Visual Analysis of the Air Pollution Problem in Hong Kong

Huamin Qu, *Member, IEEE*, Wing-Yi Chan, Anbang Xu, *Student Member, IEEE*, Kai-Lun Chung, Kai-Hon Lau, and Ping Guo, *Senior Member, IEEE*

Abstract—We present a comprehensive system for weather data visualization. Weather data are multivariate and contain vector fields formed by wind speed and direction. Several well-established visualization techniques such as parallel coordinates and polar systems are integrated into our system. We also develop various novel methods, including circular pixel bar charts embedded into polar systems, enhanced parallel coordinates with S-shape axis, and weighted complete graphs. Our system was used to analyze the air pollution problem in Hong Kong and some interesting patterns have been found.

Index Terms—Weather data visualization, polar system, parallel coordinates, air pollution, visual analytics.

1 Introduction

The rapid deterioration of the air quality in Hong Kong has aroused much attention [2]. The city is often cloaked in a heavy haze and the picturesque skyline has been barely visible over the past few years (see Fig. 1). The hazy air not only affects people's respiratory health and the tourist industry, but also significantly reduces Hong Kong's sustainable competitive advantages. A recent annual quality-of-life study [1, 4] suggested that serious air pollution has already made Hong Kong a less attractive place for expatriates. Hypotheses including external pollutants generated from mainland China, local pollution from too many vehicles, the use of high-sulphur coal in power plants, and a curtain wall effect resulting from dense buildings have been proposed.

To study this issue, the Institute for the Environment of the Hong Kong University of Science and Technology has developed a comprehensive atmospheric and environmental database on Hong Kong and the surrounding regions. The institute attempts to study the correlations between different attributes concerning air quality with classical analysis techniques. Some interesting patterns have been found, but it is difficult to obtain convincing results for high-level correlations that cannot be computed with solely numerical methods. Visualization techniques are hence required to assist in detecting trends and similarities, and in spotting possible correlations between multiple attributes which are unique for certain regions only.

Weather data possesses some special features. Weather data are usually recorded by automatic meteorological stations located in representative regions at regular time intervals, thus the data are intrinsically time-varying and contain inherited geographic information. In addition, weather data are typically multivariate and often consist of more than 10 dimensions. Wind speed and direction resulting in a vector are two of the most important attributes in weather data, which differentiates them from ordinary scalar multivariate data. These features, among others, make weather data visualization a challenging problem, especially since traditional visualization techniques like scatterplots and glyphs fail to achieve research goals.

In this case study, we integrate several well-established visualization techniques, namely the polar system, parallel coordinates, and the weighted complete graph, into a comprehensive system for weather data visualization and apply the system to analyze the air pollution problem in Hong Kong. Some novel techniques have been developed to address special challenges posed by weather data. In particular, we introduce circular pixel bar charts to detect correlations between wind direction, wind speed, and other attributes. We demonstrate how vectors and multiple scalar attributes in weather data can be effectively visualized by polar systems with embedded pixel bar charts and tailored parallel coordinates. The weighted complete graph is employed to reveal the overall correlation of all data dimensions and to determine the order of axes in parallel coordinates. Based on our system, some interesting patterns have been detected by domain scientists and valuable feedback about these visualization techniques is obtained.

The major contributions of this paper are as follows:

- We demonstrate how visualization can be used to attack a serious modern day issue. Domain scientists have gained new insights into the air pollution problem in Hong Kong with our advanced visualization system. We limit our study to Hong Kong weather data but the basic system, techniques, and lessons learned can be applied to general air quality analysis.

- We develop some novel methods to address special challenges posed by weather data, yet these novel methods are not limited to weather data visualization. For example, polar systems with embedded circular pixel bar charts can be exploited to visualize other multivariate data with both scalar attributes and vector fields. Weighted complete graphs can be used to determine the axis order of general parallel coordinates.

The remaining parts of this paper are organized as follows. Section 2 reviews related work on weather data visualization. The system overview is then described in Section 3, followed by a detailed discussion of our approach to weather data visualization in Section 4. The experimental results and discussion are presented in Section 5. Finally, we conclude the paper in Section 6.

2 Related Work

Weather data visualization is rarely considered a stand-alone problem. Instead, it is commonly addressed within the scope of multivariate data visualization, which sometimes overlooks the uniqueness of weather data including important vector values and time-series nature.

Treinish [20, 21] has conducted various research works on weather data visualization but his approaches lean more toward scientific visualization rather than information visualization. Healey et al. [9] used nonphotorealistic brush strokes for visualizing multidimensional information spaces like weather data. Tang et al. [19], on the other hand, applied a controllable texture synthesis technique to exploit natural textures for the same purpose. Although these methods yield effective and aesthetic results, they inevitably suffer from limited scalability as the number of individual visual channels of textures is believed to be around only three.

- *Huamin Qu, Wing-Yi Chan, Kai-Lun Chung, Kai-Hon Lau are with the Hong Kong University of Science and Technology, E-mail: huamin@cse.ust.hk, winchan@cse.ust.hk, cspeter@cse.ust.hk, alau@ust.hk.*
- *Anbang Xu and Ping Guo are with Beijing Normal University, E-mail: abxu@ieee.org, pguo@bnu.edu.cn.*

Manuscript received 31 March 2007; accepted 1 August 2007; posted online 27 October 2007.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

Fig. 1. Hong Kong's air pollution problem. The spectacular harbor view has been increasingly crippled by a massive haze [1, 3, 4].

.

In other cases, weather data visualization appears as a concrete application for a particular visualization tool. Luo et al. [15] extended existing methods to handle spatial distribution data including weather data. Their approach is more on visualization than analysis. Wilkinson et al. [22] proposed statistical measures for organizing multivariate displays and for guiding interactive exploration, which help users discover any patterns or outliners more easily in scatterplots. A comprehensive system VIS-STAMP [8] is especially designed for space-time and multivariate data, which consists of parallel coordinates, self-organizing maps, and pixel-based methods. However, these general approaches cannot be directly used to analyze the air pollution problem in Hong Kong. The wind factor is the most important issue in our problem. But it is not well addressed in these approaches. In this paper, we introduce an enhanced polar system along with other multi-variate data analysis methods which are tailored for air quality analysis.

3 System Overview

3.1 Data Collection and Processing

The Environmental Protection Department (EPD) and the Hong Kong Observatory (HKO) of the Hong Kong Special Administrative Region (HKSAR) have operated a number of continuous air quality and weather monitoring stations around Hong Kong (see Fig. 2). Air quality and weather information from many of these stations are published hourly and in real-time on EPD and HKO websites, and the Environmental Central Facility (ENVF) of the Hong Kong University of Science and Technology collects and archives these data for research. These datasets are then maintained within the ENVF Atmospheric & Environmental Database for future studies. This database is open to the public through the ENVF website [1] which has accumulated more than 10 million visits. Some primitive visualization techniques such as scatterplots and simple glyphs are used to display the data. The attributes recorded from these stations are summarized in Table 1. The data span more than 10 years and contain more than 13 dimensions.

3.2 Visualization Tasks

There are diverse visualization tasks for weather data analysis, which can be mainly classified into three categories:

- Finding correlations between different attributes. For example, in order to pinpoint air pollution sources, correlations between the air pollution index and any pollutants should be examined.

- Comparing the data from different stations. Similarities and differences between different regions are always of great interest for tracing informative patterns.

Table 1. Data Attributes Collected at Different Monitoring Stations.

Name	Unit
Precipitation	mm
Wind Direction	bearing
Air Temperature	Degree Celsius
Wind Speed	m/s
Dew Point	Degree Celsius
Relative Humidity	%
Sea Level Pressure	hPa
Respirable suspended particulates (RSP)	ug/m3
Nitrogen dioxide (NO_2)	ppb
Sulphur dioxide (SO_2)	ppb
Ozone (O_3)	ppb
Carbon monoxide (CO)	ppb
Solar Radiation	mw/cm2
Air Pollution Index (API)	scale 100
Contributed Pollutant to API	RSP, O_3, NO_2, SO_2 or CO

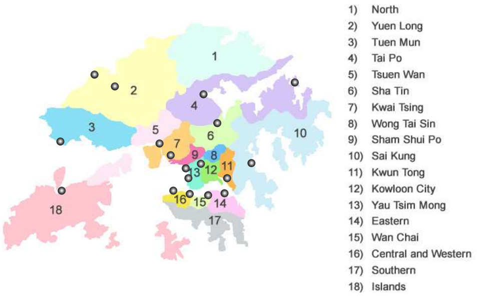


Fig. 2. Locations of different air quality monitoring stations shown as dots in 18 districts of Hong Kong.

- Detecting the trend for Hong Kong's weather and air quality. For time-series data, one important task is to predict the future based on the patterns observe today.

3.3 Visualization Modules

Based on the data we have and the visualization tasks to be conducted, we develop a comprehensive weather data visualization system. Our system consists of three major visualization modules, namely the polar system, parallel coordinates, and the weighted complete graph. These three modules all have their own strengths and weaknesses, which can be used separately or together to achieve a particular visualization goal. Some novel techniques are introduced and integrated into each module for more effective weather data visualization.

Weather data specifically contain an important vector dimension comprising wind direction and speed. Therefore, we introduce an embedded circular pixel bar chart for polar systems and a special S-shape axis for parallel coordinates to visualize this dimension in an intuitive manner. As there are complicated relationships among the attributes of multivariate weather data, weighted complete graphs are exploited to give an overview of correlations of all dimensions and help users in determining the axis order in the parallel coordinates display.

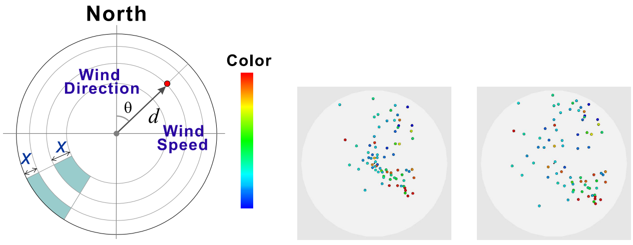
4 Visualization Techniques

In this section, we introduce the major visualization techniques employed in our system and the extensions that we have made to address the challenges posed by weather data visualization.

4.1 Polar System

In weather data, wind speed and wind direction are frequently used as the key attributes among all parameters. By visualizing the distribution of other attributes based on the wind profile, interesting patterns are more likely to be observed especially for data related to air quality,

[1] HKUST ENVF Website: http://envf.ust.hk/dataview/gts/current/



(a) (b) (c)

Fig. 3. Traditional polar system: (a) Encoding scheme; (b) Mapping radius without preserving the area; (c) Area-preserving polar system. Blue denotes low values and red denotes high values. The color bar is also used for other figures.

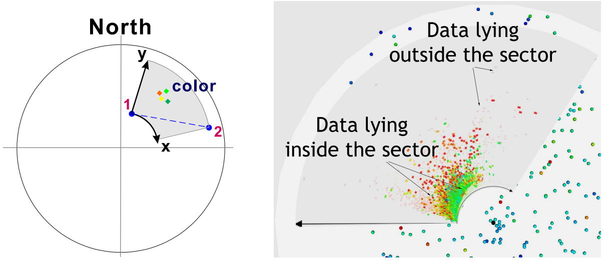
where different pollutants are correlated with the wind in some ways. Therefore, we apply the polar system, one of the most common representations for vectors, to encode wind so that the principal wind speed and direction are shown in an intuitive manner.

Fig. 3 shows the encoding scheme of the initial polar system. For a particular pixel on the polar plane, its distance from the center and the angle spanning from the north encode wind speed and direction respectively, while the pixel color encodes a third scalar attribute such that its relationship with wind speed and direction is clearly shown. To generate a more reliable representation for the underlying data, as a common practice in the environmental field, an area-preserving mapping should be applied on the distance from the center so that points located closer to the center are not overcompressed (see Fig. 3(c)). The simplest measure is to take the square root of the linearly computed distance value.

To cope with higher dimensions, we introduce the circular pixel bar chart which can be conveniently embedded into a polar system. The circular pixel bar chart is an extension to the pixel bar chart [13] which is derived from the traditional histogram. Three more attributes may be visualized with the *x*-position, *y*-position, and color of a pixel inside the circular pixel bar chart.

The initial polar system first acts as a guideline for users to explore other parameters based on the existing information provided. When users are interested in data items lying within a certain range of wind direction and/or wind speed, they can select the respective sector on the polar plane as described in Fig. 4. The wind information then becomes irrelevant and a circular pixel bar is produced only for data items falling within the sector of interest. A pixel bar is then banded and placed in the sector region. The width and height of the original pixel bar are then transformed to the arc and radius of the sector accordingly, showing the range of wind direction and speed that the pixel bar sector occupied. The x-position, y-position, and pixel color in this circular pixel bar now can encode three additional attribute values. For example, from the original polar system, users may observe that the amount of sulphur dioxide (SO_2) is remarkably high at a certain wind direction and wind speed. They may then further examine the temperature, amount of carbon dioxide (CO_2), and nitrogen dioxide (NO_2) for a higher level of correlation under those particular circumstances. It would be difficult to compute this kind of correlation by traditional correlation measures alone as they are only found in a relatively small subset of the data. Nevertheless, it can be notably important as users are more interested in extreme or abnormal cases in general.

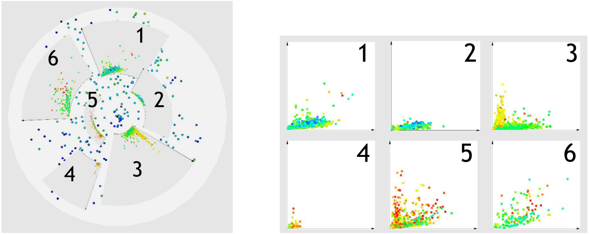
Sometimes users may want to examine the pattern of a sector against the overall one to see if the current sector exhibits similar or exceptional distribution with that of the complete dataset. We thus introduce a complementary circular pixel bar, which essentially encodes the data falling outside the sector, blended underneath the in-range sector. In effect, it highlights the selected data in the plot such that users can obtain an overall picture for tracking any provocative behavior. Subsequently, there is no re-normalization in the sector as opposed to the traditional pixel bar chart; data points are normalized based on the full range of data attribute values.



(a) (b)

Fig. 4. Polar system with circular pixel bar: (a) A sector selected by a user and a circular bar chart embedded into the sector; (b) Blending of circular pixel bar for data falling in the sector against one for its complement.

When the data size increases, multiple data items may feature identical wind direction and speed, which are hence mapped onto the same pixel location in the polar system. The display would then become cluttered and the overlapping of pixels may affect the accuracy of the visualization. Data filtering and clustering can reduce the data size to a manageable level. For example, we may only show one representative pixel with the average value of all the data items having similar wind direction and speed. After users select a sector and zoom in, a corresponding circular pixel bar chart will be drawn and the overlapping data items are usually mapped to different pixels.

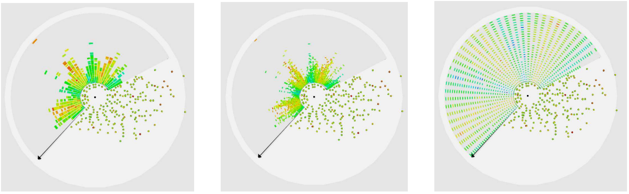


(a) (b)

Fig. 5. Comparing circular pixel bars with rectangular ones: (a) A polar system with multiple circular pixel bars; (b) Conventional pixel bars for the sectors. The overall patterns are preserved in the sector for comparison, and in-depth numerical analysis may be performed on the supplement rectangular pixel bars.

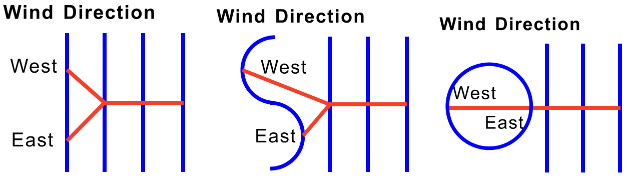
A polar system with circular pixel bars is more favorable for weather data visualization than a basic pixel bar chart mainly due to the unique vector information present in the weather data. Wind speed and direction are the major principal attributes from which further investigation begins. With circularly arranged pixel bars in a well-established polar system for handling vectors, the relationship between wind and individual attributes is clearly revealed. Such interaction also allows users to demand details for other parameters only on a subset of the featured data they are interested in, so that the result is less overwhelming and distracting information is minimized. This is especially important if multiple sectors are selected and compared by users because the wind information for these sectors is presented simultaneously in an intuitive manner (see Fig. 5(a)). One major concern for the circular pixel bar chart is that the regular shape of the basic plot is distorted, which may affect the accuracy of data analysis. To facilitate quantitative studies, the conventional rectangular pixel bars are also provided alongside the polar system, as presented in Fig. 5(b).

Our polar system can also visually represent the time information of the data. In the circular pixel bar chart, time information such as hour, day, month, and year can be treated as additional attributes of the data. For example, we can simply use the *x*-category to encode year, the *x*-position to encode month, and the *y*-position to encode day. The data pattern against time is now clearly shown, helping users detect



(a) (b) (c)

Fig. 6. Polar system with time information: (a) x-position, y-position, and color of the sector indicate the month of observation, amount of SO_2, and temperature, respectively; (b) the x-position now represents the day in which the entry was recorded; (c) the y-position now encodes the day and the x-position encodes the month.



(a) (b) (c)

Fig. 7. Different layouts of parallel coordinates: (a) Traditional layout; (b) Circular layout; (c) S-style layout.

any significant distribution at a particular point in time (see Fig. 6). Moreover, users may specify the time interval into which they wish to divide the data and subsequent polar plots will be generated to show the trend over time.

4.2 Parallel Coordinates

Parallel coordinates [11, 12] are powerful visual techniques where attributes are represented by parallel vertical axes. Each data item is represented by a polygonal line that intersects each axis at respective attribute data value. Parallel coordinates are also integrated into our system and several extensions have been made to improve their effectiveness for weather data visualization.

As discussed earlier, wind direction is one of the most important attributes in air quality analysis. However, the vertical axis used in traditional parallel coordinates is not good at encoding directions (see Fig. 7(a)). To address this problem, we experimented with different axis styles for wind direction. The polar system is widely used to encode wind. One scheme is to directly embed a polar system into the parallel coordinates (see Fig. 7(b)). However, the lines originating from the left side of a polar system will also pass through the right side, which causes ambiguity and/or too many crossing lines. The S-shape axis strikes an excellent trade-off (see Fig. 7(c)). It contains some characteristics of the polar system but can be naturally embedded into the parallel coordinates. In addition, as wind direction is an important axis, the S-shape axis stands out among all axes to attract users' attention.

Parallel coordinates have been thoroughly studied and many excellent techniques are available in reducing visual clutter caused by numerous crossing lines [6, 23]. However, the detailed information of the data might be lost during the "clustering" or "summarization" processes. To mitigate this problem, we also draw a scatterplot above every pair of neighboring axes for accurate quantitative analysis (see Fig. 8). Each line between the parallel axes then becomes a point in the scatterplot. Because points use less space than lines, the overall visual clutter is reduced.

4.3 Weighted Complete Graph

In our system, both polar systems and parallel coordinates can be used to detect possible correlations between multiple attributes. However, polar systems with embedded circular pixel bar charts can only reveal the relationship between five dimensions while only correlations between attributes represented by neighboring axes can usually be detected in parallel coordinates. Sometimes, users would want to know the overall relationship among all the data dimensions. In this paper, we propose a new technique, the weighted complete graph, as a guide map for polar systems and parallel coordinates to display the relationship between dimensions at higher levels.

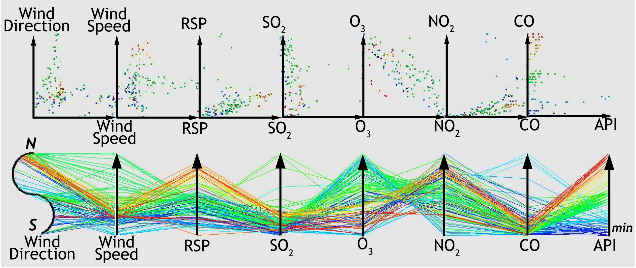


Fig. 8. Enhanced parallel coordinates with S-shape axis to encode wind direction and scatterplot to reveal bi-variate relationship between neighboring axes.

Recently, Sauber et al. [16] proposed multifield graphs as visual aids to intuitively provide information about the amount of correlations contained in each correlation field. However, one drawback of this technique is that the number of nodes in the graph grows exponentially with the dimensionality of the data. Although they proposed an optimal strategy to reduce the number of nodes to a certain degree, it is still not easy for users to perceive the correlation information of all nodes at once.

In our weighted complete graph, each node represents one dimension from the data and the weight of each edge encodes the correlation between the adjacent nodes. Compared with multifield graphs, the number of nodes in weighted complete graphs is greatly reduced. More over, the topology of the graph not only reveals the relationship between any two dimensions, but also shows the overall relationship among all dimensions. The uses of the weighted complete graph in the system are two-fold: to visualize the overview relationship among the dimensions, and to generate an optimized axis order for parallel coordinates interactively or automatically.

4.3.1 Definition and Distance Metrics

In a complete graph, every pair of graph vertices is connected by an edge. A weighted complete graph is a complete graph where each edge has an associated weight. In our application, we can use the node to represent the data attribute and the weight of the edge to encode the strength of correlations between two nodes. The weight of the edge between two nodes, which is the correlation strength, can be computed using different metrics. There are several correlation measures available for two variables. A common measure, called the correlation coefficient, can detect linear dependencies for normally distributed data. Another metric, pointwise mutual information which is based on entropy, is able to find more general dependencies but at the cost of much longer computation time. After testing with various metrics, we finally adopted the standard correlation coefficient in our system. The correlation between two dimensions X_i and X_j is defined as

$$C_s(X_i,Y_j) = \frac{\|(X_i-\overline{X}_i)(X_j-\overline{X}_j)^T\|}{((X_i-\overline{X}_i)(X_i-\overline{X}_i)^T)^{\frac{1}{2}}((X_j-\overline{X}_j)(X_j-\overline{X}_j)^T)^{\frac{1}{2}}} \tag{1}$$

4.3.2 Layout Optimization and Encoding Scheme

As viewers naturally interpret closely located nodes in a graph as strongly related [7], we need a layout algorithm for all the nodes in a weighted complete graph so as to reflect the relationships of all dimensions visually. Barnes and Hut [5] proposed a type of multi-scale algorithm for the simulation of astronomical systems, and the algorithm was then introduced to the graph drawing field [18]. On the other hand, energy based methods are popular for creating straight-line drawings of undirected graphs. Recently, Noack [17] proposed the edge-repulsion LinLog energy model whose minimum energy drawings reveal the clusters of the drawn graphs.

In our system, we use the LinLog energy model [17] with the Barnes-Hut algorithm to render weighted complete graphs. The weight of each edge is computed by the correlation metric introduced in the previous section.

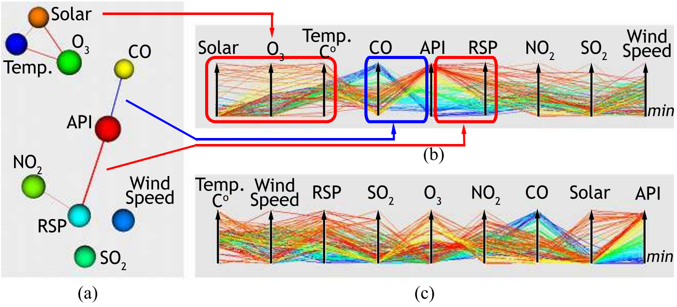


Fig. 9. Weighted complete graph: (a) Layout of weighted complete graph with node size encoding the accumulated correlation coefficients and edge encoding the correlation between two nodes. Edges with small weights are removed for clarity. (b) Parallel coordinates with a user-chosen axis order based on the weighted complete graph; (c) Parallel coordinates with a random axis order.

Furthermore, we introduce several encoding schemes to make the weighted complete graph drawing more meaningful and give a better overview of relationship information:

First, we can use the size of a node to encode the accumulated correlations between this node and all other nodes. With this strategy, the area of the node becomes proportional to the sum of the correlation measures between the node and the other nodes. Thus, a bigger node means that the node has larger accumulated correlation measures and may have a strong relationship with the other nodes.

Second, we can set a threshold for the edge weight and remove all low values and draw only the remaining ones to visually represent the strength of correlations. The width of the edge encodes the absolute value of the correlation coefficient. Positive and negative correlations can be differentiated using red edges and blue edges respectively. This filtering strategy helps users locate strong correlations easily.

Our layout optimization algorithm can also be applied to a subgraph. Sometimes users are just interested in the relationship among certain dimensions. In this circumstance, users can select certain nodes and the corresponding subgraph will be drawn using the layout algorithm (see Fig. 9(a)).

4.3.3 Axis Order Selection for Parallel Coordinates

In parallel coordinates, different orders of axes (each axis denotes one dimension in the multidimensional data) can reveal different aspects of the dataset and the order in which the axes are drawn is critically important for effective visualization. Axes representing attributes with potential correlation should be placed closely so that the relationship has a better chance of being revealed. The weighted complete graph can show the overall relationships of all attributes and those attributes having strong correlations tend to appear closer in the graph. Therefore, the weighted complete graph can be applied to guide users in adjusting the axis order of parallel coordinates. The axis order can be determined manually, semi-automatically, or automatically. Because there are only about 13 dimensions in our data and the layout can be easily grasped by users, our system thus provides a scheme for users to manually choose the axis order of the parallel coordinates. Fig. 9 shows an example of using a weighted complete graph to choose the axis order for parallel coordinates.

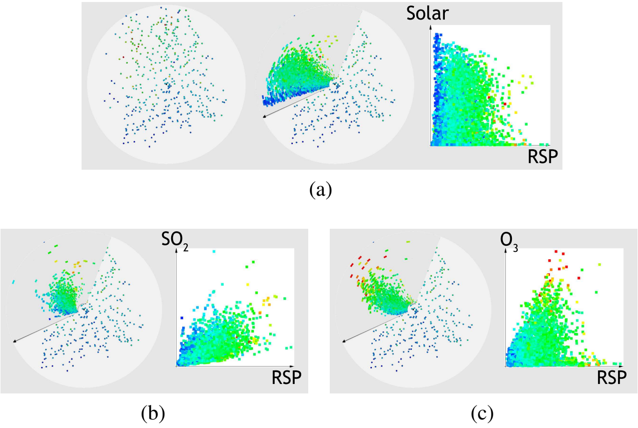


Fig. 10. Detecting the correlation between the Air Pollution Index (API) and other attributes when API is high: (a) Initial polar system with color denoting API value. The northwest sector is chosen, plotting RSP against solar radiation. (b) Plotting RSP against SO_2 instead, high API value (red pixels) are not found when SO_2 is high, revealing SO_2 contributed little to API. (c) Y-position now becomes O_3 clearly correlated with API. For (b) and (c), suspicious clusters (blue clusters behind green ones) are shown.

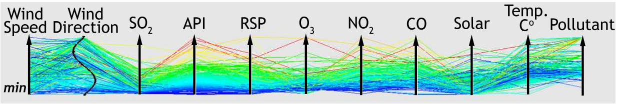


Fig. 11. Detecting correlations of the same set of data by Parallel Coordinates, with color denoting API value.

5 Experimental Results

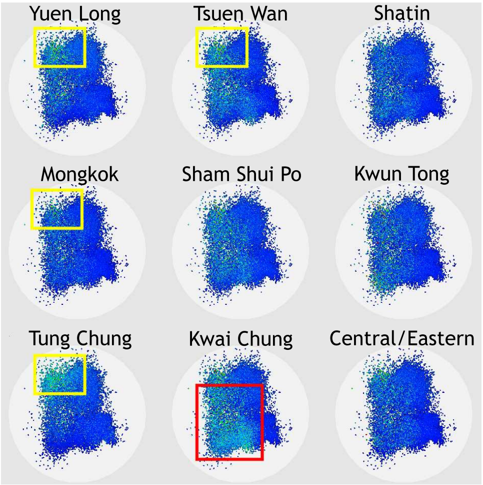
The whole system is developed and installed on a Dell Precision mobile workstation M70 with 2 GB memory and a 256M Nvidia Quadro FX Go 1400 graphics card. The VTK library is used for rendering. For all the experimental results in this section, interactive performance is achieved after the data are loaded into the memory.

5.1 Correlation Detection

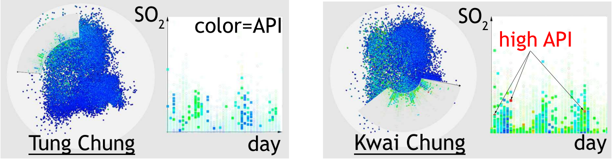
In the first set of examples, we tested the effectiveness of our method on detecting correlations between different attributes of the weather data. Finding correlations between multiple attributes is always one of the major visualization tasks for multivariate data visualization. In this section, we present how our system can assist users in detecting informative correlations.

The example in Fig. 10 aims at finding the correlation between the Air Pollution Index (API) and other dimensions. Our polar plots can help users examine any possible correlations between wind direction, wind speed, and any three other attributes chosen by users. We were only interested in serious air pollution conditions, which can be selected by picking a relevant sector. The API shows no relationship with solar radiation (Fig. 10(a)). The pixel bar here shows that API is highly correlated with Respirable Suspended Particulates (RSP). In fact, the positive correlations between the API with RSP and Ozone (O_3) in Hong Kong are known to experts [10], which the polar system demonstrates in Fig. 10(c). In addition, Fig. 10(b) suggests that SO_2, being mapped to the y-position, does not have a strong correlation with API though it is one of the major air pollutants taken into account for computing API. It has been known that the contribution of SO_2 is negligible compared with RSP and O_3 in API calculation. Significantly, two distinct blue clusters in the two latter figures become visible. The rationale behind such findings is still unknown and it is worth further study in the environmental aspect.

Next we demonstrate how similar conclusions can be drawn from the data represented in parallel coordinates in Fig. 11. A gradual color



(a)



(b)

(c)

Fig. 12. (a) Tracking the possible internal and external pollution sources through nine stations in the past three years. Pixel color represents the amount of SO_2 recorded in each individual station. (b)-(c) The detailed plots for station Tung Chung and Kwai Chung respectively.

change is perceived at the axis for RSP and O_3 as expected, indicating they are positively correlated with API. In contrast, a group of red lines passing through the SO_2 axis at a low value implies a high API reading not necessarily attributed to a large amount of SO_2. Moreover, the fact that the solar radiation and temperature are not related to API is revealed by the messy colored lines found at their vertical axes. Although it is more difficult to assess the impact of wind direction and wind speed in parallel coordinates, it is more effective to explore the correlations between multiple dimensions than a polar system. For instance, NO_2 and CO display some partial relationships in the graph that are worth investigating in the future.

5.2 Similarities and Differences

Hong Kong society mostly weighs external factors in tackling the air pollution problem. Many believe that the source of air pollutants are the factories on the Pearl River Delta, the manufacturing heart of China, located northwest of Hong Kong. Many very often ignore the pollution incurred locally. Other parties hold a conviction that the monopolistic power plants, and the excessive number of vehicles and vessels are responsible for the poor air in Hong Kong. To judge the two adverse statements, we first visualized the amount of sulphur dioxide (SO_2) recorded by nine air-monitoring stations in Hong Kong for the past three years in Fig. 12. As the energy sector and vehicular exhaust are the two major emission sources of SO_2, it may provide us clues on less apparent internal pollution. All of them exhibit a relatively high SO_2 amount when the wind speed is high and originating from the northwest direction. The high wind speed suggests that SO_2 is likely brought from the northwest region outside Hong Kong, which coincides with what most people suspect. Yet, much to our surprise, the station in Kwai Chung depicts a significantly different relationship between SO_2 and the wind. It recorded a large amount of SO_2 even with a southwest wind regardless of the wind speed. This probably implies that the pollution should be closely related to internal factors because the wind speed does not play any important role. Geographic location is then taken into consideration to explore the causes of the abnormal pattern to the southwest of Kwai Chung. It turns out that one of the world's busiest ports, the Kwai Tsing Container Terminal, is operating at the southwest of Kwai Chung. The local pollution observed can hence be attributed to cargo ships frequently entering and leaving the port. Although pollutants generated externally usually affect most stations, we should not overlook the internal factors that particularly cause severe air pollution in several districts with congested traffic on land or in the water.

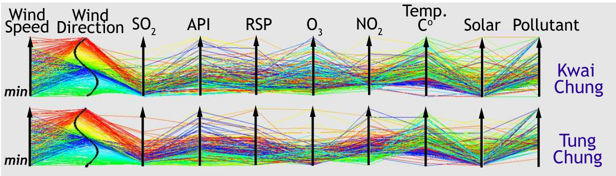


Fig. 13. Comparing two stations, Kwai Chung and Tung Chung with parallel coordinates using color to represent wind direction. Clusters of wind direction records are found in Tung Chung station but not in Kwai Chung.

For more in-depth analysis, we compared that Kwai Chung station data with another station to study how it behaves differently. In Fig. 12(b), we chose the northwest of Tung Chung station which also registers a high SO_2 reading with sector selection and plotted the amount of SO_2 against day with color encoding Air Pollution Index (API). The Kwai Chung data generally show a higher API value for higher recorded SO_2 values than the Tung Chung station. As we discussed in the previous section, SO_2 is not the main pollutant contributing to API under normal circumstances, so it again suggests that the local pollution resulting from heavy SO_2 emission by vessels is in fact a dominating factor in the Kwai Chung region [14].

Fig. 13 compares two stations, Kwai Chung and Tung Chung, using parallel coordinates. With the line color representing wind direction, the relationship between wind direction and all other dimensions can be easily reviewed. For example, at Kwai Chung, the API is not strongly related to wind direction whereas at Tung Chung, clusters of red and blue lines, representing winds from the north and northwest, can be seen at the API axis. Moreover, yellow and green lines that denote southwesterly winds are mainly connected to the lower API value in Tung Chung. However, for Kwai Chung, the color spreads diversely and a noticeable number of yellowish lines marks the highest API, which agrees with what we discovered from the polar system. Apart from that, these two stations experience fairly different distributions for O_3.

5.3 Time-Series Trend

Weather essentially varies with time on a seasonal basis, and may also signify certain trends over time when the global climate is also taken into account. Consequently, our system supports querying with time range and is able to generate desired results based on user-defined time intervals as shown in Fig. 14. In this example, data for three successive years for three different stations are visualized with time intervals set to six months. The two time periods, March to August and September to February, display similar distributions over the past three years. The directions of the winds observed are also opposite, which is typical in subtropical regions like Hong Kong which has distinguishable seasons. Furthermore, they also clearly show that the air quality is worse in the winter than in the summer.

Despite the fact that Hong Kong has experienced worse air quality in recent years, these time-series sequences do not present any growing trend in the API value. It is possible that the overall distribution remains rather constant and the variation that we are looking for is subtle and obscure. A more thorough investigation is then conducted for spotting any minor abnormality by showing the detailed distribution on a year-to-year basis with sector selection described in Fig. 15 for Kwai Chung. The red pixels clearly stand out from the first pixel bar, suggesting that local pollution from SO_2 emission was signifi-

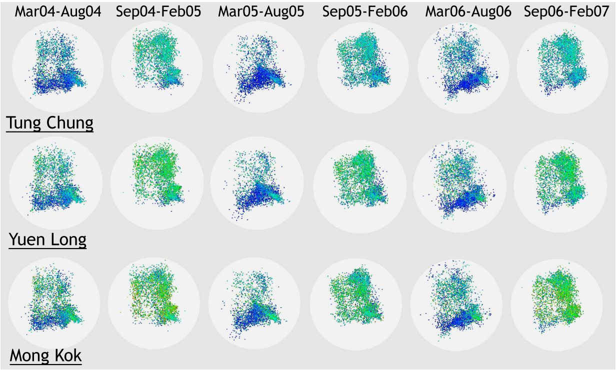


Fig. 14. Visualizing time-series data. Each row is comprised of polar plots for different stations, namely Tung Chung, Yuen Long and Mong Kok, in different periods of time from March 2004 to March 2007 at intervals of six months. The pixel color denotes the Air Pollution Index (API).

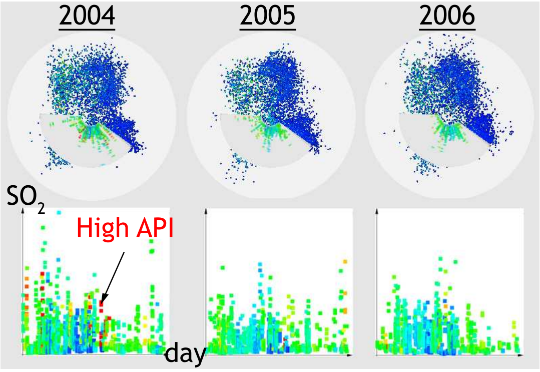
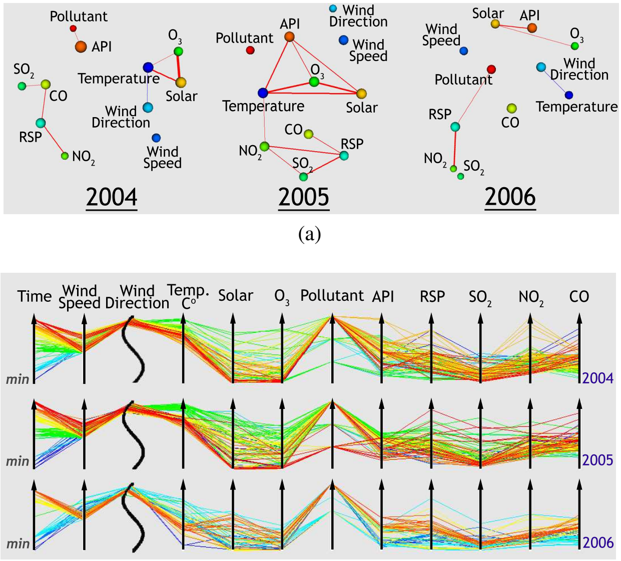


Fig. 15. Time-series polar plots for Kwai Chung station focusing on the impact of local pollution from the southwest direction. X-position, y-position, and color of the sector encode day, SO_2, and API accordingly. Prominent red pixels are mainly seen in year 2004.

cant in 2004, while a slight improvement is observed in the following years, inferring that local pollution has become less dominating in the district.

To visualize time-series data, we do not simply render a sequence of parallel coordinates. Instead, we first apply the polar system for sampling and then add a time axis in the parallel coordinates to show the time domain. As demonstrated in the previous examples, the polar system provides users an intuitive interface to select the data that they are interested in for in-depth analysis. Although we supplied the pixel bar for viewing more attributes, it was not able to give a general overview on every dimension in the data as parallel coordinates do. However, the major deficiency of parallel coordinates is that clustering and overlapping become so severe that it is difficult to spot any interesting patterns. By combining the two techniques, unnecessary data items are filtered, leaving the important time-series data effectively represented in parallel coordinates for detailed studies, as illustrated in Fig. 16(b). Users first selected the items with high RSP values, and then both the parallel coordinates and the weighted complete graph on this subset were shown for each year. The three weighted complete graphs in Fig. 16(a) generally indicate persistent correlations among SO_2, CO, RSP, and NO_2. Another group of correlated dimensions formed by temperature, solar radiation, and O_3 is also found in the graphs. With the help of the weighted complete graph (Fig. 16(a)) to arrange more correlated RSP, SO_2, NO_2, and CO closer together, one may quickly notice that polylines for the year 2006 plot are clustered together for most dimensions. Obviously, all three of them yield similar figures with the temperature varying dramatically for 2006, in contrast to the data for 2004 and 2005. In the display for 2004, unusual yellow lines are seen at high RSP and NO_2 values, resulting in the largest API in this set of data. Such abnormalities are not found in the other results, which indicates that during this particular time in 2004, some mysterious factors caused these unexpectedly high NO_2 records. For O_3, solar radiation, and temperature, they reveal a rather constant pattern through the first two years, also seen from the weighted complete graphs, that these parameters remain highly correlated throughout the two years while a decreasing trend of O_3 is observed when strong winds are blowing from the north.



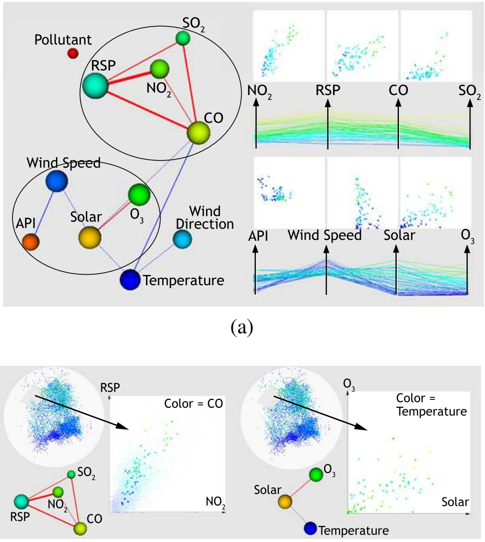
(a)

(b)

Fig. 16. 3-year time-series data of Yuen Long district that are constrained to a range of wind speed and direction by sector selection: (a) Weighted complete graph for each year with edge width encoding the correlation strength; (b) Parallel Coordinates with a time axis. Color also denotes time value for clarity.

5.4 Discussion

From the experiments, we can see that the three visualization techniques all have their advantages and disadvantages. The polar system is good at revealing the correlation of the wind and another attribute. With the embedded circular pixel bar chart, at most five attributes can be displayed to users. If users want to find the overall pattern for all attributes, parallel coordinates can be exploited. However, the effectiveness of parallel coordinates highly depends on the order of axes. To solve this problem, we compute a correlation value between every two dimensions. These dimensions and their correlation values then naturally form a weighted complete graph. After filtering out the edges with low weights and computing the layout of the graph based on a force model, the overall relationship of these dimensions can be revealed. Users will then have a big picture and can further explore the details using the polar system and parallel coordinates. Therefore, these three visualization techniques can complement each another's strengths. Next we demonstrate how the three modules can be used together to analyze the dimension correlations of the year 2006 data at the Yuen Long station. Users first select a small sector of data from the polar system with high RSP under strong northwest wind. Then, based on the corresponding weighted complete graph in Fig. 17(a), we immediately observe some strong positive correlations among the upper attributes and negative correlations among the lower attributes in the graph. When the axes in parallel coordinates are ordered accordingly, these relationships are shown clearly and visual clutters are minimized. While parallel coordinates are good at displaying the general correlations between multiple attributes, users may switch back to the polar system to plot these dimensions in the embedded pixel



(a)

(b)

Fig. 17. Using a weighted complete graph as a visual aid in exploring dimension correlations for year 2006 data of the Yuen Long station: (a) By arranging more correlated attributes together, positive and negative correlations are clearly shown in the parallel coordinates; (b) Users can also plot the attributes demonstrating interesting relationships in the weighted complete graph as the embedded pixel bar in the polar system.

bar for more quantitative analysis (Fig. 17(b)). The result suggests that the positive correlation between NO_2 and RSP found in this sector also holds for the whole dataset by comparing with the complementary plot underneath. On the other hand, the positive correlation between solar radiation and O_3 is barely revealed as the distribution is rather dispersed.

The system received very positive feedback from the domain scientists at the ENVF of the Hong Kong University of Science and Technology (HKUST). Among the three modules, the domain scientists found that the enhanced polar system with embedded circular bar charts was the most useful one. The scheme is similar to their traditional visual analysis method but nicely integrates several very useful techniques. The S-shape axis in the parallel coordinates has the characteristics of the polar system and makes wind direction stand out among all axes. Therefore, it is readily accepted by them to encode wind direction. Unlike the polar system, parallel coordinates can reveal all data attributes. However, they prefer to use scatter plots together with parallel coordinates because more accurate quantitative analysis can be conducted using scatter plots while parallel coordinates are good for qualitative analysis. The weighted complete graph can visually show the results of the traditional linear correlation analysis technique and surely provides a visual aid for users to choose the order of parallel coordinates and guides users to explore the correlation of multiple dimensions using the polar system and parallel coordinates. To get the full potential of the weighted complete graph, the domain scientists suggest that a more advanced nonlinear correlation metric should be explored. This issue is worth further study.

6 Conclusions and Future Work

In this paper we have presented a comprehensive system for weather data visualization. Many visualization techniques are integrated into our system and several novel techniques are developed. Our system has been used to analyze the air pollution problem in Hong Kong and some interesting patterns have been detected by the domain scientists using our system. In the future, we plan to continue our work with domain scientists and make our system available to the public through the website of the ENVF at the HKUST.

Acknowledgements

The authors wish to thank Dr. Zibin Yuan at the Institute for the Environment of the Hong Kong University of Science and Technology for his help in this project. The data used in this study were measured by the Environmental Protection Department and the Hong Kong Observatory of the Hong Kong Special Administrative Region. We would like to thank the anonymous reviewers for their valuable comments. Shauna Dalton helped proofread the paper. This work was supported by RGC grant CERG 618705.

References

[1] Hong Kong's air pollution hits the hightest in seven years. BBC Chinese, September 2002.

[2] Hong Kong's bad air days. Cover Story, Time Asia, 13 December 2004.

[3] Let there be light. Time Asia, 8 May 2006.

[4] HK's air pollution drags it down in eyes of expats, survey shows. South China Morning Post, 15 March 2007.

[5] J. Barnes and P. Hut. A hierarchical O(N log N) force-calculation algorithm. *Nature*, 324(4):446–449, 1986.

[6] M. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 19(3):378 – 394, 2003.

[7] E. Dengler and W. Cowan. Human perception of laid-out graphs. In *Proceedings of the International Symposium on Graph Drawing*, pages 441 – 443, 1998.

[8] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, 2006.

[9] C. G. Healey, L. Tateosian, J. T. Enns, and M. Remple. Perceptually based brush strokes for nonphotorealistic visualization. *ACM Trans. Graph.*, 23(1):64–96, 2004.

[10] J.-P. Huang, C. H. Fung, K. H. Lau, and Y. Qin. Numerical simulation and process analysis of typhoon-related ozone episodes in Hong Kong. *J. Geophys. Res., 110, D5301, doi:10.1029/2004JD004914*, 2005.

[11] A. Inselberg. Multidimensional visualization with applications to multivariate problems. *SIGGRAPH Course Notes*, 2002.

[12] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multidimensional geometry. In *Proceedings of the IEEE Symposium on Visualization*, pages 361–378, 1990.

[13] D. A. Keim, M. C. Hao, and U. Dayal. Hierarchical pixel bar charts. *IEEE Trans. Vis. Comput. Graph.*, 8(3):255–269, 2002.

[14] K.-H. Lau, W. M. Wu, C. H. Fung, R. C. Henry, and B. Barron. Significant marine source for SO_2 levels in Hong Kong. *Civic-exchange Environmental and Conservation reports*, 2005.

[15] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. *Proceedings of the Symposium on Data visualisation*, pages 29–38, 2003.

[16] H. T. Natascha Sauber and H.-P. Seidel. Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):917 – 924, 2006.

[17] A. Noack. Energy-based clustering of graphs with nonuniform degrees. In *Proceedings of the International Symposium on Graph Drawing*, pages 309–320, 2005.

[18] A. J. Quigley and P. Eades. Fade: Graph drawing, clustering, and visual abstraction. In *Proceedings of the International Symposium on Graph Drawing*, pages 197– 210, 2000.

[19] Y. Tang, H. Qu, Y. Wu, and H. Zhou. Natural textures for weather data visualization. *Proceedings of the International Conference on Information Visualization*, pages 741–750, 2006.

[20] L. A. Treinish. Multi-resolution visualization techniques for nested weather models. *Proceedings of IEEE Visualization*, pages 513–516, 2000.

[21] L. A. Treinish. Visual data fusion for applications of high-resolution numerical weather prediction. *Proceedings of IEEE Visualization*, pages 477–480, 2000.

[22] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

[23] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 105 – 112, 2003.