Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 4, April -2017

The application of bivariate polar plots and k-means clustering to analysis air pollution in Taoyuan, Taiwan

Ming-Hung Shu [1], Dinh-Chien Dang [1,*], Thanh-Lam Nguyen [2], Bi-Min Hsu [3] and Ky-Quang Pham[4]

[1] *Department of Industrial Engineering and Management, National Kaohsiung University of Applied Sciences, Kaohsiung 80778, Taiwan*
[2] *Office of Scientific Research, Lac Hong University, Dong Nai, Vietnam*
[3] *Department of Industrial Engineering and Management, Cheng Shiu University, Kaohsiung 83347, Taiwan*
[4] *Office of Science and Technology, Vietnam Maritime University, Hai Phong, Vietnam*

Abstract - *In this paper, we apply k-means clustering techniques directly to bivariate polar plots to identify and group similar features. Bivariate polar plots method is one of the tools in open-air packagefor source detection and characterisation. Bivariate polar plots provide an effective graphical means of discriminating different source types and characteristics. Importantly, this paper links identified clusters to known emission characteristics to confirm the inferences made in the analysis.The combination between k-means clustering and bivariate polar plots helps to avoid making arbitrary decisions about how to extract and analyse different source features. Using this approach, we initially understand and properly evaluate the level of pollution to improve air quality in Taoyuan city. This work not only provides more cases for the approach to analysis of air pollution monitoring data but also use as a reference for further research.*

Index Terms- *Openair, Air Quality, Taoyuan, Bivariate polar plots, k-means clustering.*

I. INTRODUCTION

1.1 Background

Open-air is an R package primarily developed for the analysis of air pollution measurement data but which is also of more general use in the atmospheric sciences [1].

The open-air software is freely available as an R package. Details on installing R and optional packages including open-air can be found at R Core Team (2014) and http://www.r-project.org. R will run on Microsoft Windows, Linux and Apple Mac computers. No special hardware is required to run open-air other than a standard desktop computer. Some large data sets or complex analyses may require a 64-bit platform. Ref: R Core Team (2014). R: A language and environment for statistical computing. RFoundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/ [2].

Another key strength of R is its package system. The base software, which is in itself highly capable (e.g. offering for example linear and generalized linear models, non-linear regression models, time series analysis, classical parametric and non-parametric tests, clustering and smoothing), has been greatly extended by additional functionality. Packages are available to carry out a wide range of analyses including generalized additive models, linear and nonlinear modeling, regression trees, Bayesian statistics etc. Currently there are over 2500 packages available and this number continues to grow. These packages are readily available through a global network of repositories called the Comprehensive R Archive Network (CRAN) [1].

Openair provides several functions to help users import data that ensures a format consistent for use in all other openair functions. In this study, data is imported from monitoring site and used for all figures.

There are many papers used openair functions to combine other methods for their research. Such as, David C. Carslaw and Sean D. Beevers (2012). Characterising and understanding emission sources using bivariate polar plots and k-means clustering. Iratxe Uria-Tellaetxe and David C.Carslaw (2014) Conditional bivariate probability function for source identification. etc…

1.2. Description of study area

Taoyuan *(Chinese: 桃園市)*, is a special municipality in northwestern Republic of China, neighboring New Taipei City, Hsinchu County, and Yilan County. Taoyuan District is the seat of the municipal government and that which, along with Zhongli District, forms a large metropolitan area. Taoyuan developed from a satellite city of Taipei metropolitan area to be the fourth-largest metropolitan area, and fifth-largest populated city in Taiwan. Since commuting to the Taipei metropolitan area is easy, Taoyuan has seen the fastest population growth of all cities in Taiwan.

Taoyuan is located approximately 40km (25 mi) southwest of Taipei, in northern Taiwan, and occupies 1,220km^2 (470 sq mi). It is made up of low-lying plains, interconnected mountains and plateaus. Its shape has a long and narrow southeast-to-northwest trend, with the southeast in the Xueshan Range and the far end on the shores of the Taiwan Strait [3].

"Taoyuan" means "peach garden," since the area used to have many peach trees. The city is home to many industrial parks and tech company headquarters. Taipei Taoyuan International Airport, which serves the capital, Taipei and the rest of northern Taiwan, is located in this city.

The city of Taoyuan has been elevated to special municipality status since 2014 from the original Taoyuan County. At the same time, the former county-controlled city of Taoyuan was also promoted to Taoyuan District within the new municipality [3].

The concentration of air pollution in our environment depends on both the amount of pollution produced and the rate at which pollutants disperse. This depends largely on wind (both strength and direction). In areas where the wind is very strong, pollution is dispersed and blown away. In areas where there is little or no wind, air pollution accumulates and concentrations can be high. However, local factors such as topography (hills and mountains), proximity to the coast, building height and time of the year all affect local wind conditions and can play a role in increasing air pollution levels. Two common pollutants particulates ($PM_{2.5}$, PM_{10}) and nitrogen dioxide (NO_2, NO_x) seemed to be particularly important.

The location of the study site on the country map is given in ***Fig.1***. As shown in the wind rose diagram constructed based on our data, the dominant wind directions for the site are from Southwest and South-southwest. Weak winds prevail in the northwest and northeast directions. The site is located about 23km southeast of Taoyuan International Airport.

II. METHODS

2.1 Data preparation

According to the results of previous research – apply k-means clustering techniques directly to bivariate polar plots to identify and group similar features – we applied for Taoyuancity to analysis the air quality. We ourselves collected the data in website **http://aqicn.org/city/taiwan/taoyuan/m/**5 times every day (12am, 6am, 12pm, 6pm and 11pm) from January 1st 2016 to October 31st 2016. According to that website, data was records concentration of O_3, SO_2, CO, NO_2, PM_{10} and $PM_{2.5}$. We also collected the data about temperature, relative humidity and solar radiation as well as wind speed and wind direction.

2.2 Bivariate polar plots

Bivariate polar plots show how a concentration of a species varies jointly with wind speed and wind direction in polar coordinates. The plots have proved to be useful in a range of settings e.g. to characterize airport sources and dispersion characteristics in street canyons [5][6]. Wind direction together with wind speed can be highly effective at discriminating different emission sources [5]. By using polar coordinates, the plots provide a useful graphical technique, which can provide directional information on sources as well as the wind speed dependence of concentrations

Fig.1. Map showing the location of TaoyuanCity and WindRose diagram

Briefly, bivariate polar plots are constructed in the following way. First, wind speed, wind direction and concentration data are partitioned into wind speed direction bins and the mean concentration calculated for each bin. The wind components, u and v are calculated

$$\boldsymbol{u} = \overline{\boldsymbol{u}}.\sin\left(\frac{2\pi}{\theta}\right), \boldsymbol{v} = \overline{\boldsymbol{u}}.\cos\left(\frac{2\pi}{\theta}\right)$$

With $\overline{\boldsymbol{u}}$ is the mean hourly wind speed and $\boldsymbol{\theta}$ is the mean wind direction in degrees with 90 degrees as being from the east.

The calculations above provides u, v, concentration (C) surface. While it would be possible to work with this surface data directly a better approach is to model the surface to describe the concentration as a function of the wind components u and v to extract real source features rather than noise. A flexible framework for fitting a surface is to use a Generalized Additive Model (GAM) e.g [7]. The GAM can be expressed as follow

$$\sqrt{\boldsymbol{C_i}} = \boldsymbol{\beta_0} + \boldsymbol{s}(\boldsymbol{u_i}, \boldsymbol{v_i}) + \boldsymbol{\epsilon_i}$$

Where $\boldsymbol{C_i}$ is the ith pollutant concentration, $\boldsymbol{\beta_0}$ is the overall mean of the response, $\boldsymbol{s}(\boldsymbol{u_i}, \boldsymbol{v_i})$ is the isotropic smooth function of ith value of covariate u and v, and $\boldsymbol{\epsilon_i}$ is the ith residual. Note that $\boldsymbol{C_i}$ is square-root transformed as the transformation generally produces better model diagnostics e.g. normally distributed residuals. Moreover, the smooth function used is isotropic because u and v are on the same scales. The isotropic smooth avoids the potential difficulty of smoothing two variables on different scales e.g. wind speed and direction, which introduces further complexities [5].

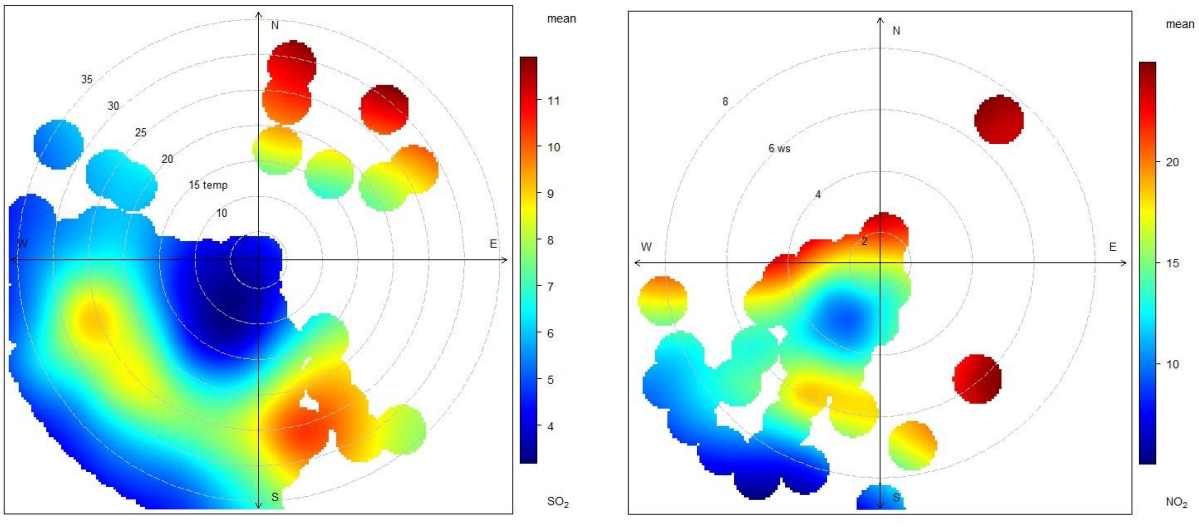
Bivariate polar plots have proved to be extremely valuable for identifying and understanding sources of air pollution [1][6]. Fig.2shows the bivariate polar plot for $SO_2$and NO_2concentrations in Taoyuan city. The plot was created by:

```
polarPlot(Taoyuan.City, pollutant = "so2", x = "temp", wd = "wd", cols = "jet", fontsize = 18)

polarPlot(Taoyuan.City, pollutant = "no2", cols = "jet", fontsize = 18)
```

Where Taoyuan.Cityrepresents the imported data 10 months period from January 2016 to October 2016 from http://aqicn.org.

In Fig.2(a) shows a bivariate polar plot for SO_2 concentrations in Taoyuan city as a function of wind direction and surface temperature. It is apparent that there is a clear dependence of SO_2 concentrations with increasing ambient temperature. The reason why concentrationsincrease with increasing temperature is that dispersing plumes from distant chimneystacks ($\approx$30 km) are brought down to ground level under unstable atmosphericconditions when thermal



(a) (b)

Fig.2. (a) Bivariate polar plot of SO_2concentrations ($\mu g\, m^{-3}$), the radial scale shows the temperature (degrees C),

(b) Bivariate polar plot of NO_2 concentrations ($\mu g\, m^{-3}$)in Taoyuan City. The radial scale shows the wind speed, which increases from the center of the plot radially out-wards.

turbulence is increased.

In Fig. 2(b) shows the bivariate polar plot of NO_2 concentration. The most obvious feature are the higher concentration of NO_2 at low wind speeds, which would typically be expected at urban background-type sites where higher concentrations results from more stable atmospheric conditions and reduced advection that exist under low wind speed conditions. However, there is also an indication of elevated concentrations of NO_2 to the south-west and north-west.

2.3k-means clustering

K-means clustering is one method in which bivariate polar plot features can be identified and grouped. The main purpose of grouping data in this way is to identify records in the original time series data by cluster to enable post-processing to better understand potential source characteristics [5]. Central to the idea of clustering data is the concept of distance i.e. some measure of similarity or dissimilarity between points. Clusters should be comprised of points separated by small distances relative to the distance between the clusters.

Given a data set $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_1}, \dots, \mathbf{x_n}\}$ in $\mathbb{R}^{\mathbf{A}}$, i.e. $\mathbf{n}$ points (vectors) each with $\mathbf{A}$ attributes (components), hard partitional algorithms divide X into K exhaustive and mutually exclusive clusters $\mathbf{C} = \{\mathbf{c_1}, \mathbf{c_1}, \dots, \mathbf{c_K}\}$, $\cup_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{C_i} = \mathbf{X}$, $\mathbf{P_i} \cap \mathbf{P_j} = \emptyset$ for $\mathbf{1} \leq \mathbf{i} \neq \mathbf{j} \leq \mathbf{K}$. These algorithms usually generate clusters by optimizing a criterion function [8]. The basic k-means algorithm for K clusters is obtained by minimising:

$$\sum_{i=1}^{K} \sum_{x_j \in P_i} \|x_j - \mu_i\|_2^2$$

Where $\|\boldsymbol{x_j} - \boldsymbol{\mu_i}\|_2$ is a chosen distance measure, $\boldsymbol{\mu_i}$ is the mean of cluster c_K.

After choosing the initial cluster centers, the other examples are assigned to the cluster center that is most similar or nearest according to the distance function. If *n* indicates the number of features, the formula for Euclidean distance as follows:

$$\boldsymbol{dist}(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{y_i})^2}$$

Where *x* and *y* are two i dimensional vectors, which have been standardized by subtracting the mean and dividing by the standard deviation. In the current case *j* is of length three i.e. the wind components u and v and the concentration C, each of which is standardized. Standardization is necessary because the wind components *u* and *v* are on different scales to C. In principle, more weight could be given to the concentration rather than the *u* and *v* components, although this would tend to identify clusters with similar concentrations but different source origins, which is not the aim in the current work [5].

III. RESULTS AND DISCUSSION

Cluster analysis has been carried out for the NO2 surface shown in Figure 2(b) for clusters between 2 and 10. For 4 clusters the complex-shaped feature of higher NO_2 concentrations is shown by cluster number 2. This feature remains as the number of clusters used increases until there are 8 clusters and the feature to the south-west is split into two groups. Similarly, the higher concentrations to the north-east identified in cluster 5 remain until cluster 9 where they are further split. Seven clusters are required before the high concentration region with low wind speed is separately identified.

There are many methods available for determining whether the different clusters represent different source types. As an example of analysing different cluster characteristics, we consider the important temporal variations in concentration by cluster. The analysis of cluster 2 in the 10 clusters solution in Fig. 3 showed that many of the temporal components of concentration differed markedly from other clusters and in particular showed potential aircraft-influenced characteristics. Cluster 1 of 10 clusters solution is also likely to show strong aircraft characteristics. There are several reasons why cluster 2 shows different characteristics to other clusters. First, the narrow directional range of the cluster maximises the signal from aircraft. Second, the relatively high wind speed associated with cluster 2 also maximises the aircraft signal. Finally, a narrow range of meteorological conditions (wind speed and direction) reduces the effect of meteorology on concentrations [5].In the example given for cluster 2 of the 10 clusters solution, further analysis could confidently provide specific information on aviation emissions as a source. For example, it would be possible to use the same cluster as a means of analysing other species such as PM_{10}, $PM_{2.5}$ or SO_2 to determine whether aviation emissions are important for those species.

REFERENCES

[1] David C. Carslaw, Karl Ropkins (2012) openair – An R package for air quality data analysis. Environmental Modelling & Software 27-28, 52-61.
[2] Iratxe Uria-Tellaetxe, David C. Carslaw (2014) Conditional bivariate probability function for source identification. Environmental Modelling & Software 59, 1-9.
[3] https://en.wikipedia.org/wiki/Taoyuan,_Taiwan. Accessed 2017 April 05
[4] https://en.wikipedia.org/wiki/Taoyuan_International_Airport. Accessed 2017 April 05
[5] Carslaw, David C., and Sean D. Beevers (2013) Characterising and understanding emission sources using bivariate polar plots and k-means clustering. Environmental Modelling & Software 40, 325-329.
[6] Carslaw D.C, Beevers S.D, Ropkins K, Bell M.C (2006) Detecting and quantifying aircraft and other on-airport contributions to ambient nitrogen oxides in the vicinity of a large international airport . Atmospheric Environment 40 (28), 5424-5434.
[7] Wood S.N (2006) Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC.
[8] M. Emre Celebi, Hassan A. Kingravi, Patricio A. Vela (2013) A comparative study of efficient initialization methods for the k-means clustering algorithm. Expert Systems with Applications 40, 200 - 210.

