ISSN 2349-0292
Impact Factor 2.675

GLOBAL JOURNAL OF ADVANCED ENGINEERING TECHNOLOGIES AND SCIENCES

USING OPEN-AIR PACKAGE FOR STATISTIC OF AIR QUALITY DATA: STUDY IN KAOHSIUNG, TAIWAN

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

This paper used some functions of open-air package to analysis air quality index (AQI) in Kaohsiung. The package includes many tools for manipulating data to understand about air pollution data. The time series concentration of PM10 and $PM_{2.5}$ are be used for evaluation and analysis of AQI. Air pollution data is collected from monitoring sites in the world. There are two principal approaches are used. First, a graphical technique using bivariate polar plots to discriminate between source types. Bivariate polar plots is a method which plotting pollutant concentration in polar coordinates showing concentration by wind speed (or another numeric variable) and direction. Second, using timeVariation and timePlot functions to analyze and assess the level of air pollution. For presented case study, we initially understand and properly evaluate the level of pollution to improve air quality.

KEYWORDS: Openair, Air Quality, Kaohsiung, Statistic.

INTRODUCTION

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].

ISSN 2349-0292
Impact Factor 2.675

Table 1 Summary of Open-air functions

Function	Purpose	Multiple pollutants	Type option
calcPno2	estimate primary NO_2 emissions ratio from monitoring data; see package help file for details and Carslaw and Beevers (2005)	no	no
calendarPlot	calendar-type view of mean values	no	no
conditionalQuantile	for model evaluation and air pollution forecasting evaluation	no	yes [2]
cutData	partition data into groups for conditioning plots and analysis	yes	yes [≥1]
importADMS	import outputs from the ADMS suite of dispersion models (McHugh et al., 1997)	NA	NA
importAURN	import hourly data from the UK air quality data archive (http://www.airquality.co.uk/data_and_statistics.php)	NA	NA
importKCL	import data from King's College London databases (http://www.londonair.org.uk/)	NA	NA
kernelExceed	bivariate kernel density estimates for exceedance statistics	no	yes [1]
linearRelation	explore linear relationships between variables in time	no	limited
MannKendall	calculate MannKendal slopes and Sen-Theil slope estimates	no	yes [2]
modStats	calculate a range of model evaluation statistics	no	yes [≥1]
percentileRose	plot multiple percentiles by wind direction	yes	yes [2]
polarAnnulus	polar annulus plot for temporal variations by wind direction	yes	yes [2]
polarFreq	alternative to wind rose/pollution rose	no	yes [2]
polarPlot	bivariate polar plot (Carslaw et al., 2006)	yes	yes [2]
pollutionRose	pollution rose	no	yes [2]
scatterPlot	traditional scatter plots with enhanced options	no	yes [2]
smoothTrend	non-parametric smooth trend estimates	yes	yes [2]
summaryPlot	summary view of a data frame with key statistics	yes	no
TaylorDiagram	taylor Diagram used for model evaluation (Taylor (2001))	no	yes [2]
timePlot	flexible time series plotting	yes	yes [1]
timeVariation	diurnal, day of week and monthly variations	yes	yes [1]
trendLevel	level plots with flexible conditioning	no	yes [2]
windRose	traditional wind rose	no	yes [2]

Fig.1. Map showing the location of Fengshan district and WindRose diagram

A summary of most open-air functions is shown in Table 1 [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…

Description of study area

Kaohsiung City (Chinese: 高雄市) is a special municipality in Taiwan. Located in southern-western Taiwan and facing the Taiwan Strait, it is by area the largest municipality, at 2,951.85 km2, and second most populous (by urban area) with a population of approximately 2.77 million. Since its start in the 17th century, Kaohsiung has grown from a small trading village, into the political, economic, transportation, manufacturing, refining, shipbuilding, and industrial center of southern Taiwan. It is a global city with sufficiency as categorized by Globalization and World Cities Research Network in 2012 [3].

Fengshan District (Chinese: 鳳山區) is a district located in southern Kaohsiung, Taiwan. Fengshan is one of the administrative centers of Kaohsiung and is home to the Chinese Military Academy. There are three military units currently located in Fengshan. Both Chinese Military Academy and R.O.C. Army Infantry School were migrated from mainland China and re-established here in 1950. Chung Cheng Armed Forces Preparatory School was established in 1976. These three units used to be the main economic driving force, but their importance seems to diminish gradually as Fengshan has established itself as a conjunction between Pingtung City and Kaohsiung. Although there are several industrial zones at the rim or outskirt of the city, the major life style in Fengshan seems to be very residential [4].

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 (PM2.5, PM10) and nitrogen dioxide (NO2, NOx) 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 South-southeast, South and South-southwest. Weak winds prevail in the northwest and northeast directions. The site is located about 7km northwest of the Kaohsiung Port, approximately 5km south of Kaohsiung International Airport.

METHODS

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 Kaohsiung City (Fengshan District) to analysis the air quality. We ourselves collected the data in website **http://aqicn.org/city/taiwan/fongshan/m/** 5 times every day (12am, 6am, 12pm, 6pm and 11pm) from November 1st 2015. According to that website, data was records concentration of O3, SO2, CO, NO2, PM10 and PM2.5. We also collected the data about temperature, relative humidity and solar radiation as well as wind speed and wind direction.

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

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

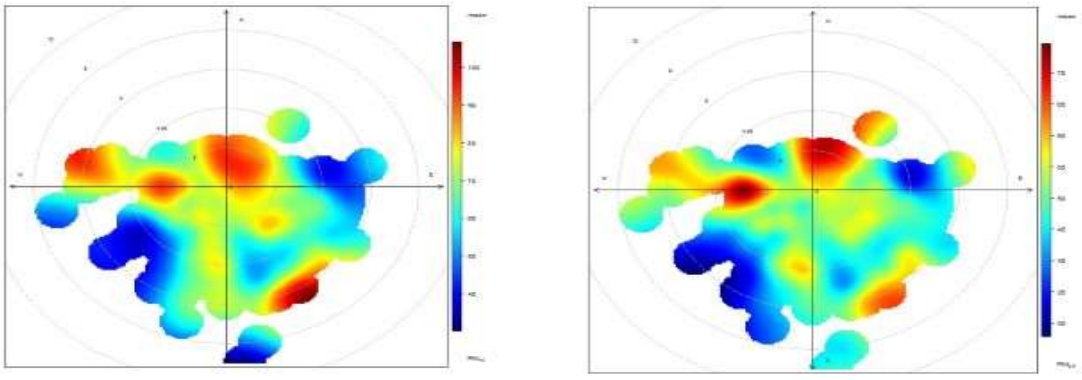
$$\boldsymbol{u} = \overline{\boldsymbol{u}}.\sin\left(\frac{2\pi}{\boldsymbol{\theta}}\right), \boldsymbol{v} = \overline{\boldsymbol{u}}.\cos\left(\frac{2\pi}{\boldsymbol{\theta}}\right)$$

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.

ISSN 2349-0292
Impact Factor 2.675

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

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



(a) (b)

Fig.2. (a) Bivariate polar plot of PM10 concentrations $(\mu g\ m^{-3})$
(b) Bivariate polar plot of PM2.5 concentrations $(\mu g\ m^{-3})$ ***at Fengshan district.***

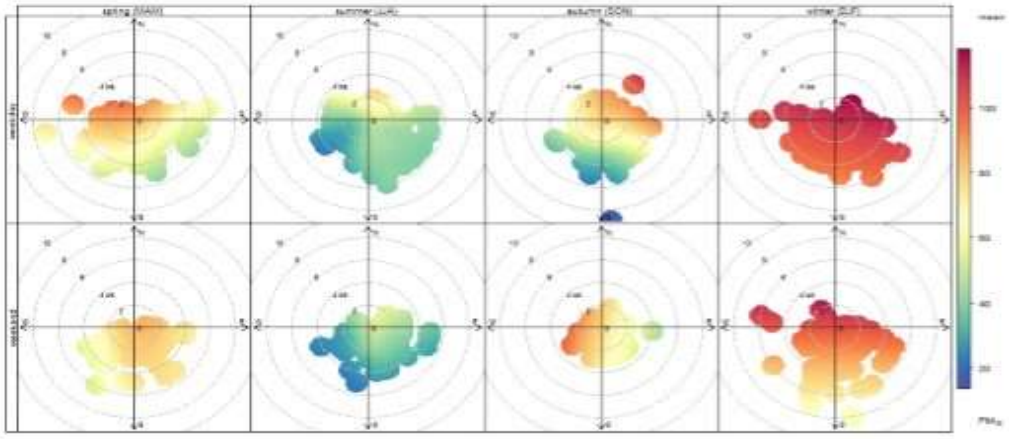


Fig.3. Bivariate polar plot of PM10 concentrations $(\mu g\ m^{-3})$ ***at Fengshan district. The center of each plot represents a wind speed of zero. The plot has been 'conditioned' by two variables: season and whether the day is weekend or weekday.***

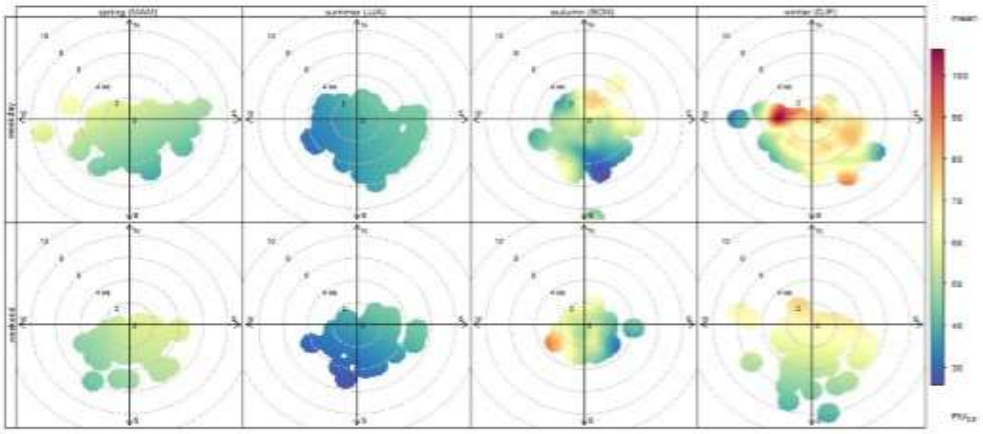


Fig.4. Bivariate polar plot of PM2.5 concentrations $(\mu g\ m^{-3})$ ***at Fengshan district. The center of each plot represents a wind speed of zero. The plot has been 'conditioned' by two variables: season and whether the day is weekend or weekday.***

ISSN 2349-0292
Impact Factor 2.675

Where C_i is the ith pollutant concentration, β_0 is the overall mean of the response, $s(u_i, v_i)$ is the isotropic smooth function of ith value of covariate u and v, and ϵ_i is the ith residual. Note that 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. 2 shows the bivariate polar plot for PM_{10} and $PM_{2.5}$ concentrations at Fengshan. The plot was created by:

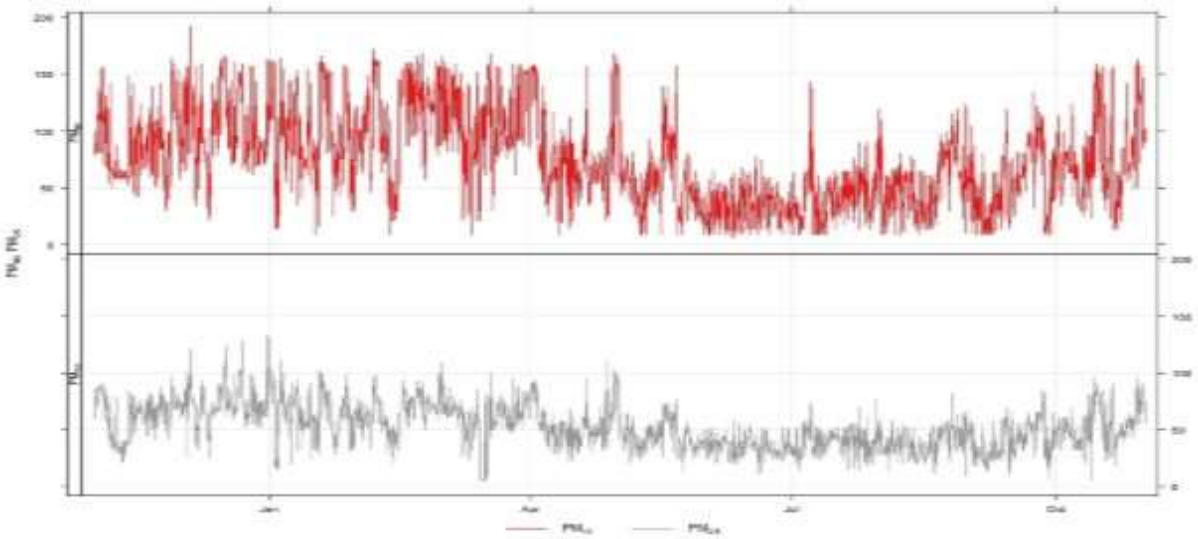


Fig.5. Plot time series of PM10 and PM2.5 concentrations $(\mu g\ m^{-3})$ ***at Fengshan district***

polarPlot(Fengshan.District, pollutant = "pm10", cols = "jet", fonsize = 18)

polarPlot(Fengshan.District, pollutant = "pm25", cols = "jet", fonsize = 18)

Where Fengshan.District represents the imported data one year period from November 2015 to October 2016 from http://aqicn.org.

In Fig. 2, the most obvious feature are the higher concentration of PM_{10} and $PM_{2.5}$ at low wind speed, which would typically be expected at urban background. However, there is also an indication of elevated concentration of PM_{10} and $PM_{2.5}$ to the south-west and south-east. These features would be good to investigate further.

In openair there are several in built ways of conditioning data; two of which were used in the polarPlot call "season" and "weekend. In Fig. 3 and Fig. 4 we demonstrated type = c("season" and "weekend") option of the function. The plot itself was produced by:

polarPlot(Fengshan.District, pollutant = "pm10", type = c("season", "weekend"), k = 70)

polarPlot(Fengshan.District, pollutant = "pm25", type = c("season", "weekend"), k = 70)

We clearly seen that, the concentrations tend to be higher during the working week than at weekends. The weekend-weekday differences can often reveal information about source activities. The first characteristic to note is that there is a wind speed and direction dependence on the concentration of PM_{10} and $PM_{2.5}$. Concentrations are generally higher for higher wind speed conditions in spring and winter shown by the elevated PM_{10} and $PM_{2.5}$ concentrations to the south-east. However, the patterns of concentration are different. For high wind speeds (>about 5 m s^{-1}) there is more evidence of high PM_{10} and $PM_{2.5}$ concentrations during the wintertime than spring. This type of analysis can help narrow-down the number of likely source origins and characteristics, but does not always provide a definitive answer. For this reason, it is important to analyse the data in other ways. A subset of wind speed and direction conditions can be used to isolate various features for analysis.

ISSN 2349-0292
Impact Factor 2.675

TimeVariation and timePlots

The *timeVariation* function has been extended to allow users to consider the median and quantiles rather than only the mean and 95% confidence interval in the mean. There is a new option statistic that can either be 'mean' or 'median'. If the statistic is 'median' then the median line is plotted together with the 5/95 and 25/75th quantiles are plotted. Users can control the confidence intervals with conf.int. For example, conf.int = c(0.25, 0.99) will show the median, 25/75th and 1/99th quantile values. The statistic = "median" option is therefore very useful for showing how the data are distributed - somewhat similar to a box and whisker plot. Note that it is expected that only one pollutant should be shown when statistic = "median" is used due to potential over-plotting; although the function will display several species of required. The result is shown in Fig. 6 for PM_{10} and $PM_{2.5}$ concentrations by using:

timeVariation(Fengshan.District, pollutant = c("pm10", "pm25"))

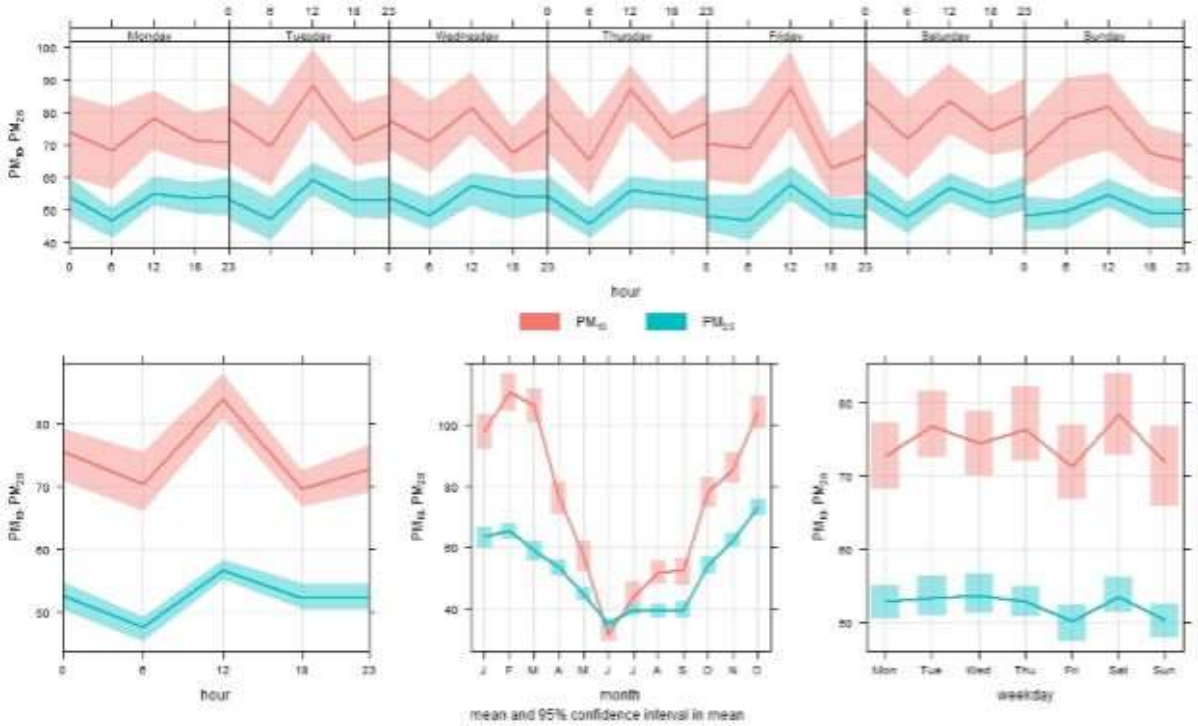


Fig.6. TimeVariation of PM10 and PM2.5 concentrations $(\mu g\ m^{-3})$ ***at Fengshan district.***

The *timePlot* is the basic time series plotting function in openair. Its purpose is to make it quick and easy to plot time series for pollutants and other variables. The other purpose is to plot potentially many variables together in as compact a way as possible. The function is flexible enough to plot more than one variable at once. If more than one variable is chosen plots it can either show all variables on the same plot (with different line types) on the same scale, or (if group = FALSE) each variable in its own panels with its own scale. The general preference is not to plot two variables on the same graph with two different y-scales. It can be misleading to do so and difficult with more than two variables. If there is in interest in plotting several variables together that have very different scales, then it can be useful to normalise the data first, which can be down be setting normalise = TRUE. This option ensures that each variable is divided by its mean and makes it easy to plot two or more variables on the same plot - generally with group = TRUE. The user has fine control over the choice of colours, line width and line types used. This is useful for example, to emphasise a particular variable with a specific line type/colour/width. timePlot works very well with select by Date, which is used for selecting particular date ranges quickly and easily. By default, plots are shown with a colour key at the bottom and in the case of multiple pollutants or sites, strips on the left of each plot [8]. Sometimes this may be overkill and the user can opt to remove the key and/or the strip by setting key and/or strip to FALSE. In this case, plot was created by:
timePlot(Fengshan.District, pollutant = c("pm10", "pm25"))

The result shown in Fig. 5; it seems that in wintertime tend to be higher concentrations than summertime at both PM_{10} and $PM_{2.5}$. To clarify the level of air pollution at Fengshan, we used timeAverage to calculate the average value, the result of this analysis is shown in Fig. 7 by using:

timeAverage(Fengshan.District, avg.time = "year", statistic = "mean")

ISSN 2349-0292
Impact Factor 2.675

This function to flexibly aggregate or expand data frames by different time periods, calculating vector-averaged wind direction where appropriate. The averaged periods can also take account of data capture rates.

```
        date     pm10    pm25       o3      no2      so2       co      uvi     temp
      <dttm>    <dbl>   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
1 2015-01-01 94.96066 67.6459 25.09836 25.83279 7.918033 7.645902 1.760656 21.63279
2 2016-01-01 70.43738 49.3259 19.67520 16.61902 6.513443 6.019016 2.369836 24.61902
# ... with 5 more variables: pres <dbl>, hum <dbl>, ws <dbl>, wd <dbl>, X <dbl>
```

Fig.6. TimeAverage for Fengshan data frames

The average concentration of PM_{10} is $\boldsymbol{m_{pm10}} = \frac{\sum_{i=1}^{n}(m_1, m_2, \ldots m_n)}{n} = \mathbf{82.69902}$ μg / m^3

The average concentration of $PM_{2.5}$ is $\boldsymbol{M_{pm2.5}} = \frac{\sum_{i=1}^{n}(M_1, M_2, \ldots M_n)}{n} = \mathbf{58.4859}$ μg / m^3

These results compare with Air Quality Standards of Taiwan from Taiwan Air Quality Monitoring Network (65 μg / m^3 for PM_{10} and 15 μg / m^3 for $PM_{2.5}$ annual average), we can clearly see that the level of Air Pollution at Fengshan in the period November 2015 to October 2016: "***Moderate".*** Impact on Human Health: "*Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution".*

CONCLUSIONS

There are many methods exist to analysis air quality but in this study has combined two approaches commonly used for source identification. The bivariate polar plot provides additional information on how sources disperse and timeVariation, timePlot give an easy way to plot time series for pollutants and other variables. Combining these two methods provides detailed plot for analyzing data and comparing with Air Quality Standards of Taiwan. Based on the comparison results, we assess the level of air pollution at Fengshan.

The approach can therefore provide a more comprehensive understanding of a very wide range of sources affecting a particular monitoring site. There are many potential uses of these methods in this paper. This is the foundation for our further research.

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/Kaohsiung. Accessed 2017 March 22
[4] https://en.wikipedia.org/wiki/Fongshan_District. Accessed 2017 March 22
[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] https://www.rdocumentation.org/packages/openair/versions/0.3-8/topics/timePlot. Accessed 2017 April 02