Air Quality Data Analysis Report

Zhengyan Shi 2023/1

1. Introduction

This is an application for data analysis and visualization of air quality data and the functions are listed as follows:

1. **Load data**
   Users are allowed to load `*.csv` data to the application for further analysis.
2. **Display the spatial-temporal pattern of air pollution**
   - **Top 10 stations with most selected pollutant**: Users are allowed to set the geographical range, the air pollutant, and the starting and ending timestamp of the time period. A bar chart is shown on the right side of the page to demonstrate the top 10 stations with most selected pollutant.
   - **Concentration of a specific air pollutant**: Users are allowed to set the air pollutant, the station, and the starting and ending timestamp of the time period. A line chart is shown to demonstrate the trend and the average concentration is also calculated.
3. **Displays comparison information**
   - **Difference between two stations**: Users are allowed to set the stations, the air pollutant, and the starting and ending timestamp of the time period. Two line charts are shown to demonstrate the difference and the average concentration is also calculated.
   - **Difference between two air pollutants**: Users are allowed to set the air pollutants, the geographical range, and the starting and ending timestamp of the time period. The average concentration is calculated and shown both in the bar chart and the line editor.
4. **Prediction-based analysis**
   - **Concentration of air pollutants at a specific station**: Users are allowed to set the station. The predicted concentration of the next month is shown.
   - **Concentration of air pollutants averaged over stations**: Users are allowed to set the geographical range. The predicted concentration of the next month is shown.
5. **Similarity-based analysis**
   - **Similarity between two stations**: Users are allowed to set the stations, and the similarity of all pollutants is shown.
   - **Similarity between two air pollutants**: Users are allowed to set the pollutants, and the similarity in half a year is shown.

2. Implement details

2.1. Multi thread

The `*.csv` file is large and it will make the computer stuck when loading it. In order to avoid the situation, I use a new thread to load the data. In that way, users can freely drag the window, change tabs or set restrictions. Also, a progress bar is used to show the rate of loading process. If the process is done, there will be a message box to inform you.

Figure 1. Progress bar

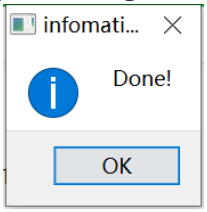

Figure 2. Information message box

2.2. Data structure

- **Vector**: There are mainly two vectors used in my program. First of them is used to store the "ID-name" pair of stations, and the second one is used to store "ID-information" pair, which connect stations with concentration of pollutants.
  In fact, the information in both the two vectors can also be built as `std::unordered_map`, but it takes plenty of time to build the two hashmaps, thus making the window stuck. So, to reach a balance between loading time and calculating time to satisfy users, I choose to build it as `vector`.
- **KDTree**: KDTree is built to store the "ID-location" pair information. Traverse through all the data may cost O(n) time, but KDTree can provide an O(sqrt(n)) cost search.

2.3. Robustness

To make sure that users' input is legal and improve the robustness, I add some checks, including:

- Check whether the set timestamp is legal.
- Check whether the set geographical range is legal.
- Check if users choose the correct number of pollutants.
- Check whether there is any data in the range that users have set.

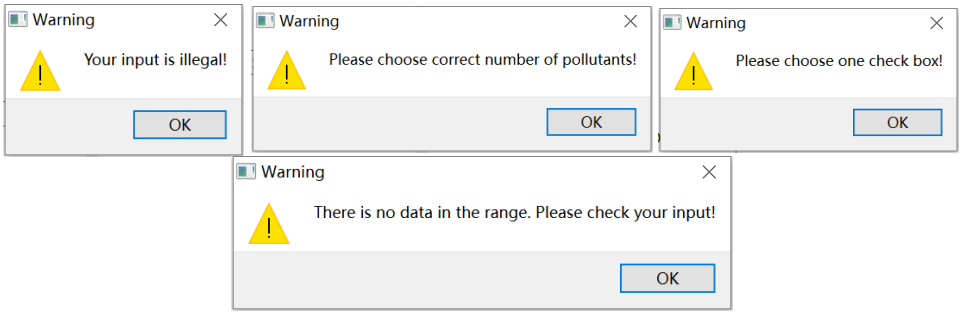

Figure 3. Warning message box

2.4. Demonstration

Various charts and other forms are applied in the interface to give users a clear understanding of the information, including:

- **Bar chart**: The top 10 stations with most pollutants are listed in a bar chart. Moreover, the difference between two pollutants is demonstrated by a bar chart.
- **Line chart**: The trend of concentration of pollutants are shown by a line chart. Also, the difference between two stations is shown by line charts.
  In fact, in Qt, there exist a class called `QSplineSeries`, which can draw smooth curves. But I still choose to show the line chart by fold lines, because there are lots of data points in a

narrow range of timestamp. If I use smooth lines to connect these points, it will be hard to distinguish between different lines.

- **Text editor**: The results of prediction-based analysis and similarity-based analysis are directly demonstrated in text editors.

2.5. Interpolation

In the dataset `air_quality.csv`, there are many data points which contain `NULL`. To make the curves more accurate, I apply linear interpolation to these kinds of data.

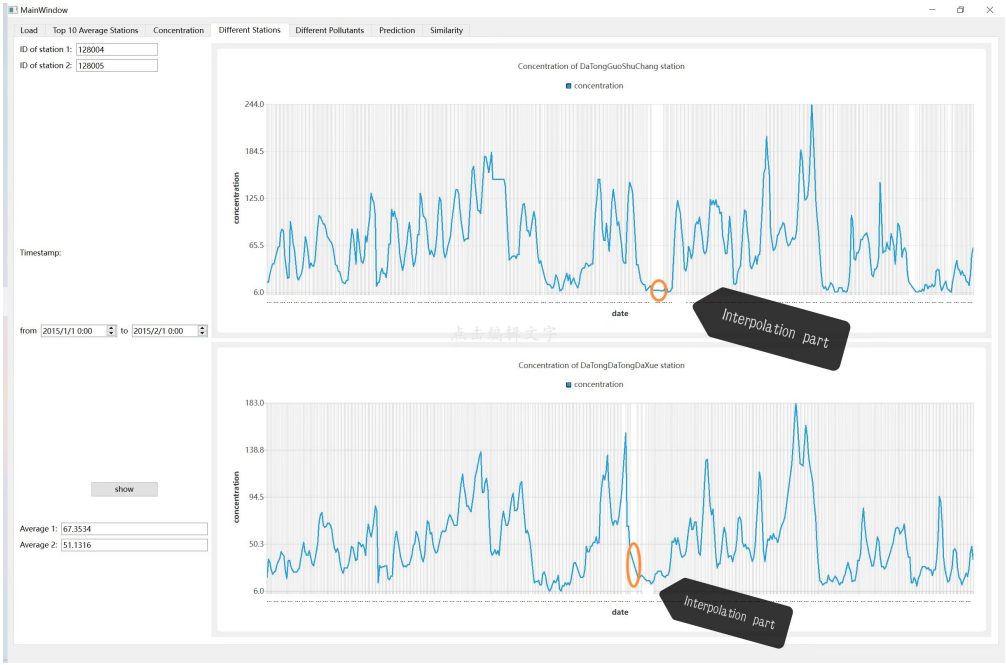

Figure 4. After interpolation

3. Results

Screen shots of the results are listed as follows. Due to the limitation of the length of the report, I will only choose some of them to explain. For more information about how to use the APIs, please refer to `README.md`.

Figure 5. Results

The concentration of pollutants is more likely to be similar with recent data. The predicted concentration is calculated with the data of recent two months. If the timestamp of data is closer to the date of target , the data will get higher weight. And an example of results are shown in Figure 6.

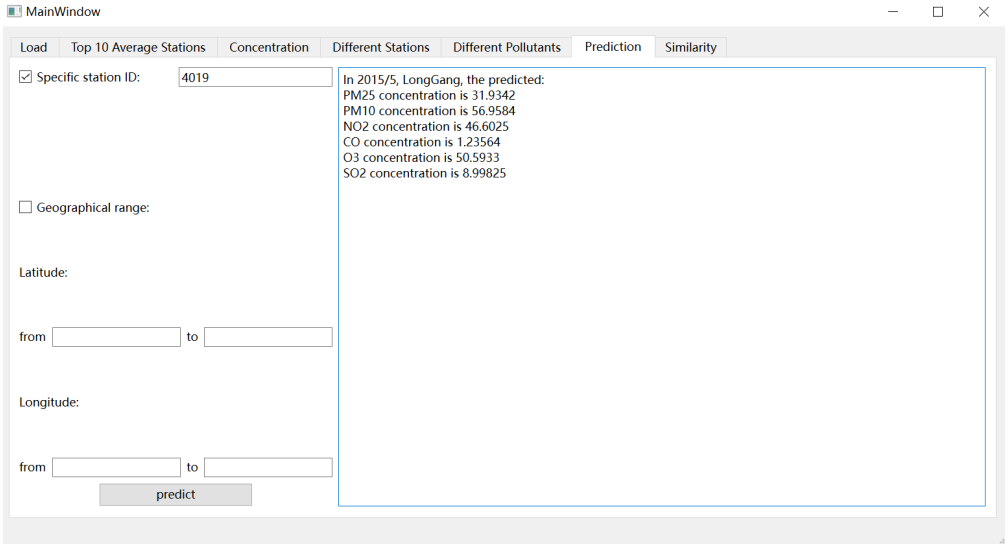

Figure 6. Prediction

The similarity is calculated with the average of relative gap between the data of the same timestamp. And an example of results are shown in Figure 7.

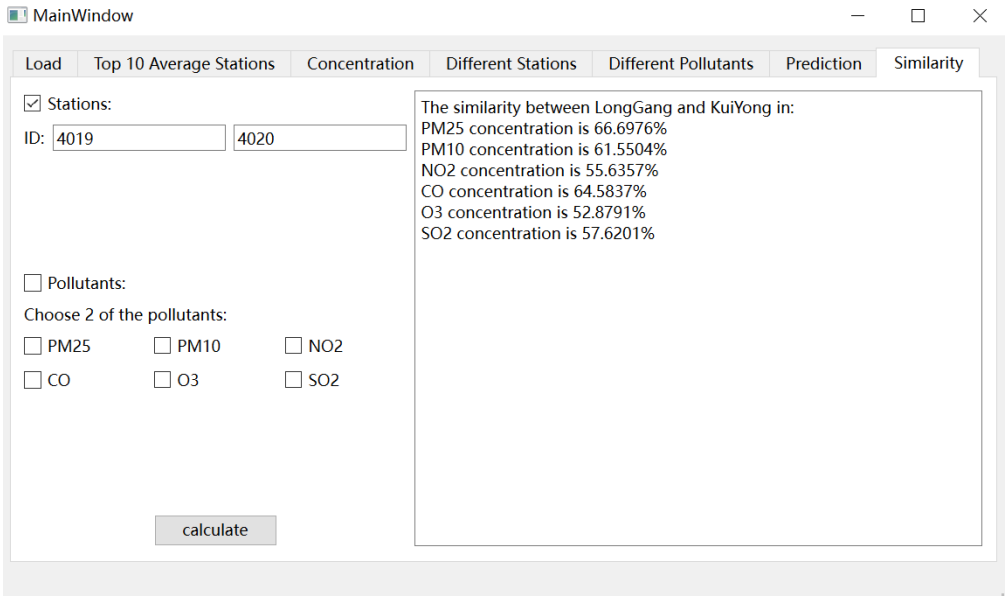

Figure 7. Similarity

4. Discussion

4.1. Pollutant with most concentration

In our daily life, we often regard PM25 as the pollutant with most concentration. But in this dataset which contains large range of data, PM10 seems to take the lead when it comes to concentration, and exceeds PM25 by nearly 40%, as is shown in Figure 8.
Moreover, I find that the concentration of PM25 and O3 are usually at the same level, regardless of the geographic location or timestamp. NO2 and SO2 have a similar relationship, too. This can also be shown by the "Similarity" function.
CO seems to be at low level concentration at all times and locations. Its concentration may be just 1% or less of other pollutants at the same condition.

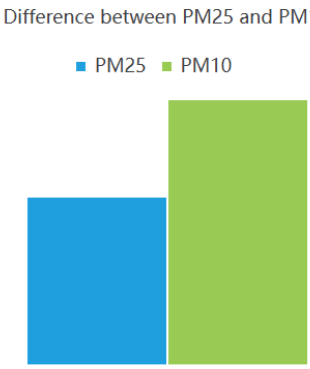

Figure 8. Concentration of PM10 vs PM25

4.2. Concentration change with seasons

As is shown in Figure 9, concentration of pollutants in winter is much higher than that in other seasons, which is consistent with our cognition.
Also, we can easily discover from the figure that there is a cliff-like rise/drop of concentration when winter begins/ends. It may indicate a quick change of temperature, because people need to consume fuel to fight against the cold weather. In that case, more pollutants are released to the air.

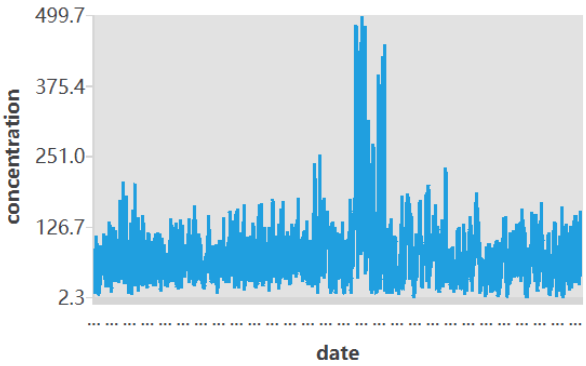

Figure 9. Season change

4.3. Performance

The loading process of this program occupies most resources. As is shown in Figure 10, when loading data, the RAM occupied by the program is 548.1MB, and the peak CPU consumption is 16.9%, which indicates a good performance. And after loading data, the consumptions are so little that we can dismiss them.

Figure 10. Performance

5. Acknowledgement

- Thanks to Teacher Zhao and Teacher Jin for their excellent lectures.
- Thanks to all the TAs for their patient instructions.