Visualization of Multivariate Time Series pollutant variables in Malaysia

`Ulya Abdul Rahim1 and Nurulkamal Masseran2
1,2 Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia

2E-mail: kamalmsn@ukm.edu.my

Abstract. Visualization and exploratory analysis is a crucial preliminary part of any data analysis process. Several visualization approaches have been introduced to evaluate the behaviors of time-dependent data. However, the visualization technique tends to be challenging when the data are high-dimensional and voluminous. Environmental data such as pollutant variables are often collected in multi-variables form and over time, resulting in a form of multivariate time-series data. To deal with this issue, this study provides several graphical approaches and methods which include the plots of multiple individually on a time-series, correlation matrix visualization and smoothing multivariate time-series. A case study involving data on air-pollution variables in Klang, Malaysia have been analyzed. The results found the all the methods able to provide an informative visualization on the behavior of multivariable time series of pollutant data.

1. Introduction
Multivariate time-series data are a type of data derived from a combination of related variables of any phenomenon being studied. At present, multivariable time-series data are available in various fields of study, particularly in science, engineering, and business. Thus, a good visualization technique should be adopted to provide good preliminary insights into the behavior of multivariable time-series data. In general, the visualization approach aims to exploit the phenomenal abilities and capabilities of the human eye to detect the structures and arrangement in images [1]. The visualization of time-dependent series data has a long history. According to Tufte [2], time-series plots were performed for the first time in the 10th or 11th century. They appeared in an illustration of planetary orbits in a text from a monastic school. In science, time-series charts were rediscovered by Lambert around the 17th century. Lambert [3] applied line graphs to present the periodic variation in soil temperature in relation to depth under the surface. At present, a wider repertoire of methods is used and applied to visualize time-dependent data in many areas.

The data on air pollution in Malaysia consist of five main variables, namely, carbon monoxide (CO), nitrogen dioxide (NO₂), oxide (O₃), particulate matter (PM₁₀), and sulfur dioxide (SO₂). Several researchers have analyzed and visualized the multivariable pollutants in Malaysia using different approaches. For example, Al-Dhurafi et al. [4], Sánchez-Balseca and Pérez-Foguet [5] used compositional approach in which the proportion of each pollutant is considered as a component of compositional data. Masseran and Safari [6], Martins et al. [7], Masseran and Mohd Safari [8] modeled and visualized the behavior of rare occurrences of air pollution indices in term of Extreme-Value model. Al-Dhurafi et al. [9] described and visualized the multivariable pollutants in terms of its distributional form, which was derived from its structure and descriptive status. Chen and Wu [10], and Masseran and Safari [11] evaluated and visualized the fluctuation of pollutant variables which corresponds to five
determinant states of the Markov chain. Kumar and Ridder [12], and Masseran [13] modelled and visualized the fluctuation of PM10 data using ARIMA–GARCH model. Al-Dhurafi et al. [14] analyzed and visualized the dependency of multivariable pollutants corresponding to their unhealthy air-quality status using generalized Pareto distribution and Pickands dependence function plots. Masseran and Hussain [15] modelled and visualized the dependency fluctuation among the air pollution variables using dynamic Copula approach.

In this study, we present alternative visualization techniques to analyze multivariate time-series data. In Section 2, we describe the study area and main components of the variables involved. An overview of the visualization techniques for this air-pollution field are plotted and discussed in Section 3. Section 4 concludes the study with general explanations and remarks on using the described techniques and suggestions for further research.

2. Data and area of study

The air-pollution index data were obtained from the Department of Environment, Malaysia from January 1, 2007 to December 31, 2016. This study focuses on Klang, a large city with land area of approximately 573 sq. km. As an urban and industrial region, Klang is densely populated. It is recognized as the 13th busiest transshipment port and 16th busiest container port in the world. The major industrial activities in Klang are importation and exportation. However, due to industrial activity and rapid economic growth, the city’s air quality can be affected negatively if not monitored properly [16]. Thus, this study proposes a multivariable visualization as an alternative tool for monitoring pollutant behavior in Klang. The data has a small percentage of missing values at random points. Thus, to estimate these missing values, the method of single imputation based on the average of the last known and next known observations is used. This method is easy to implement and can provide good results for missing data with random behavior [17].

3. Visualization techniques for multivariable air-pollution data

The multivariate plots are useful in visualizing the relationships or interactions among attributes. The purpose of this plot is to learn about the trends, correlations, and distributions among the variables. Furthermore, using the bivariate scatter plot is important to determine the spread over groups of data, typically for a pair of variables. Both of these plots are available in R programming software.

3.1. The function of autoplot

The autoplot() function from ggplot2 package was used to create a ggplot version. Previously, we transformed the mts object into a zoo object through the as.zoo function. The present study illustrates the use of the autoplot() function to plot the trends of multivariate time-series data of pollutant variables in Malaysia. By producing standard time-series plots of the data, we can conduct visualizations either simultaneously on a single plot or individually on a time-series plot. For example, figure 1 illustrates 10 years of daily air-pollution data on five main pollutant variables (CO, NO2, O3, PM10, and SO2) in Klang.

In addition, figure 2 shows the trends of overall variables in a single plot. It does not present the empirical distribution of the data, the correlation among the variables, or the significance levels of the series. As shown in figures 1 and 2, most of the air-pollution data fluctuate around their means. However, for the data on variables CO, PM10, and SO2, several “shock points” or outliers are distant from the mean. This result implies that for a particular period, the volatility effect occurred in the CO, PM10, and SO2 data. Although figures 1 and 2 indicate the simultaneous fluctuations of multivariable pollutant data, these graphs lack information on the behaviors of time series.
Figure 1. Individual plot of daily air-pollution data in Klang for 2007–2016

Figure 2. Overall variables on a single plot of daily air-pollution data in Klang for 2007-2016
3.2. Correlation matrix visualization
This section presents the `ggcorrplot()` function for plotting the correlation analysis among the variables. A correlation matrix describes the correlation coefficients among the variables in the form of a table. Usually, a correlation matrix is presented in square form, as shown in figure 3 that the columns and rows with the same variables.

Figure 3 presents the visualization of the correlation matrix between the variables. The line of 1.00 s is the main diagonal, which goes from top left to bottom right. This line indicates that each variable is perfectly associated with itself. The value of the correlation shown above the main diagonal is a mirror image of those below the main diagonal, which implies the symmetrical matrix.

![Correlation Matrix Visualization](image)

**Figure 3.** Visualization of correlation matrix between variables.

In statistics, the value of correlation coefficient, $r$ is always between +1 and –1. Table 1 presents the range of the $r$ value, which is used to interpret the correlation coefficient.

| $r$-value | Negative linear relationship         | $r$-value | Positive linear relationship |
|----------|--------------------------------------|-----------|-----------------------------|
| 1        | Perfect downhill                      | 0.7       | Strong uphill               |
| -1       | Perfect uphill                        | 0.5       | Moderate uphill             |
| -0.7     | Strong downhill                       | 0.3       | Weak uphill                 |
| -0.5     | Moderate downhill                     |           |                             |
| -0.3     | Weak downhill                         |           |                             |
| 0        | No linear relationship                |           |                             |

3.3. Correlation matrix corresponds to their distribution plot
This section illustrates a correlation matrix that corresponds to a distribution plot. The function `chartCorrelation()` based on PerformanceAnalytics package, which can be used to generate this plot. Figure 4 illustrates the chart correlation of five main pollutant variables in Klang.
According to figure 4, the diagonal matrix represents the distribution of each variable (CO, NO$_2$, O$_3$, PM$_{10}$, and SO$_2$). From this plot, we can conclude that not all pollutant variables follow a normal distribution. The bivariate scatter plots with a fitted line are presented at the bottom of the diagonal matrix. The value of the correlation plus the significance level indicated with asterisks are shown on top of the diagonal. Note that the asterisk values **, and ** indicate the significant variables at 5%, and 1% significance level, respectively.

3.4. \texttt{mvts()} plot function

Peng \cite{18} proposed the idea behind the \texttt{mvtsplot()} function and package for the purpose of revealing the underlying patterns in the data by using color. The \texttt{mvtsplot} function provides a useful summary of multiple parallel time series. An image plot is the basic plot of the time-series matrix. This plot uses images to provide quick insights into whether the values of each time series are low, medium, or high. By default, this function uses three colors, namely, purple, grey, and green, to divide the data into low, medium, and high categories, respectively. Based on the work of Harrower and Brewer \cite{19}, the colors used in this plot are obtained from palettes provided by the RColorBrewer package. On average, most of the variables have higher and lower values of each time series based on the color shown. For the time-series plot of the median value, the data have lower variation around their medians according to the plot at the bottom. By using \texttt{mvtsplot()}, the graph in figure 5 shows the “raw” data of each pollutant variable simply coded for each variable (CO, NO$_2$, O$_3$, PM$_{10}$, and SO$_2$) independently from low (magenta) to high (green). The plots on the margins are boxplots for the individual series on the right and a time series of the median values across all series on the bottom, which aim to summarize the
overall level of the series. This figure provides a visualization of the variation that presence among the five compositional pollutant variables; CO, NO₂, O₃, PM₁₀, and SO₂. The figure also detecting some clustered events in term of a high, medium, and low values on each pollutant variable. However, this information is not clearly explained in figure 5. Thus, to overcome this problem, a smoothing technique need to be adopted.

The number of arguments for the `mvtsplot` function are given as:

```r
library(mvtsplot)
#loading required package: RColorBrewer)
mvtsplot(Data, group = NULL, xtime = NULL, norm = c("internal", "global"), levels = 3, smooth.df = NULL, margin = TRUE, sort = NULL, main = "", palette = "PRGn", rowstat = "median", xlim, bottom.ylim = NULL, right.xlim = NULL, gcol = 1)
```

**Figure 5.** Multivariate time-series data plot

### 3.5. Smoothing multivariate time-series data

Smoothing technique is a useful visualization that helps to spot trends in noisy data and compare the trends between two or more fluctuating time series. Although the smoothing does not provide a model, it can be a first step in describing several components of the series. When the plot is slightly noisy, it
can be simplified by smoothing the individual series. This plot provides an idea of how the data are organized along the time axis. Figure 6 shows the smoothing multivariate time-series data in Klang. Based on figure 6, the variation that among the five compositional pollutant variables accompanied with their clustered events in term of a high, medium, and low values are clearly represented. On the other hand, the `mvtsplot()` function has an option called `smooth,df` argument, which can be used to apply a natural spline smoother in the time-series matrix for each of the time series. The `smooth,df` argument specifies the number of degrees of freedom to be used in the natural spline smoother for each time series. Each observed value is substituted by its fitted value once the smoother is fit to the data and then plotted. The smoothing trend line indicates that the level of air quality influenced by these clustered events.

![Figure 6. Smoothing multivariate time-series](image)

Apart from that, by increasing the value of smoothing parameter and normalizing the data, a rougher visualization can be obtained. Figure 7 provides further simplification by “normalizing” the time-series data, and then converting and assigning colors individually based on the range of values in each series. The normalized time-series data in figure 7 clearly show the trend and pattern of the level and clustered events respectively.
4. Conclusion
This study has successfully presented several visualization methods for multivariate time series that are useful for exploratory analysis prior to formal model fitting. Several steps of data visualization on multivariable air pollutants data has been proposed. First, plot for a trend of pollutant variables by using autoplot() function on a plot individual graph and a single plot. Next, ggcorrplot() and chartCorrelation() function have been propose to be used to plot the correlation matrix and its distribution. Apart from that, the use of mvtsplot() function has been proposed in order to reveal the underlying patterns in the data by using color, and smoothing argument was added to apply a natural spline smoother in the time-series matrix. This study revealed and provided the visualization function and applied it to ambient air-pollution data. For future study, we encourage interested users to apply their own functions and modifications to the existing code to suit their own research purposes.

Acknowledgement
The author is indebted Malaysian Department of Environment for providing air pollution data. This research would not be possible without the sponsorship from the Universiti Kebangsaan Malaysia (grant number FRGS/1/2014/SG04/UKM/03/1 and DIP-2018-038)

References
[1] Baughman A K, Gao J, Pan J Y and Petrushin V A. 2015 Multimedia data mining and analytics: disruptive innovation (Cham, Switzerland: Springer)
[2] Tufte ER 1983 *The visual display of quantitative information* (Cheshire, CN: Graphics Press).
[3] Lambert J H 1779 *Pyrometrie, oder Vom Maasse des Feuers und der Wärme* (Berlin, Germany: Haude und Spener)
[4] Al-Dhurafi N A, Masseran N and Zamzuri Z H 2018 Compositional time series analysis for air pollution index data *Stochastic Environmental Research and Risk Assessment*** 32** (10) 2903–2911
[5] Sánchez-Balseca J and Pérez-Foguet A 2020 Spatio-temporal air pollution modelling using a compositional approach *Heliyon*** 6** (9) e04794 1–13
[6] Masseran N and Safari M A M 2020 Risk assessment of extreme air pollution based on partial duration series: IDF approach *Stochastic Environmental Research and Risk Assessment*** 34** (3–4) 545–559
[7] Martins L D, Wikuats C F H, Capucim M N, de Almeida D S, da Costa S C, Albuquerque T, Carvalho V S B, de Freitas E D, de Fátima Andrade M and Martins J A 2017 Extreme value analysis of air pollution data and their comparison between two large urban regions of South America *Weather and Climate Extremes*** 18** 44–54
[8] Masseran N and Mohd Safari M A 2020 Intensity–duration–frequency approach for risk assessment of air pollution events *Journal of Environmental Management*** 264** 110429
[9] Al-Dhurafi N A, Masseran N, Zamzuri Z H and Safari M A M 2018 Modeling the Air Pollution Index based on its structure and descriptive status *Air Quality, Atmosphere and Health*** 11** (2) 171–179
[10] Chen J-C, Wu Y J 2020 Discrete-time Markov chain for prediction of air quality index *Journal of Ambient Intelligence and Humanized Computing* (Online first article).
[11] Masseran N and Safari M A M 2020 Modeling the transition behaviors of PM$_{10}$ pollution index *Environmental Monitoring and Assessment*** 192** 441
[12] Kumar U and Ridder K D 2010 GARCH modelling in association with FFT–ARIMA to forecast ozone episodes *Atmospheric Environment*** 44** (34) 4252–4265
[13] Masseran N 2017 Modeling fluctuation of PM$_{10}$ data with existence of volatility effect *Environmental Engineering Science*** 34** (11) 816–827
[14] Al-Dhurafi N A, Masseran N, Zamzuri Z H and Razali A M 2018 Modeling Unhealthy Air Pollution Index Using a Peaks-Over-Threshold Method *Environmental Engineering Science*** 35** (2) 101–110
[15] Masseran N and Hussain S I 2020 Copula modelling on the dynamic dependence structure of multiple air pollutant variables *Mathematics*** 8** (11) 1910
[16] Masseran N 2021 Power-law behaviors of the duration size of unhealthy air pollution events *Stochastic Environmental Research and Risk Assessment* (Article in press).
[17] Masseran N, Razali A M, Ibrahim K, Zaharim A and Sopian K 2013 Application of the single imputation method to estimate missing wind speed data in Malaysia *Research Journal of Applied Sciences, Engineering and Technology*** 6** 1780-1784
[18] Peng R D 2008 A method for visualizing multivariate time series data *Journal of Statistical Software*** 26** (7) 1–24
[19] Harrower M and Brewer C A 2003 ColorBrewer. org: an online tool for selecting colour schemes for maps *The Cartographic Journal*** 40** (1) 27–37