Quantitative model developer

Code test report

Applicant: Zixin Zhang

Email: zhangzixin.dalian@gmail.com

Date: 14 January 2021

1 Task 1 - Coding

Code and README file are presented in github. The link is `https://github.com/7cats/AirmineTest`.

Programming language is python.

2 Task 2 - Forecasting air quality

2.1 Model design

The flow chart of the model is presented in Figure 1.

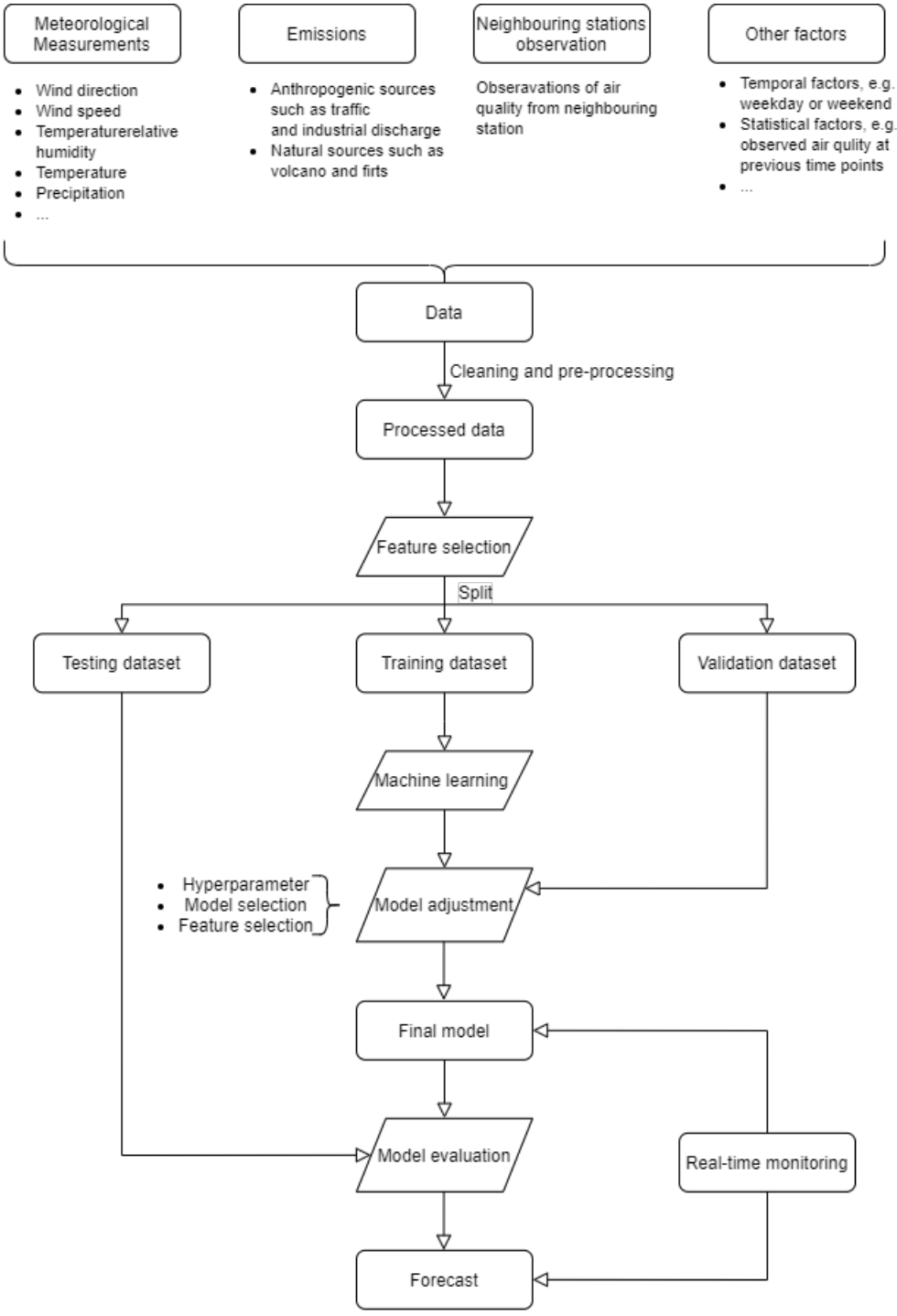

Figure 1: Model structure for forecasting air quality of output 1

This model is designed to produce the output1 forecast, in other words, a time series prediction of one location.

2.1.1 Model input

The modelling includes in the following inputs:

- Meteorological measurements
- Emission
- Neighbour station monitoring
- Other data input

2.1.2 Model structure

This model is a statistical model which estimates the parameters of variables for PM forecast. The model structure has these several parts:

1. Data collection, cleaning and split
   Influencing factors should be selected based on a carefully preliminary research. Collected data should be properly cleaning, processed and divided into three groups: training data, testing data and validation data. This process is vital for avoiding the "garbage in, garbage out".

2. Model construction, adjustment and selection
   The models are built by using the training dataset. Best model and parameter settings are selected by applying the validation dataset to all models. Finally the performance of the model is evaluated by test dataset.

3. Real-time monitoring
   The final model and the prediction should be updated and modified continuously based on a real-time monitoring system and a closely diagnosis.

2.2 Key challenges

There are several challenges for building a forecasting model for air quality:

2.2.1 The uncertainty of emission sources

There are two kinds of variations of the emission which needs to be taken into consideration when building the forecasting model. This first kind is that the emission has its own temporal pattern, which also varies in different places. For example, traffic-related emission varies across the day and behaves differently in urban area or rural area. The other example is the pollen concentration has a seasonal pattern. The other kind is the emission could suddenly increase. The natural hazard such as volcano eruption and forest fire might suddenly increase the PM concentration in the air.

2.2.2 Accuracy relying on the weather forecast accuracy

It is known to people that meteorological measurements have a significant influence on the air quality. Therefore, it is inevitable that the error from weather forecast will reduce the accuracy of the air quality forecast. The weather forecast should be updated timely and its error should be minimised by statistical methods when including it into the model input.

2.2.3 Necessity of real-time diagnosis and adjustment

The model should include a comprehensive observing system that is used to evaluate the forecast accuracy and precision. This is beneficial for identifying the weak points of the model.

2.3 Differences between models for two outputs

2.3.1 The boundary conditions

The boundary condition should be set when the forecast is for an area. The total particles of the study area at one time stamp should equal to amount of particles at last time stamps adding the in-flux of particles and emission from local sources, then

minus the out-flux of the particles. Clear boundary conditions should be defined to control the total PM inside the area.

2.3.2 Extra input variables

Forecast for an area requests more variables for ensuring the accuracy of the models. For example, include the NDVI value of that area because the density of the vegetation has an influence on the air quality.

2.3.3 Vertical distribution of the pollutants

The weather condition is different at one certain location in the vertical direction which further affects the distribution of the pollutants across a certain area.

2.3.4 Model structure and grided system

It would be more desired to use a three-dimensional deterministic model to forecast for an area. A grided system should be created and cell sizes should be carefully designed. There are two ways to model the movement of the particles: Lagrangian and Eulerian models. A hybrid model which is an ensemble of deterministic models and parametric models of is possible of improving the performance of the model's accuracy.

3 Task 3 - Forecasting pollen

The allergic pollen is the result of the flowering of plants. This determines that most of the differences of pollen forecast and air quality forecast, which are discussed as follows.

3.1 Seasonal pattern

The most distinct difference between forecasting pollen from air quality is that the content of pollen grains in the air has a seasonal pattern fluctuation, as it is the direct result from the season circle of vegetation. This poses a requirement on the models,

which is that the model should focus on medium-range forecast. Monitoring and early warning system of pollen is beneficial for improving forecast accuracy and reduce the risk of allergy for people.

The seasonal pattern gives researchers more freedom to exploit historical data since the growth circle of vegetation for one area is more stable comparing to other emission sources. However, if the species of vegetation or the climate have changed greatly, historical data should be incorporated with extra attention and adjustment.

The model should include factors that have an effect on the growth of plants, such soil moisture and temperature. Relevant phenological models should be informative.

3.2 Extent of study area

Pollen is possible of being transport for a long distance, therefore the study area should be large enough to include all sources of emission. In the same time, the terrain of study area also needs to be taken into consideration.

3.3 Mechanism

Modelling pollen should consider pollen grain size, which is relatively larger than other pollutants. And pollen modelling does not need consider chemical transformation.