Parallel Coordinates Visualization Tool on the Air Pollution Data for Northern Malaysia

J. Joshua Thomas[1], Raaj Lokanathan[2], Justtina Anantha Jothi[3]

{joshopever@yahoo.com[1], raaj5671.r1@gmail.com[2] , justtina@kdupg.edu.my[3]}

Department of Computing, School of Engineering, Computing and Built Environment,
KDU Penang University College, Penang , Malaysia

Abstract. The paper describes based on techniques on the air pollution data visualization. This approach involves graph plotting algorithms with parallel coordinates that explore the raw data with interactive filtering that facilitates the insight of the materials that mixed and harm the population in northern Malaysia. By presenting the parallel coordinates method to visualize the parameter space that influence and visually identify the hazardous, moderate, unhealthy gaseous content in the air. The paper discusses the performance of the chosen visualization method and tested with northern region datasets.

Keywords: data visualization, parallel coordinates, visual representation

1 Introduction

The continuous bad air quality in Malaysia has got much attention especially the North of Malaysia which are Penang, Perlis and Kedah has got more attention during last year haze [1]. These states were affected by heavy haze where many victims were admitted to the hospital due to bad air condition. Most of the people had to wear a face mask before going to work. The tourist industry of Malaysia was badly affected because of the haze issue [2]. The city was totally covered by haze where ferries travelling along the Penang Island was not visible on the other side of the city.

In order to study the bad air quality issue, a visualization prototype which can be used to visualize the air pollution data in Malaysia. The air pollution data are in a CSV file format which was retrieved from University Sains Malaysia (USM) data monitoring station. For the current situation, benchmark data was used to visualize the air pollution data in Malaysia. Through the prototype, there are some interesting pattern have been found where usually cannot be found through the numerical values. Comparing the other state of the air pollution data, the researchers could come to a conclusion about the gaseous contents.

The weather data, in Malaysia are recorded automatically by the data monitoring station which are located in various places around each states. There are total of 52 data monitoring stations in Malaysia which records the weather data during each time interval for at least 6 times a day [3]. Moreover, it is also known that the weather data are multivariate [4]. This means usually there are more than 10 dimensions of data.

The current weather data in Malaysia has the same 10 dimension of data this includes the wind speed and the wind direction.

For our case, we planned to integrate the parallel coordinate's data visualization method, which is able to visualize the multivariate datasets. This method can be used to visualize and analyse the air pollution data in Malaysia.

The remaining section of this paper are organized as follows. Section 2 discusses about the related work done by other researchers on data visualization prototype. In the Section 3, the overview of the prototype was further discussed. The Section 4, the parallel coordinates data visualization method is discussed too. The process flow of the APVT is discussed on the Section 5 of this paper. The results are shown in Section 6 of this paper. At last, the conclusion of this paper is shown in Section 7.

2 Related Work

For this past years, a lot of researchers are developing a stand-alone prototype which is used to visualize the air pollution data. There are different data visualization methods have been used to develop the prototype to visualize the air quality. We have published parallel coordinates visualization for examination timetabling data [14] and it is more suitable for the categorical data to be visualized before execute the computational algorithms.

Fig. 1. Air pollution data recorded in observation center from Island Penang to Butterworth

Based on the research done from University of Seoul, Korea, they have developed a prototype which uses the cloud computing technology to visualize the urban air quality data [5]. This prototype is more towards the predicting the air pollution using cloud computing technology. Moreover, there is another research done by student of Central South University, China which collaborated with the Arizona State University, where they have developed a visual analytics system which is a web based. The web based system uses the parallel coordinates method to visualize the air quality data. However, a GIS view is used for spatial analysis which includes a scatter plot view too [6]. This research paper has been taken as an inspiration to develop a visualization prototype. On the other hand, the Microsoft Research students have developed a mobile app which forecasts the air quality based on the big data [7]. Data visualization research are being widely carried out by researchers all around the world. One of the problem universities

faces is creating an examination timetable. During the research phase, we have studied the research paper done by Joshua Thomas [16]. The author has visualized the student and subject data from the university and created a parallel coordinates visualization framework. This helped to get more idea on the implementation of the APVT software. After reading their research paper, we came to our conclusion that visual cues play a major role in data visualization software. We have applied some of the visual cues on our APVT software. This makes the APVT software more realistic. Besides that, Treinish [17] [18] also have carried out multiple researches on weather data visualization. However, his research was more towards the scientific visualization. Our APVT software was more towards information visualization. But we have modified to develop the APVT software. Whenever it comes to data visualization, data mining plays an important role. Previously, Karina Gibert [19] have proposed a data mining system which analyzes the air quality data. However, it is clearly known that using data mining techniques has its own weakness. Using data mining to analyze the data it is hard to understand the data because the analyzing process is almost same to the "black box" testing. Luo et al. [4] have extended the existing methods to handle the spatial distribution data which includes the weather data. Most of the approach which are used are more towards visualization rather than analysis. However, these related works done by other researchers, cannot be fully implemented in a visualization prototype. This is because, the APVT prototype can visualize the gaseous content of the air pollution data in Malaysia.

3 Prototype Overview

3.1 Data Collection

The Department of Environment (DOE) have installed the monitoring stations all around Malaysia which it retrieves the weather data real time. This weather data is then calculated and shown in the APIMS website which is part of the DOE. The APIMS website does shows only the Air Pollution Index (API) of each monitoring stations only. This details are available to public, whereas the other dimensions of the weather data are not available for the public. An official letter needed to be written to the DOE before accessing the real time weather data. Table shows the current 10 dimensions of data are being retrieved by the data monitoring station around Malaysia.

Tabel 1. List of Weather data collected at the monitoring stations in Malaysia

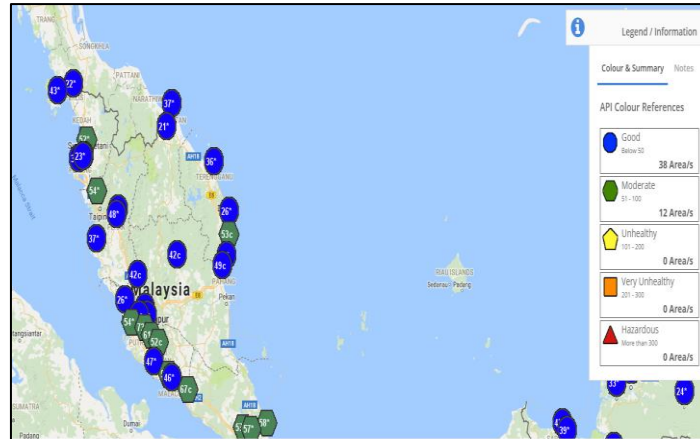|	
Wind Speed	Units m/s
Wind Direction	Bearing
Year	y
Month	m
Day	D
CO	ppb
NO	ppb
O_3	ppb
PM_{10}	μg/m3
Air Pollution Index	scale 100

Fig. 2. The ASMC haze map of Malaysia [20]

3.2 Prototype Feature

The APVT prototype is mainly focussed on visualizing the air pollution data only. Below are the list of the features the APVT has:

- Visualizing the gaseous content by different ranges of colours, the prototype has five different types of colours. Each colours has its own meaning the list of colours and the descriptions of it.
- Visualizing the air pollution data from different monitoring stations around Malaysia
- Filtering the air pollution data by date

Table 2. Gas ranges colors with its descriptions.

Colour	Descriptions
Black	Default
Blue	Good
Green	Moderate
Pink	Unhealthy
Orange	Very Unhealthy
Red	Hazardous

3.3 Prototype Code

Drawing the axis and plotting the CSV data takes alot of time. The code block below shows the partial code on how the parallel coordinates are plotted. Drawing the axis and plotting the CSV data takes alot of time. The code block below shows the partial code on how the parallel coordinates are plotted.

```
Public void paintComponent (Graphics g) {
   super.paintComponent (g);
   Graphics2D g2d = (Graphics2D) g;
   g2d.setRenderingHint(RenderingHints.KEY_ANTIALIASING,
RenderingHints.VALUE_ANTIALIAS_ON);
   g2d.setColor(Color.WHITE);
   g2d.fillRect(0, 0, getWidth(), getHeight());
   g2d.setFont(font);
   line2ds = new HashMap();
// draw the data as parallel coordinates from data
      if (csvData != null) {
      if (plot == TWO_AXIS_PLOT) {
      plotTwoAxis(g2d);
      } else if (plot == PARALLEL_AXIS_PLOT) {
      if (selectedAttributes.size() == 1) {
      selectedAttr = selectedAttributes.get(0);
      plotAttribute(g2d);
         } else {
      plotAllParallelAxis(g2d);
      }
         } else if (plot == SINGLE_ATTRIBUTE_PLOT) {
         plotAttribute(g2d);} } }
```

4 Prototype Overview

In this section, we will further discuss on the data visualization methods has been choosen for this prototype.

4.1 Parallel Coordinates

Parallel coordinate is one of the most popular tool which is used for visualizing and analyzing data [8]. Most of the researchers are using this visualization method to carry out their research analysis. Based on the article from Wikipedia, parallel coordinate was used before 1885 by Henry Gannetts. He used parallel coordinate to show the rank of states by ratios. Later after 79 years, which is during the 1959, a researcher developed the parallel coordinate as a system [9]. After that, many applications were developed by developers to do data mining or to analyze the data. Examples of application created was a collision avoidance algorithm for air traffic control.

In order for the Parallel Coordinate to work data is needed to be feed to it. Without data the parallel coordinate could not plot the coordinates. Given a small sets of data, the user can displayed it using a table form. Unless the size of the data is huge, which means it has more than million rows, the user cannot view it. A data sets with multiple attributes are named as multivariate data sets [10]. First of all, the user needs to add in the raw data into the visualization tool. After that, the preparation process will be started. This is where the process of filtering, extracting and calculating will be carried out. Filtering is a process where the amount of data

will be reduced. Finally, the plotted coordinates can be viewed. However, it might be hard to find out the data which has been left out [10]. This is because raw data has hundreds of rows.

The user can interact with the data using some of the interaction tool which is available for the parallel coordinate data visualization method. The reason interaction tool is created is because, users can view or analyze the information at their first sight. There are many interaction tool such as brushing, and axis. These interaction are one of the important interaction. Brushing is a process where the user is able to select a sample of subset which originally referred to an axis aligned rectangle for selection [11]. The other interaction tool is axis. The user is allowed to move the axis of the parallel coordinate by flipping it or reordering it. The value of the axes will be changed according to the way the user flips the axis [12].

However, there are some limitations for the parallel coordinate data visualization method. First of all, if the number of data is huge, then there is a chance of over plotting. This can cause the user hard to see. Besides that, if the data is in a categorical data format then the user could not view the plotted coordinates. This is because parallel coordinates only works well with numerical values. Thus, interaction needed to be used in order to view the parallel coordinate.

For our prototype, we have added a tooltip interaction for the parallel coordinates. The researchers can hover on the data line to know more details about the selected gaseous contents. Moreover, the way the APVT prototype plots the parallel coordinates is similar to the Alfred Inselberg plots the data. The each rows of the CSV data are stored in an array list. The stored data are plotted using the Java Line2D method. We have implemented the HashTable data structure for our prototype. The way we draw the axis is at first, we read the CSV file to get the columns name and then we draw the axis using a oval shape. After that, we plot the lines from the arraylist. Further plotting the parallel coordinate's works the same way we plot the raw data.

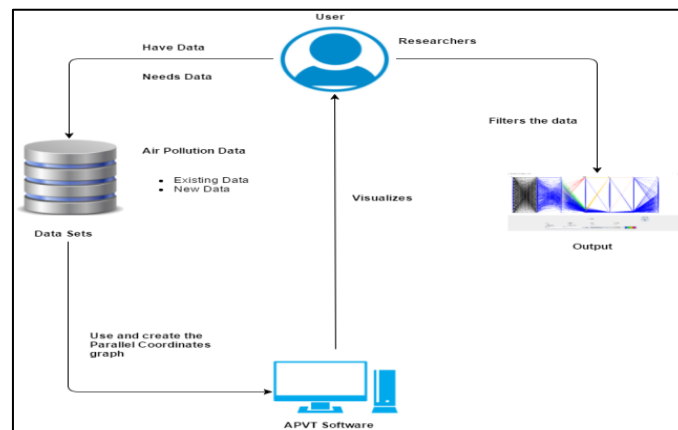## 5 Methodology

5.1 APVT Process Flow

Fig. 3. The process flow diagram of APVT

6 Experimental Results

The whole prototype has developed using Java programming language. A computer with JDK1.8 installed can be used to run the APVT prototype. Minimum of 2GB of memory has needed for the computer. This is because, parallel coordinates involves a lot of background processing. By installing the .jar file the APVT prototype has initiated successfully. During the testing phase, we have tested our prototype with two different data monitoring station data loaded into the prototype. From here we did some comparison on the selected gaseous content. First of all, we load the USM benchmark data and then we further plot the parallel coordinates. Below shows, the result of the USM parallel coordinates data plotted.
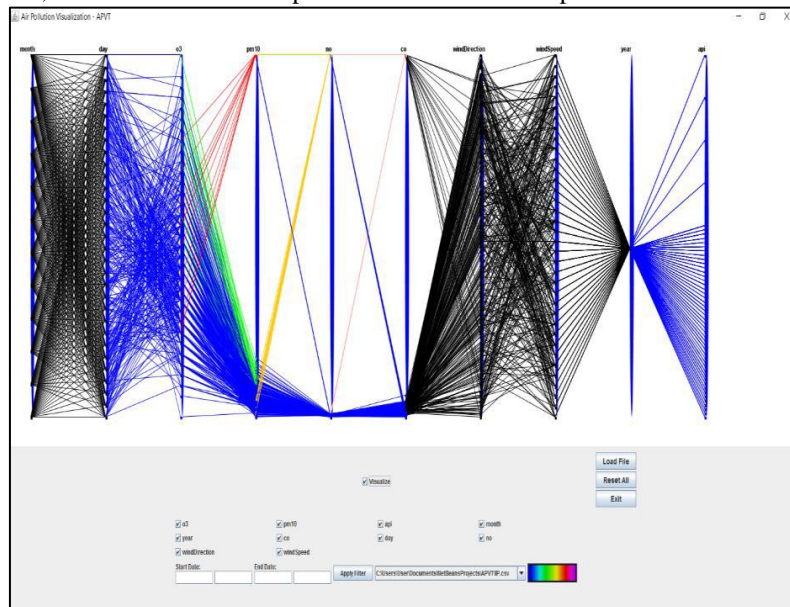

Fig. 4. The loaded USM benchmark data into the APVT prototype

In order to get deeper insights of the data, we have implemented the filter data function. While comparing the data we filter both of the CSV data first. Then we were able to get deeper insights on the plotted axis. For example, we filter the PM_{10} gas by date. Then we hover on the dataline in the PM_{10} axis so that we can know the accurate value of the line. As the PM_{10} of the USM data was showing in red color, we can conclude that the air quality was bad during the month of december. The same we done with the Kuala Kangsar CSV data. From here, we found out that, the gaseous content were very good in the month of December.

7 Conclusion and Future Work

In this paper we have shown a complete prototype which visualizes the air pollution in Malaysia. For this first implementation we have implemented only the parallel coordinates techniques. For the future recommendation, we have planned to implement the prototype into real time data retrieval. This means by collecting the weather data in real time we can even start to predict the air pollution in Malaysia for the next year. The main required method is regression needed to be used. From here we can analyze the trend of the air pollutioon data and then predict the API value. Moreover, we also have planned to integrate more data visualization techniques into our prototype. Once all this features are implemented, we can make our prototype into a web based system, where publics can start to use visualize the air pollution data. Besides that, we also planned to developed our own method to visualize the air pollution data. Currently, we are using the existing approach to visualize the gaseous contents. Developing our own method makes the APVT more reliable.

References

[1] Haze: Six areas in Perlis, Kedah and Penang record very unhealthy air quality (2015) Themalaymailonline.com. Available at: http://www.themalaymailonline.com/malaysia/article/haze-six-areas-in-perlis-kedah-and-penang-record-very-unhealthy-air-quality (Accessed: 30 May 2016).

[2] Tourist arrivals to Malaysia affected by haze, says deputy Tourism minister (2015) Themalaymailonline.com. Available at: http://www.themalaymailonline.com/malaysia/article/tourist-arrivals-to-malaysia-affected-by-haze-says-deputy-tourism-minister (Accessed: 30 May 2016).

[3]Malaysia Air Pollutant Index (n.d.) Apims.doe.gov.my. Available at: http://apims.doe.gov.my/v2/faq.html (Accessed: 30 May 2016).

[4] Qu, H., Chan, W., Xu, A., Chung, K., Lau, K. and Guo, P. (2007) "Visual Analysis of the Air Pollution Problem in Hong Kong", IEEE Trans. Visual. Comput. Graphics, 13(6), pp. 1408-1415. doi: 10.1109/tvcg.2007.70523.

[5] Won Park, J., Ho Yun, C., Sun Jung, H. and Woo LEE, Y. (2011) "Visualization of Urban Air Pollution with Cloud Computing", 2011 IEEE World Congress on Services, pp. 578-583. doi: 10.1109/SERVICES.2011.111.

[6] Liao, Z., Peng, Y., Li, Y., Liang, X. and Zhao, Y. (2014) "A web-based visual analytics system for air quality monitoring data", in 2014 22nd International Conference on Geoinformatics. Changsha: IEEE, pp. 1-6.

[7] Chang, E., Li, T., Yi, X., Li, M., Li, R., Zheng, Y. and Shan, Z. (2015) "Forecasting Fine-Grained Air Quality Based on Big Data", in KDD'15. China: ACM – Association for Computing Machinery.

[8] Zhang, J., Wang, W., Huang, M., Lu, L. and Meng, Z. (2014) "Big Data Density Analytics using Parallel Coordinate Visualization", International Conference on Computational Science and Engineering, p. 1115.

[9] Inselberg, A. (1985) "The Plane with Parallel Coordinates", The Visual Computer, 1(2), pp. 69-91.

[10]. Steinparz, S., Aßmair, R., Bauer, A. and Feiner, J. (n.d.) InfoVis – Parallel Coordinates. Graz University of Technology.

[11] Becker, R. and Cleveland, W. (1987) "Brushing Scatterplots", Technometrics, 29(2), p. 127. doi: 10.2307/1269768.

[12] Hauser, H., Ledermann, F. and Doleisch, H. (2002) "Angular Brushing of Extended Parallel Coordinates", IEEE Computer Society, p. 127.

[13] Nambiar, P. (2015) Penang air quality at unhealthy levels, NST Online. Available at: http://www.nst.com.my/news/2015/10/penang-air-quality-unhealthy-levels (Accessed: 30 May 2016).

[14] Thomas, J. J., Khader, A. T., & Belaton, B. (2011). A parallel coordinates visualization for the uncapaciated examination-timetabling problem. In Visual Informatics: Sustaining Research and Innovations (pp. 87-98). Springer Berlin Heidelberg.

[15]Thomas, J., Khader, A., Belaton, B. and Ken, C. (2012) "Integrated Problem Solving Steering Framework on Clash Reconciliation Strategies for University Examination Timetabling Problem", Springer Berlin Heidelberg, 7666, pp. 297-304. Available at:
http://link.springer.com/chapter/10.1007%2F978-3-642-34478-7_37#page-1 (Accessed: 3 September 2016).