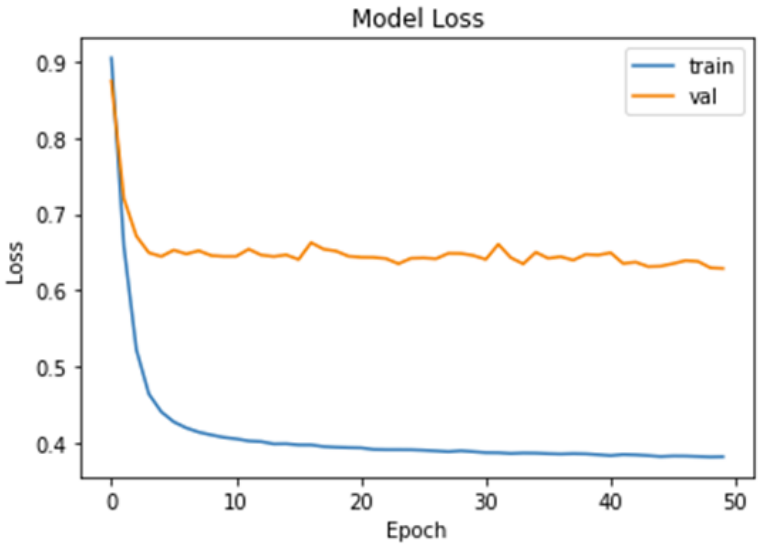
The testing R-squared is 0.25, the testing MSE is 0.003 and the testing MAE is 0.037. According to the metrics evaluation, LSTM is doing a better job than DNN. This could also be

visually reflected when we plotted the predicted results and the true results of the last 60 days in the testing dataset.

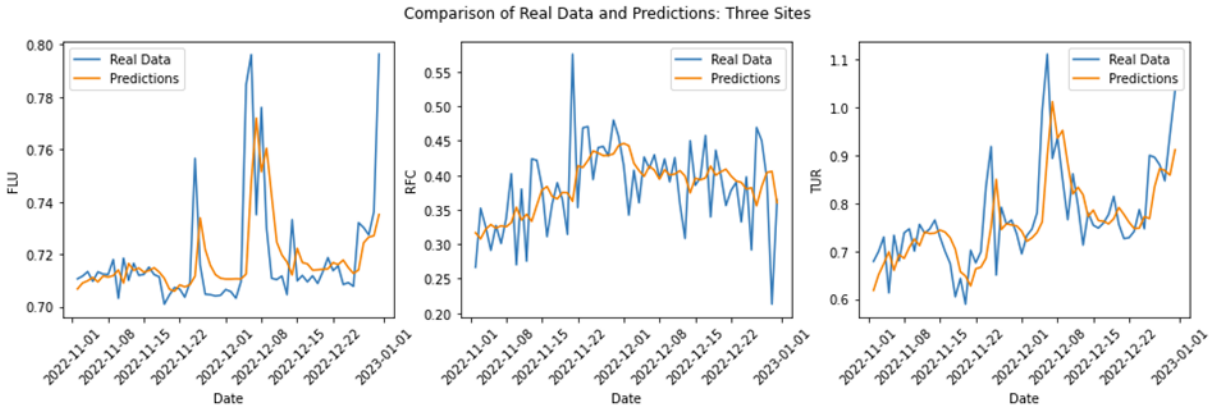


We also checked the feature importance of the LSTM model. Interestingly, the average temperature and precipitation of the given day are listed as two of the top three most important features. The FLU, RFC, TUR, and precipitation of the previous day are also listed in the top 10 important features. This indicates that average temperature and precipitation are also playing important roles in water quality.

With the same sequential data, we also built an RNN model with 128 units and Sigmoid as activation function. Again, we tried over different hyperparameters and decided that the learning rate is 0.0001 and the epoch is 50.



The RNN model has a testing R-squared of 0.25, a testing MSE of 0.003 and testing MAE of 0.037, which is very similar to the evaluation result of LSTM. Again, we plotted the trend of the predicted data and the real data.



Conclusion:

In conclusion, since LSTM and RNN are designed to process sequences of data and are particularly suited for time series data, they outperform other types of models in our case and provide us with good predictions. Based on the feature importance of the LSTM model, we also

find out that the average temperature and precipitation of the given day as well as the previous day are important predictors of the water quality. Improving water quality prediction is a crucial but challenging task that must be addressed to ensure public health, promote outdoor activities, and reduce greenhouse gas emissions. It is essential to emphasize the importance of data transparency and making government data more accessible to the public to promote research on public health-related matters. By doing so, we can achieve a better understanding of the factors that affect water quality and develop more effective models for predicting it. Overall, further research and development of adequate models are needed to improve water quality prediction and ensure safe and environmentally friendly outdoor experiences for everyone.

Citation:

Benoit-Bird, K. J., & Gagnon, A. G. (2015). A review of water quality monitoring techniques. Journal of the American Water Resources Association, 51(2), 326-339. doi: 10.1111/jawr.12217

DrinkingFountains.csv: New York City Department of Parks & Recreation. (2015-2022). Drinking Fountains - Geographic Locations. Retrieved from https://data.cityofnewyork.us/Recreation/Drinking-Fountains-Geographic-Locations/rkxx-2v2y

WaterQualityMonitoringData.csv: New York City Department of Environmental Protection. (n.d.). Water Quality Data Portal. Retrieved from https://data.cityofnewyork.us/Environment/Water-Quality-Data-Portal/7yqb-9df6

WaterQualitySamplingSites.xlsx: New York City Department of Environmental Protection. (n.d.). Water Quality Data Portal. Retrieved from https://data.cityofnewyork.us/Environment/Water-Quality-Data-Portal/7yqb-9df6

df_prcp.csv and df_tavg.csv: National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI). (n.d.). Climate Data Online (CDO) - Search and Order. Retrieved from https://www.ncdc.noaa.gov/cdo-web/webservices/v2