Water Quality Prediction

HarvardX Data Science Professional Certificate

Remi Courtellemont

2022-09-29

Contents

1 Introduction

"Accurate water quality prediction is the basis of water environment management and is of great significance for water environment protection. Water quality information exist in the form of multivariate time-series datasets. There is no doubt that the accuracy of water quality prediction will be improved if the multivariate correlation and time sequence data of water quality are fully used." [1]

This project is a requirement for the HarvardX Data Science Professional Certificate Program which requires to use a publicly available dataset to solve the problem of our choice. Water Prediction Data available in UCI Machine Learning is chosen. The report is created using R Markdown in RStudio that covers Data Preparation for Model Building, Data Exploration with common visualization techniques, Model Development using train and test sets, Model Evaluation using validation set and Concluding Remarks.

1.1 Objective

The goal is to predict the spatio-temporal water quality in terms of the power of hydrogen (pH) value for the next day based on the historical data of water measurement indices. Root Mean Squared Error (RMSE) is used to evaluate the accuracy of the model by comparing the predicted values with the actual outcome on the testing set. The best model having the lowest RMSE value is used to predict the water quality.

1.2 Water Prediction Data

UCI Machine Learning Repository, is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms. UCI Machine Learning Repository provides the data sets of daily samples for 37 sites, providing measurements related to pH values in Georgia, USA. The input features consist of 11 common indices including volume of dissolved oxygen, temperature, and specific conductance. The output to predict is the measurement of **'pH, water, unfiltered, field, standard units (Median)'**

2 Data Preparation

In the data preparation step, the Water Prediction data is downloaded and prepared for exploration, modeling and model evaluation using needed packages and libraries.

2.1 Install Required Packages

To prepare and transform the data, required packages are installed and necessary libraries are loaded.

2.2 Download Water Prediction Data

First, the Water Prediction data set is downloaded from UCI Machine Learning Repository.

The Water Prediction data is a list. The names of the different variables are: X.tr, X.te, Y.tr, Y.te, location.group, features, location.ids.

The readme.docx available in the zip data can be useful to really get the meanings of each list.

2.3 Creation of training and testing sets

2.3.1 Extraction

The training (.tr) and testing sets (.te) are already built. Features (X) and output (Y) are in separated lists.

The variable name of the training set in the code is **train_set**. The variable name of the testing set in the code is **test_set**.

2.3.2 Stations

The training set has 15,651 instances and the testing sets 10,434 instances. According to the readme document, there are 37 stations so we add the stations number in the training and testing sets.

2.3.3 Dates

For the training set, the readme document indicates that the dates start from 2016-01-28 with 423 contiguous dates. For the testing set, it indicates that the dates finish at 2018-01-01 with 282 contiguous dates. It means that the training period finish the 25th of march 2017 and the testing period starts from the next day.

2.3.4 input Y.J0

As a reminder, the goal is to predict the spatio-temporal water quality in terms of the power of hydrogen (pH) value for the next day **Y.J1** based on the historical data of water measurement indices.

For this study, it is considered that the value for the same day **Y.J0** is known, and used as an input for the prediction.

We extract and add Y.J0 (value of pH for the **same** day) at each instances of the training and testing sets.

2.3.5 Output Y.J1

For each instances, we also add the output to predict **Y.J1** (value of pH for the **next** day), which is the value we want to predict, to the training and testing sets.

Once **Y.J1** is added to the date just before, the rows with the latest date of training and testing sets are removed. It is necessary because there is no output in that case.

2.3.6 Predictors

The data includes 11 predictors.

Their names in the training and testing sets are X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11.

The real names of the predictors are listed in the variable "features" of the data:

X1 -> Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees Celsius (Maximum)

X2 -> pH, water, unfiltered, field, standard units (Maximum)

X3 -> pH, water, unfiltered, field, standard units (Minimum)

X4 -> Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees Celsius (Minimum)

X5 -> Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees Celsius (Mean)

X6 -> Dissolved oxygen, water, unfiltered, milligrams per liter (Maximum)

X7 -> Dissolved oxygen, water, unfiltered, milligrams per liter (Mean)

X8 -> Dissolved oxygen, water, unfiltered, milligrams per liter (Minimum)

X9 -> Temperature, water, degrees Celsius (Mean)

X10 -> Temperature, water, degrees Celsius (Minimum)

X11 -> Temperature, water, degrees Celsius (Maximum)

Instead of giving their full names to the predictors of the training and testing sets, we build 3 groups of predictors, which include maximum, minimum and mean values of a certain parameter:

Group P1: X1, X4, X5

Group P2: X2, X3

Group P3: X6, X7, X8

Group P4: X9, X10, X11

It has to be noted that group P2 correspond to the max and min of the median value we want to predict.

We can rename the predictors in the training and testing sets with the group names, as follow:

P1.max, P2.max, P2.min, P1.min, P1.mean, P3.max, P3.mean, P3.min, P4.mean, P4.min, P4.max.

2.3.7 Water Systems

The stations are divided into 3 water systems. In each system, the stations are connected.

It is interesting on a modeling point of view to add the system each station belongs (This information indicates spatial dependency among different locations which are important to the forecast).

G1 system includes the following station: 35

G2 system includes the following stations: 1, 2, 3, 31, 32, 33, 34, 37

G3 system includes the following stations: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 36

3 Data Exploration and Analysis

Next step is to analyze the content of the data, to see if we can arrange it differently to facilitate the understanding and modeling.

3.1 Data Overview

The training set is built with 15,614 rows and 16 columns.

An instance represents 11 predictors, the station number, the date of measurement, the output value of the day, output value of the next day (to be predicted), and the water system it belongs to.

Table 1: Train Set Overview, columns 1 to 8

P1.max	P2.max	P2.min	P1.min	P1.mean	P3.max	P3.mean	P3.min
0.0011306	0.8846154	0.0011198	0.0011133	0.6776316	0.8414634	0.7651515	0.7874016
0.0011696	0.8717949	0.0011591	0.0011523	0.7039474	0.8292683	0.7727273	0.7952756
0.0013255	0.8846154	0.0011984	0.0012500	0.6776316	0.8536585	0.7500000	0.7559055
0.0140936	0.8589744	0.0012377	0.0039258	0.6973684	0.8292683	0.7727273	0.7716535
0.0881092	0.8589744	0.0107662	0.0292969	0.6842105	0.8536585	0.7651515	0.7559055

Table 2: Train Set Overview, columns 9 to 16

P4.mean	P4.min	P4.max	stations	date	Y.J0	Y.J1	System
0.293750	0.2980769	0.2761628	1	2016-01-28	0.6481481	0.6481481	G2
0.293750	0.3012821	0.2761628	2	2016-01-28	0.6481481	0.6481481	G2
0.300000	0.2980769	0.2877907	3	2016-01-28	0.6481481	0.6481481	G2
0.296875	0.2948718	0.2790698	4	2016-01-28	0.6388889	0.6388889	G3
0.296875	0.2916667	0.2819767	5	2016-01-28	0.6481481	0.6574074	G3

3.2 Vizualizing output

The input **Y.J0** and output **Y.J1** are representing the median value of pH respectively the day and the next day of measurement (date).

the minimum value of Y.J1 in the training set is 0.5740741

the maximum value of Y.J1 in the training set is 0.962963

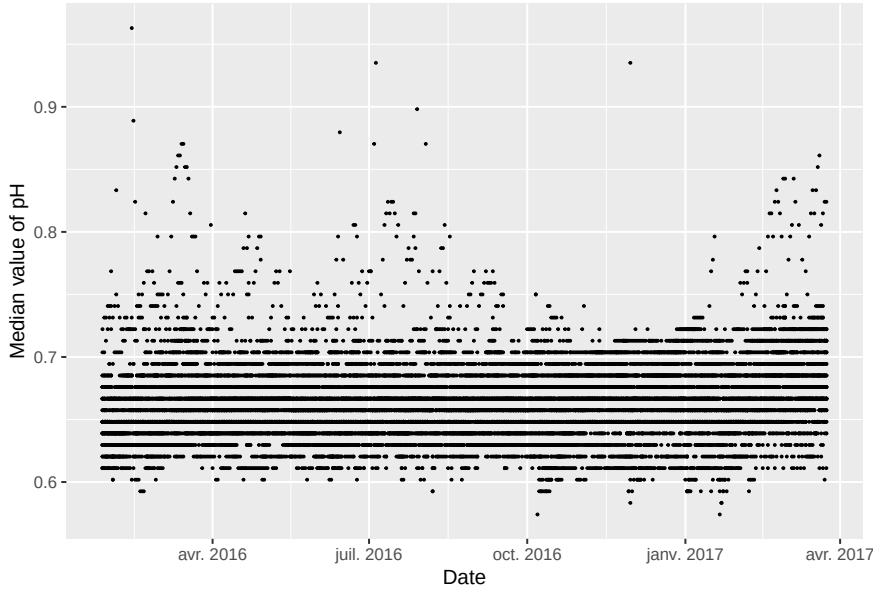


Figure 1: Y.J1 versus time

On the plot above, the output has a fixed step between each value.

In the training set, we have 38 levels.

The distance between each level is a constant step, which value is 0.0092593

This real values of the outputs are very small so the RMSE calculated will be hard to interpret. So Y.J0 and Y.J1 are divided by this distance in order for the levels and RMSE to be more understandable: if RMSE is higher than 1, it means that there is a global error between prediction and output of 1 level.

The new outputs have the following characteristics:

- the minimum value of Y.J1 in the training set is 62
- the maximum value of Y.J1 in the training set is 104

In the density curve below, we can see that, if we consider all the values of Y.J1,the distribution of values has 2 modes. The shape is not a Gaussian.

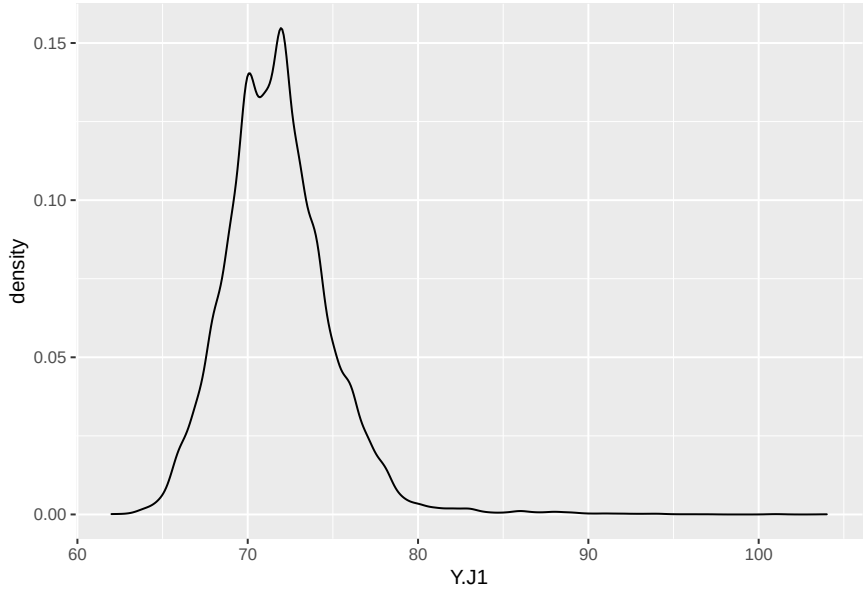


Figure 2: Output Y.J1 density curve of training sets

3.3 Vizualizing predictors

In the training set, we have the following mean and standard deviation for each predictor:

Table 3: Mean and SD of predictors of training set

	Mean	Standard.Deviation
P1.max	0.0651527	0.1613308
P2.max	0.8893670	0.0348073
P2.min	0.0289607	0.1208246
P1.min	0.0431850	0.1332288
P1.mean	0.5676978	0.1205787
P3.max	0.8585580	0.0310493
P3.mean	0.6047913	0.1469089
P3.min	0.5782328	0.1720104
P4.mean	0.5571312	0.2042844
P4.min	0.5368688	0.2125321
P4.max	0.5532272	0.1903459

If we generate a boxplot, we obtain the following drawing:

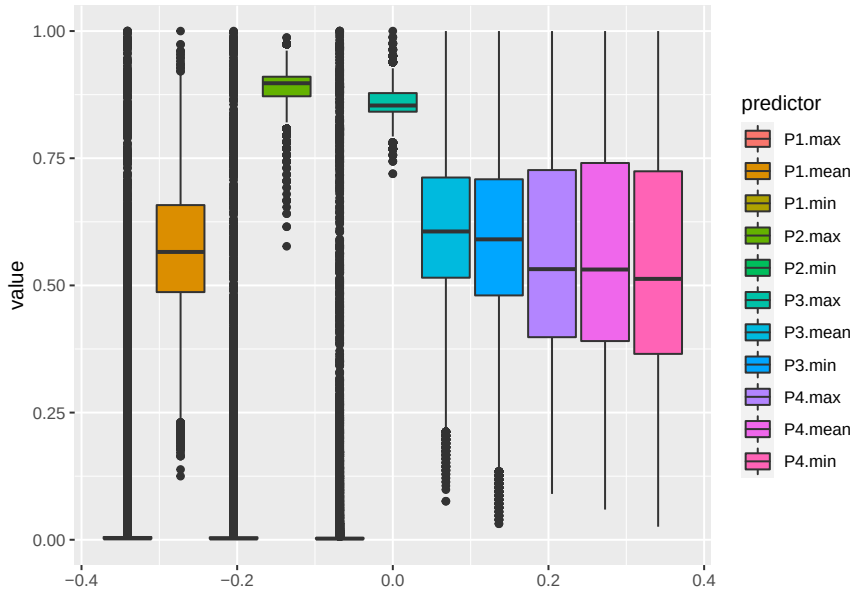


Figure 3: Boxplot of the predictors of the training set

4 groups of predictors can be identified, with similar variations:

P1.max + P1.min + P2.min are almost always at 0 but with outliers. They do not seem useful for our prediction.

P1.mean + P3.mean + P3.min have a median value between 0.56 and 0.60 and a medium SD between 0.12 and 0.17.

P4.max + P4.mean + P4.min have a median value between 0.53 and 0.55 and a large SD around 0.20

P2.max + P3.max have a high median value between 0.85 and 0.88 and a small SD around 0.03.

3.4 Relation between predictors, input and output

The data frame of the predictors and outputs of the training set is scaled.

Then the distance matrix is calculated and the heatmap of this matrix is plotted.

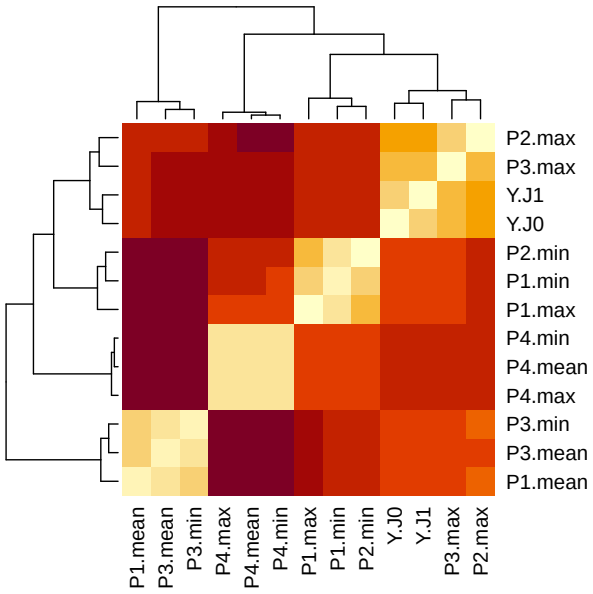


Figure 4: Heatmap of distance matrix of the training set

The heatmap of distance matrix show a connection between P2.max, P3.max, Y.J0 and Y.J1.

```
## $`1`
## [1] "P1.max" "P2.min" "P1.min"
##
## $`2`
## [1] "P2.max" "P3.max" "Y.J0"   "Y.J1"
##
## $`3`
## [1] "P1.mean" "P3.mean" "P3.min"
##
## $`4`
## [1] "P4.mean" "P4.min"  "P4.max"
```

In the clustering above,P2.max, P3.max, Y.J0 and Y.J1 are together in a cluster.

This cluster is confirmed by the correlation matrix below.

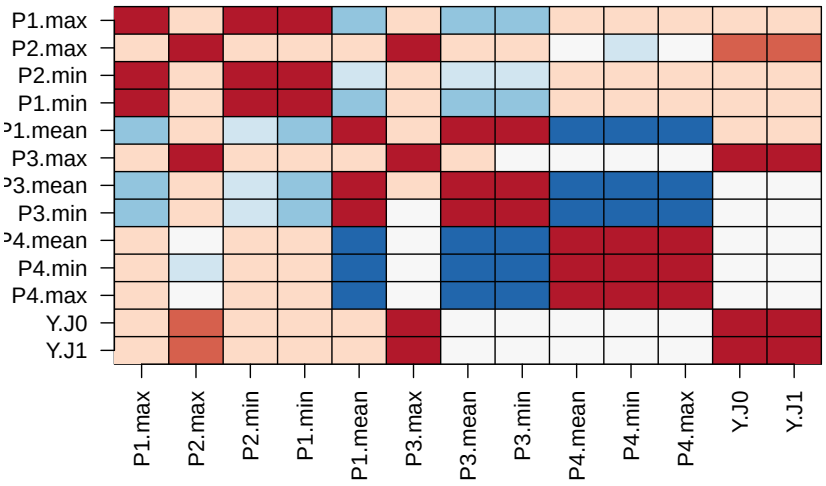


Figure 5: Correlation image of the training set

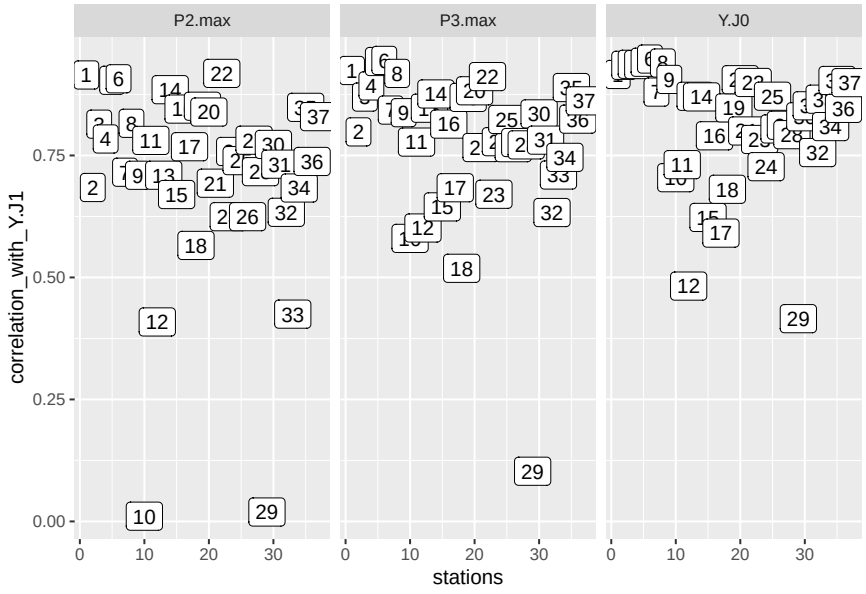
On the plot above, the output Y.J1 is significantly correlated to predictors Y.J0, P2.max and P3.max.

In details, here is the correlation of Y.J1 with all predictors:

Table 4: Correlation of Y.J1 with predictors and Y.J0 on training set

	correlation_with_Y.J1
P1.max	0.2940548
P2.max	0.7232727
P2.min	0.2672407
P1.min	0.2959526
P1.mean	0.1608006
P3.max	0.8485883
P3.mean	0.0443944
P3.min	-0.0207616
P4.mean	0.0103630
P4.min	-0.0024601
P4.max	0.0255076
Y.J0	0.9035911
Y.J1	1.0000000

The correlation between each station and their predictors P2.max, P3.max and input Y.J0 can be calculated.



It appears that the correlation with P2.max, P3.max and Y.J0 is not high for all stations. There are some stations with low correlation, for example station 29 is not correlated at all with P2.max and P3.max, and not much with Y.J0. The variance of correlation within stations might have an impact on the accuracy of group models for some stations.

The principal component analysis of the distance scaled matrix gives the following repartition of importance of components:

```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4407 2.0110 1.3799 0.61805 0.60134 0.33132 0.30121
## Proportion of Variance 0.4582 0.3111 0.1465 0.02938 0.02782 0.00844 0.00698
## Cumulative Proportion  0.4582 0.7693 0.9158 0.94520 0.97302 0.98146 0.98844
##                            PC8     PC9    PC10    PC11    PC12    PC13
## Standard deviation     0.27562 0.23198 0.10521 0.08011 0.05091 0.02044
## Proportion of Variance 0.00584 0.00414 0.00085 0.00049 0.00020 0.00003
## Cumulative Proportion  0.99428 0.99842 0.99927 0.99977 0.99997 1.00000
```

PC1 and PC2 explain more than 75% of the cumulative proportion of variance.

Adding PC3 make it up to more than 90%.

When plotting PC1 vs PC2, it appears that the output is high when PC2 is high and low when PC2 is low. However, it is hard to detect the value of output just with PC1 and PC2.

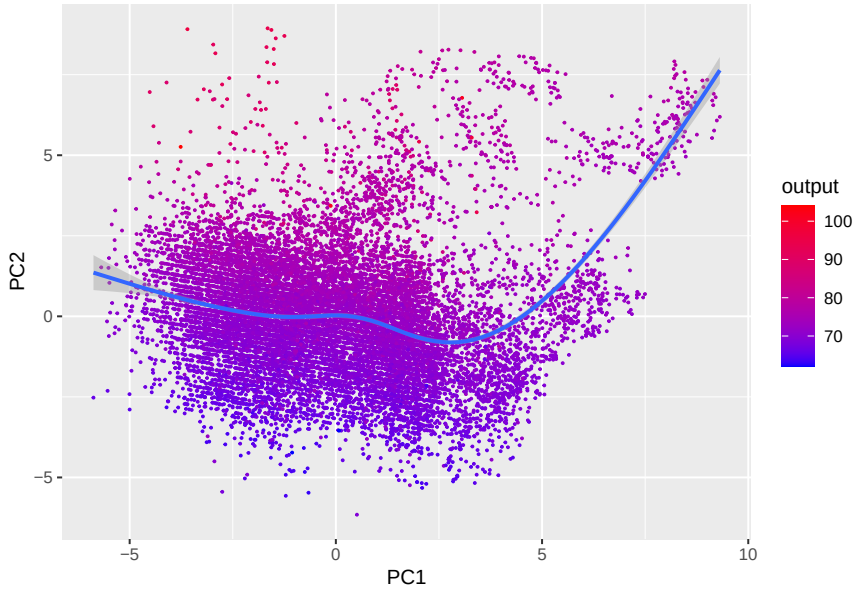


Figure 6: The 2 first principal components of the training set

Last observation: if the difference between Y.J1 and Y.J0 is plotted for all instances of the training set, the output is the same as the input in the majority of the cases.

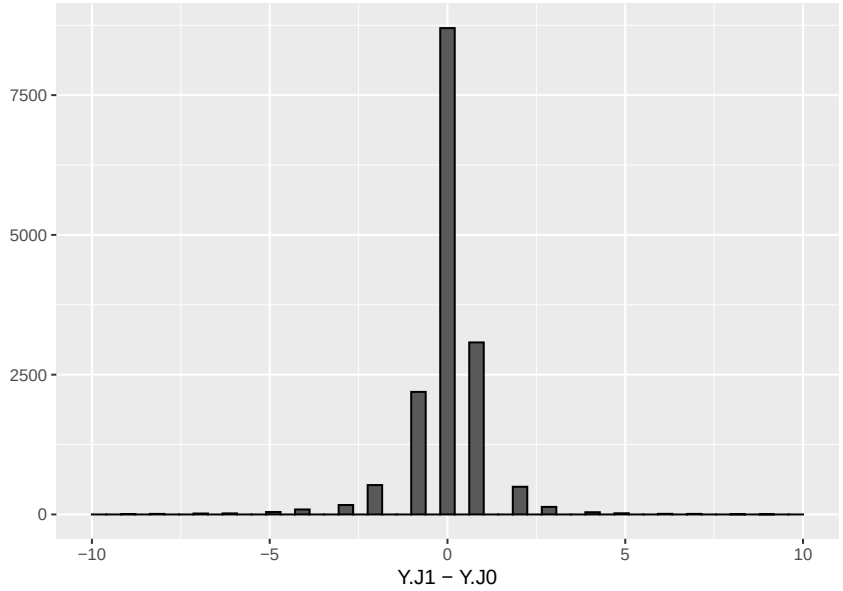


Figure 7: Difference Y.J1 - Y.J0 on training set

4 Model Development and final holdout test

The water prediction model is developed using the training set (train_set) and the final holdout test is performed on the reserved testing set (test_set).

It is decided to create models for each group of water systems (G1, G2 and G3). It is considered that in each group of water systems, the interconnection of stations creates similarities in the results. This way, only 3 models are needed for each of the 3 groups of water systems. This is the best way to quickly obtain results. However, it is expected that the models built from water systems might not have the same efficiency on all stations of the group (remember the correlation variance with P2.max, P3.max and Y.J0 at stations level).

4.1 Splitting training set into train and validation sets

The training set is split into a train set and a validation set with 80% and 20% of the original training set (train_set) respectively. In case of validation, the 80/20 split is a usual value to have sufficient material for training (80%) and a minimum amount of material to test (20%). Moreover, the test set is already built separately, so we do not need more split of the training set that could justify to reduce the previous validation split of 20%. We select the stations feature to make the split, so that there are the same proportion of stations in the 2 sets. The train set is called **train** in the code, and the validation set is called the **validation** in the code. The model is developed using **train** and tested with **validation** before the final hold-out test on the testing set (test_set).

We divide the data (train and validation) into sets of water systems.

4.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error shows how far the model prediction falls from the actual outcome. In other words, it is the standard deviation of the prediction errors which indicates how spread the data is. The error term is the difference between the predicted value and the actual outcome (Glen 2020).

RMSE for a model is calculated using the formula below:

$$RMSE = \sqrt{\frac{1}{N}\sum(\hat{y} - y)^2}$$

4.3 Model Development and RMSE Calculation

We apply regression models to the 3 different water systems G1, G2 and G3.

4.3.1 Selected models

The influence of interconnected water systems and environment on the output can be complex. The training set has been recorded for 1 year and 3 months, whereas the testing set last for 9 months. The water systems situation can be very different between years, and external influence might occur. The output could be outside of the range of the training set. We need models with good extrapolation. Then, it is decided to choose the following models:

- The model 1 is simply based on previous observations.
- The model 2 is a simple linear regression.
- The model 3 is a more sophisticated machine learning algorithm based on neural networks.
- The model 4 is the principal component regression, when unknown coefficients apply.

4.3.2 Model 1: Y.J1 = Y.J0

As noticed previously, in majority of the cases, the output remains the same the next day.

So the first model can be Y.J1 = Y.J0. This model measured with the validation set gives the following RMSE.

Table 5: RMSE for Model 1 - Validation Stage

Water_System	RMSE_Model_1
G1	0.8677218
G2	1.2631660
G3	1.3315815

4.3.3 Model 2: LM

The linear regression from Caret package is used for Model 2. In order to avoid over-training, we use the 3 predictors that are correlated to Y.J1, which are : P2.max, P3.max, Y.J0 and the stations' number (except group 1 which includes only 1 station).

Table 6: RMSE for Models 1/2 - Validation Stage

Water_System	RMSE_Model_1	RMSE_Model_2
G1	0.8677218	0.7571547
G2	1.2631660	1.2158416
G3	1.3315815	1.2490528

4.3.4 Model 3: BRNN

The Bayesian Regularized Neural Networks (BRNN) is used for model 3. The 3 predictors that are correlated to Y.J1 are used for the training : P2.max, P3.max, Y.J0 and stations. The best tune of parameter 'neurons' is searched between 1 and 2 with the validation set.

The best tune of 'neurons' parameter for G1 is 1 and the minimum of rmse is 0.7489084.

The best tune of 'neurons' parameter for G2 is 2 and the minimum of rmse is 1.1914407.

The best tune of 'neurons' parameter for G1 is 2 and the minimum of rmse is 1.2433349.

Table 7: RMSE for Models 1/2/3 - Validation Stage

Water_System	RMSE_Model_1	RMSE_Model_2	RMSE_Model_3
G1	0.8677218	0.7571547	0.7489084
G2	1.2631660	1.2158416	1.1914407
G3	1.3315815	1.2490528	1.2433349

4.3.5 Model 4: PCR

The Principal Components Regression from Caret package is used for model 4. The 3 predictors that are correlated to Y.J1 are used for the training : P2.max, P3.max, Y.J0 and stations. The best tune of parameter ncomp is searched with the validation set.

The best tune of ncomp parameter is 2 and the minimum of rmse for G1 group is 0.7424585.

The best tune of ncomp parameter is 1 and the minimum of rmse for G2 group is 1.2552691.

The best tune of ncomp parameter is 1 and the minimum of rmse for G3 group is 1.319119.

Table 8: RMSE for Models 1/2/3/4 - Training Set

Water_System	RMSE_Model_1	RMSE_Model_2	RMSE_Model_3	RMSE_Model_4
G1	0.8677218	0.7571547	0.7489084	0.7424585
G2	1.2631660	1.2158416	1.1914407	1.2552691
G3	1.3315815	1.2490528	1.2433349	1.3191190

4.3.6 Details for each station

Let's have a look deeper inside the RMSE results of each station for each model.

First, we need to build the models 3 and 4 with the best tune.

Then, we generate the predictions for all models.

The data is then combined into a single table.

The boxplot of this table is shown below:

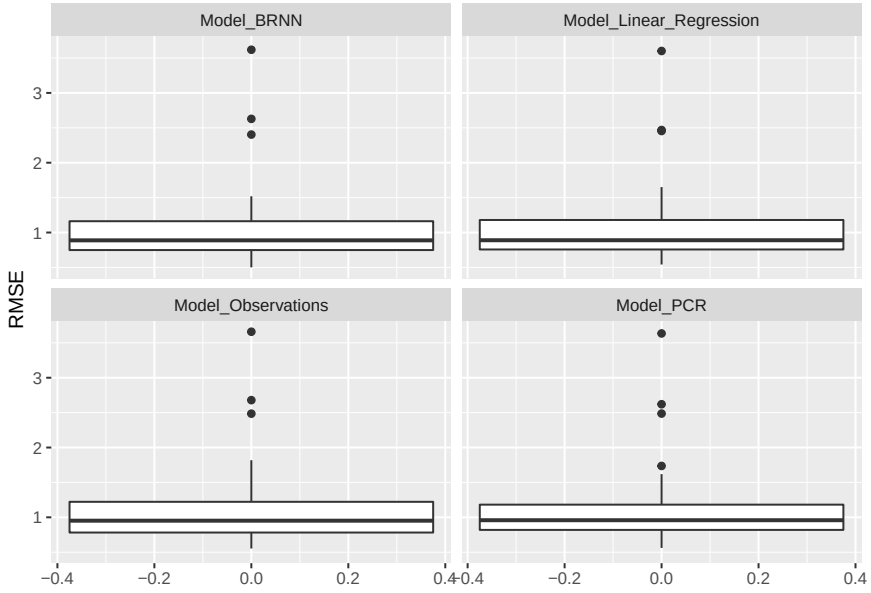


Figure 8: Boxplot of models for all stations on training set

We can see that we have outliers and the median is below 1 for each model.

In the training set, the best model for each station is the following:

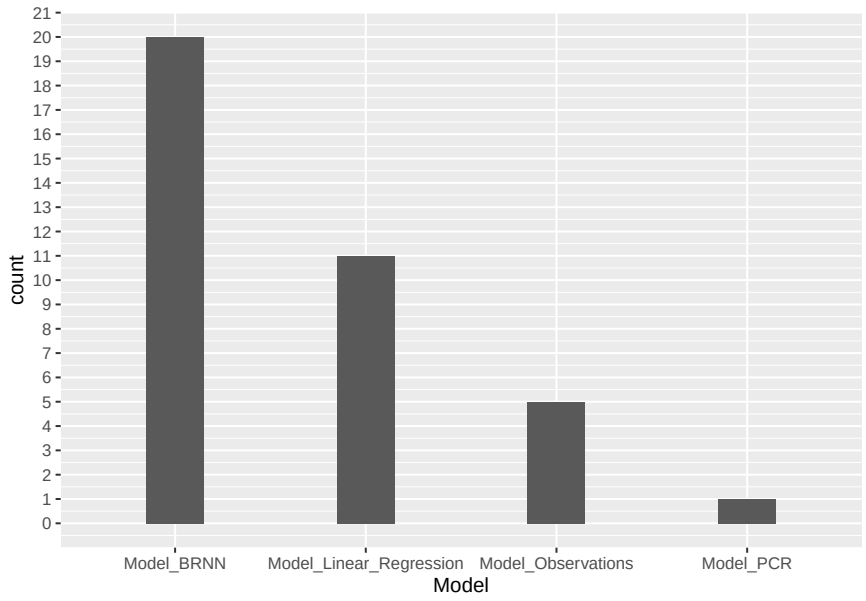


Figure 9: Total number of best models for all stations

We can see that in some stations, keeping the same output Y.J1 as the input Y.J0 is the best model! The model 3 (brnn) is the most efficient model on stations. Model 2 (lm) is the second most efficient model and sometimes can not be beaten. Model 4 (pcr) is not a very performing model.

4.4 Final holdout Test with test_set

4.4.1 Validation and choice of model

We have tested 4 models. The difference in RMSE between models is not so high.

It has to be noted that linear regression model is performing quite good, and it is hard to have a model performing better than linear regression for this kind of prediction.

From the training set, the best models for each water systems are :

- Model 4 (pcr) for G1
- Model 3 (brnn) for G2 and G3

We will use those models to perform the final holdout test of the 3 water systems.

4.4.2 Modifications of train and test sets

To apply those models to the testing set, the following modifications must be done:

- Standardization (dividing input and output by 1 step)
- Separating the data of train and test set into groups of water systems (G1, G2 and G3)

4.4.3 Final Holdout Test

We apply the models on each group of water systems of the test set, and we get the following table:

Table 9: Final Holdout Test

RMSE_G1_Test_Set	RMSE_G2_Test_Set	RMSE_G3_Test_Set
0.7759364	1.483657	1.405286

The RMSE of G1 is similar to the validation test. The RMSE obtained for G2 and G3 are acceptable, however a bit higher than expected in comparison to validation sequence: it could be the result of some stations that do not behave like the others (different correlations, external influence,...) and impact the RMSE of the group.

5 Conclusions

A water quality prediction model is created, with data from UCI Machine Learning Repository. The data is built into train and test set from the package received after downloading. A quick look at distance and correlation shows that the predictors P2.max and P3.max are key to predict the output, together with input Y.J0. The sets are divided into groups of water systems, because the interconnection between the stations in each group generates relation between outputs. It is also easier and faster to have a total of 3 models, one for each group of water systems, rather than having models for each of the 37 stations. 4 models are applied to train_test :

- Same prediction for tomorrow as today
- Linear Regression
- Bayesian Regularized Neural Networks
- Principal Components Regression

The models are then tested with a validation set (20% of the train set). Finally, the final holdout test is performed on the test_set.

The results are encouraging. The G1 group has a RMSE close to the one during validation phase. The G2 and G3 group have a RMSE slightly higher than the validation phase, but their groups are much bigger and some stations might be less responsive to the built models.

Limitations

As seen during validation phase, some stations in groups 2 and 3 of water systems do not behave like the other stations. So, their RMSE have a negative and significant impact on the RMSE of the group. 2 predictors and 1 input are used for all groups. Some stations might be less sensitive to these predictors compare to other stations. Moreover, the gain with advanced models (brnn, pcr) compare to linear regression is small.

Future Scope

One idea would be to extract some of the stations from the groups, for example the stations with little correlation or the outliers on the boxplot of RMSE, and to treat them separately. Another idea would be to build models for each and every station, but it would need much more time to build and to run. We could also try to "play" more with the parameters from the selected models (tuneControl, tuneGrid), or we could investigate other types of models in caret package. Finally, we did not use the time sequence, and its influence could be checked (as recommended by [1]).

6 References

1. Jian Zhou, Yuanyuan Wang, Fu Xiao, Yunyun Wang, Lijuan Sun. August 2018. "Water Quality Prediction Methode Based on IGRA and LSTM" MDPI Water Journal

2. Irizarry, Rafael A. 2020. Introduction to Data Science: Data Analysis and Prediction Algorithms with r. CRC Press