



# Spatial identification and temporal prediction of air pollution sources using conditional bivariate probability function and time series signature

Omar F. Althuwaynee<sup>1</sup> · Badal Pokharel<sup>2</sup> · Ali Aydda<sup>3</sup> · Abdul-Lateef Balogun<sup>4</sup> · Sang-Wan Kim<sup>1</sup> · Hyuck-Jin Park<sup>1</sup>

Received: 27 January 2020 / Revised: 17 September 2020 / Accepted: 18 September 2020 / Published online: 6 November 2020  
© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

## Abstract

Accurate identification of distant, large, and frequent sources of emission in cities is a complex procedure due to the presence of large-sized pollutants and the existence of many land use types. This study aims to simplify and optimize the visualization mechanism of long time-series of air pollution data, particularly for urban areas, which is naturally correlated in time and spatially complicated to analyze. Also, we elaborate different sources of pollution that were hitherto undetectable using ordinary plot models by leveraging recent advances in ensemble statistical approaches. The high performing conditional bivariate probability function (CBPF) and time-series signature were integrated within the R programming environment to facilitate the study's analysis. Hourly air pollution data for the period between 2007 to 2016 is collected using four air quality stations, (ca0016, ca0058, ca0054, and ca0025), situated in highly urbanized locations that are characterized by complex land use and high pollution emitting activities. A conditional bivariate probability function (CBPF) was used to analyze the data, utilizing pollutant concentration values such as Sulfur dioxide (SO<sub>2</sub>), Nitrogen oxides (NO<sub>2</sub>), Carbon monoxide (CO) and Particulate Matter (PM<sub>10</sub>) as a third variable plotted on the radial axis, with wind direction and wind speed variables. Generalized linear model (GLM) and sensitivity analysis are applied to verify and visualize the relationship between Air Pollution Index (API) of PM<sub>10</sub> and other significant pollutants of GML outputs based on quantile values. To address potential future challenges, we forecast 3 months PM<sub>10</sub> values using a Time Series Signature statistical algorithm with time functions and validated the outcome in the 4 stations. Analysis of results reveals that sources emitting PM<sub>10</sub> have similar activities producing other pollutants (SO<sub>2</sub>, CO, and NO<sub>2</sub>). Therefore, these pollutants can be detected by cross selection between the pollution sources in the affected city. The directional results of CBPF plot indicate that ca0058 and ca0054 enable easier detection of pollutants' sources in comparison to ca0016 and ca0025 due to being located on the edge of industrial areas. This study's CBPF technique and time series signature analysis' outcomes are promising, successfully elaborating different sources of pollution that were hitherto undetectable using ordinary plot models and thus contribute to existing air quality assessment and enhancement mechanisms.

**Supplementary information** The online version of this article (<https://doi.org/10.1038/s41370-020-00271-8>) contains supplementary material, which is available to authorized users.

✉ Abdul-Lateef Balogun  
alateef.babatunde@utp.edu.my

<sup>1</sup> Department of Energy and Mineral Resources Engineering, Sejong University, 209 Neudong-ro Gwangjin-gu, Seoul 05006, Republic of Korea

<sup>2</sup> School of Civil and Environmental Engineering, University of New South Wales, Sydney, NSW, Australia

## Introduction

Air pollution represents the condition of air pollutants in the atmosphere at high enough concentrations within serious or above normal ambient levels, and it often has a dominant effect on the quality of life [1]. Normal and above normal

<sup>3</sup> Department of Geology, Faculty of Sciences, Ibn Zohr University, B.P 8106, 80000 Agadir, Morocco

<sup>4</sup> Geospatial Analysis and Modelling (GAM) Research Laboratory, Department of Civil and Environmental Engineering, Universiti Teknologi PETRONAS (UTP), Seri Iskandar, 32610 Perak, Malaysia

levels are mainly measured with air pollutant index (API) which is defined in terms of the effects of pollutants on human health [2, 3]. Generally, modeling the receptor to identify and characterize emission sources is typified by difficulties like incomplete information of the sources and the complexities of modeling boundary layer processes. Therefore, identifying effective modeling tools for receptor data record is essential in detecting local, distant, and diverse emission sources in ambient air pollution management studies [4, 5].

The methods used to identify, quantify, characterize and predict pollutants' emission sources in literature can be grouped into two main classes—receptor models (RMs) [6, 7] and dispersion models (DMs) [8]. RMs are statistical approaches used for identification and quantification of pollutants' sources at receptor site. DMs are mathematical simulation approaches used to predict pollutants' dispersion in the atmosphere. In a comprehensive review of RM models, Pokorná et al. [9] identified the following prevalent methods: chemical mass balance, multivariate methods (including target transformation factor analysis, unmix, positive matrix factorization, and constrained models), other complex models (e.g., multiple sample type data and time synchronization model), multiway data (e.g., spatially distributed data and size–composition–time data), ensemble methods, methods using local wind data (e.g., conditional probability function, nonparametric regression, nonparametric wind regression (NWR) and sustained wind incidence method), methods incorporating back trajectories (e.g., potential source contribution function), simplified quantitative trajectory bias analysis and Future directions.

Holmes and Morawska [10] presented an overview of DMs applied to the dispersion of particulate matters (PM). These models were separated into four main classes: box models, Gaussian models, Lagrangian models and computational fluid dynamic models. Generally, the large number of pollutants around the monitoring stations makes the identification process more difficult [1]. Moreover, the distant locations of major resources reduce the detection sensitivity while the minor elements that are locally available close to monitoring stations also impede clear detection. To overcome these limitations, some studies have developed combined models in order to detect the pollutants' emission sources with high precision. Qin and Oduyemi [11] combined the receptor model (positive matrix factorization) and dispersion model (Gaussian plume dispersion model) to identify aerosol sources and estimate source contributions to air pollution in Dundee, UK. The dispersion model identified sources that were undetected by the receptor model. Kim and Hopke [12] combined positive matrix factorization (PMF) and conditional probability function (CPF) to identify sources of PM in Spokane, an

arid city in the Northwestern part of U.S. This approach enhanced the identification of emission sources.

The models which incorporate meteorological data (e.g., clear wind patterns) produce the most reliable results [10], particularly when integrated with CPF for apportionment of pollutants' emission sources. CPF is an effective statistical technique for providing directional information for pollutants' sources as well as isolate specific source types using wind direction, wind speed and concentration of specific pollutants [13, 14]. CPF analysis with wind speed, direction and time of day can help isolate specific source types for further analysis [13]. Using CPF, Henry et al. [15] concluded that large sources of pollution might have a minor contribution to concentrations of SO<sub>2</sub>. Kim and Hopke [12] investigated the advantages of using CPF approach over the non-parametric regression approach, while Bae et al. [16] applied the CPF method to identify the direction of major sources of pollutants in New York State. Furthermore, optimizing data collected by few monitoring stations in complex study areas is crucial. In this context, [17] developed a new approach which combines bivariate polar plots with a CPF to detect and characterize source contributions. This approach, called the conditional bivariate probability function (CBPF), is an extension of the commonly used CPF. The CBPF offers more information on the type of sources being identified by highlighting important dispersion characteristics [17].

Other studies have identified pollutants' sources from the time series of PM concentrations using RMs or artificial neural network (ANN) modeling methods. Marmur et al. [18] used RMs to determine pollutants' sources emission from the time series of PM concentrations in the southeastern United States. The study concluded that pollutants' sources have significant temporal variation with the likelihood of shortcomings with respect to spatial representation. Elangasinghe et al. [19] developed a new approach based on ANN, bivariate concentration-wind speed-wind direction relationships and k-means clustering for identifying source performance signals from the time series of ambient PM concentrations in a coastal area in New Zealand. They emphasized that the identification of periodic and aperiodic source performance signals is a useful technique to enhance the prediction accuracy of hourly average concentrations.

Most recently, Jeričević et al. [4] used a combination of different source apportionment methods to determine the major pollutant sources of PM, Hydrogen sulfide (H<sub>2</sub>S), NO<sub>2</sub> and SO<sub>2</sub> in a complex urban area in Croatia (Slavonski Brod). The methods included chemical PM speciation, time series of pollutant concentrations, PMF and CBPF. The study confirmed the importance of the CBPF method for the identification of pollutants' sources and noted a larger uncertainty of CBPF method, which is related

to the representation of peak concentrations transported with wind speeds higher than 8 m/s. Similarly, determination of the extent of influence of the pollutants' emission in ambient atmosphere and its impact on the urban areas are important for environmental management planning.

In addition to these approaches, Geographic Information Systems (GIS) modeling techniques have the capability to support pollutants' source detection [20, 21]. Wang et al. [22] proved that GIS is useful for identifying and mapping the potential sources of pollutants. Although a number of researches have integrated these techniques for optimal performance, many of the studies focus on developed cities, mostly in America and Europe. Attempts to implement such advanced statistical and programming algorithms in developing cities, particularly in Asia and Africa, are limited. Mukherjee and Agrawal [23] combined CBPF, land use regression (LUR) and trajectory statistical models (TSM) to detect the sources of PM and their relationships with gaseous pollutants (NO<sub>2</sub>, SO<sub>2</sub>, O<sub>3</sub>, and CO) and meteorological parameters (wind direction, wind speed, relative humidity, and temperature) in middle Indo-Gangetic plain, India. The study concluded that meteorological parameters aided increase in PM concentration than associated gaseous pollutants. Further, the traffic in the northwestern part of India was identified as the most probable source of PM. A combination of CBPF and statistical models was also used to investigate the effect of meteorological parameters on PM concentration in four cities in the Yangtze River Delta, China [24]. Rana and Khan [25] used CBPF and Concentration Weighted Trajectory (CWT) methods to assess the impact of congenial meteorology on PM concentrations in six urban areas in Bangladesh while Kang et al. [26] confirmed a negative relationship between PM concentrations and meteorological parameters in Nanjing city, China. Also, Jain et al. [27] indicated that the traffic emission and biomass burning contribute greatly to PM concentration in Delhi, India, especially during winter. Despite these studies, implementation of the integrated spatial- CBPF technique in Asia remains relatively limited. Therefore, it is imperative to extend these techniques to cities in the high-risk developing region (UN News, 2012), in order to ascertain their suitability for adoption therein.

Based on the foregoing, this current research aims to determine the source direction of the major pollutants in the study area and investigate possible relationships between these pollutants. We also aim to forecast potential emission of pollutants in the near future. Therefore, we present an integrated approach based on CBPF, sensitivity analysis, time series signature and GIS to identify, characterize, and model pollutants' emission sources in a complex region in Malaysia (Kuala Lumpur city and surroundings areas). The selected pollutants are Sulfur dioxide (SO<sub>2</sub>), Nitrogen oxides (NO<sub>2</sub>), Carbon monoxide (CO), and Particulate

Matter (PM<sub>10</sub>). Also, the correlation between PM<sub>10</sub> and other pollutants was explored using generalized linear model (GLM).

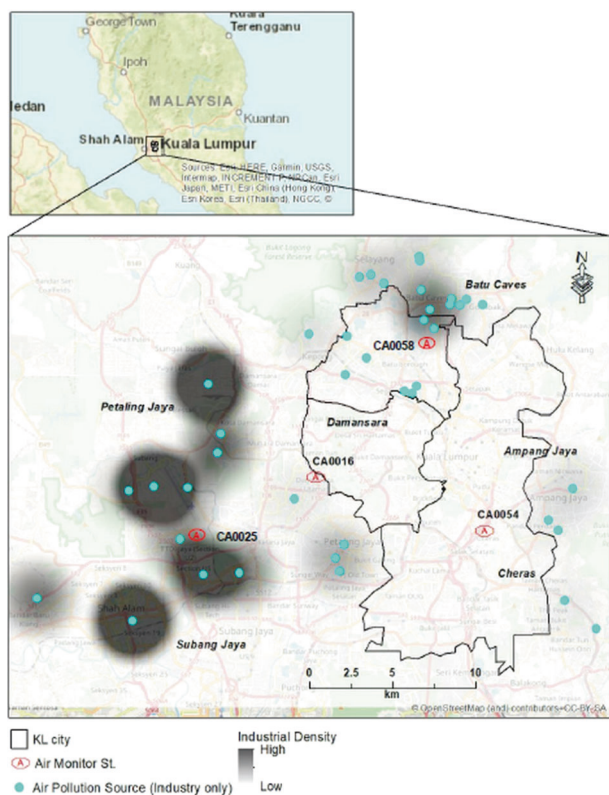
## Study area

Kuala Lumpur metropolis is the federal and economic capital of Malaysia and is surrounded by the state of Selangor. It has six strategic zones with a total area of 242.8 km<sup>2</sup> [28]. Air pollution due to airborne particulate matter (PM<sub>10</sub>) is an environmental issue in the Southeast Asia (SEA) region, particularly Indonesia, Singapore, Brunei, and Malaysia. Because of its rapid urbanization, industrialization and increased vehicular traffic, Kuala Lumpur has witnessed unprecedented infrastructural development, causing alteration of its landscape and contamination of the environment [29]. Accelerated urbanization, continuous industrial emission, vehicle emission and re-suspension of soil dust have increased the volume of PM<sub>10</sub> mix in Malaysia's atmosphere and exacerbated toxicity [30]. Smoke-haze episodes are regular occurrences in Kuala Lumpur, which cause frequent emissions of hazardous particles and gases into the surrounding atmosphere [31]. Particle pollution from urban activities and transboundary inputs are major sources of PM<sub>10</sub> pollution in the city [32]. In its study of risks posed by major air pollutants in Kuala Lumpur, Tajudin [33] noted hospitalizations due to cardiovascular diseases in relation to exposure to NO<sub>2</sub> and Ozone (O<sub>3</sub>). This sustained emission of pollutants from multiple sources make Kuala Lumpur an ideal study area as shown in Fig. 1. The figure also shows the land use and relative locations of the air quality stations in the study area.

## Methodology

### Data used

Four stations were used to collect the data from 2007 to 2016 (Fig. 1). These include station Ca0025 located at Taman Tun Dr Ismail Jaya, Shah Alam. The area is surrounded by a complicated industrial land use, the Sepang Airport and heavy industries in the southern zone. Station Ca0054, which is the source of 2009 to 2017 hourly data, is located at Seri Permaisuri, KL and is surrounded by a number of industries. It is near the city center which experiences heavy traffic. Station Ca0016 is located at Petaling Jaya, western part of KL with industries situated on the southern and western parts. Station Ca0058 is in the northern part of KL, specifically at Batu Muda, and surrounded by some industries on the north and north eastern part. The stations are surrounded



**Fig. 1 Study area.** The study area showing industrial land use locations, density, and locations of the air quality monitoring stations.

by complex industrial areas and were selected for this study because of ease of detection by distant resources due to surrounding flat terrain. Thus, wind flows are not affected by issues related to terrain.

We collected data from approximately 40 sites using land use map provided by the Federal Department of Town and Country Planning (PLANMalaysia), and digitized locations using Google earth service. The selected data comprise of industrial zones only as it is the most likely to emit the biggest share of pollutants. Roads and other landuse were ignored due to data limitation. Figure 1 shows distribution of the air gauges. CA0058 and CA0025 are surrounded by multiple sources of pollutants, while CA0054 and CA0016 are situated in neighborhoods with limited emission of pollutants. Sources of pollution in the study area include food and beverage companies, automobile maintenance and services, chemical industry, and a complex road network. ESRI's ArcGIS software was used to map the study area and relative locations of the roads, main emission sources and industrial areas to the monitoring site.

The maximum values of 1-hour concentration of SO<sub>2</sub>, NO<sub>2</sub>, CO, and PM<sub>10</sub> are found in stations CA0016 and CA0025 (Table 1). This can be attributed to the location of these stations within the city center and in close proximity

to highly dense industrial zones. All pollutants were measured in volumetric units' parts per million by volume (ppm), except PM<sub>10</sub> which was measured in Gravimetric unit (µg/m<sup>3</sup>). Wind speed was measured in meter/second (m/sec), with the Max. value (26.5) registered at CA0054 and Min. value (0.7) registered at CA0016 and CA0025. The duration of all the data measurements collected from the site and analyzed in this study was from 1st January 2006 to 31st December 2016, except data at station CA0054, which covered the period from 2007 to 2016.

### Missing data imputation

The acquired data records were examined to detect missing data that need to be imputed before proceeding as recommended by Shah et al. [34]. Using 4 stations encircling the study area, we collected daily record covering a period of 10 years. The largest amount of missing data, represented by wind direction, reached a maximum of 28% in CA0025, and the maximum missing wind speed data was 20% in CA0054 as shown in Fig. 2. The other variables (pollutants) vary but do not exceed 10%. Generally, the missing data is not very high, and the data imputation technique is applicable to cover this source of uncertainty. Figure 3 shows the data summary after data imputation using predictive mean matching (PMM) in MICE package of R programming environment.

It presents the mean concentration values of the pollutants combined with wind direction, plotted on radial axis, and indicates the most critical directions of the dominant pollutant sources around the stations. The meteorological data and air condition are also summarized in the summary plot. Since the land use data does not contain detailed information on the industrial land use, it is difficult to compare the research outcomes with acceptable or dominant range of pollution concentration for each industry (Table 2).

### Conditional bivariate probability function

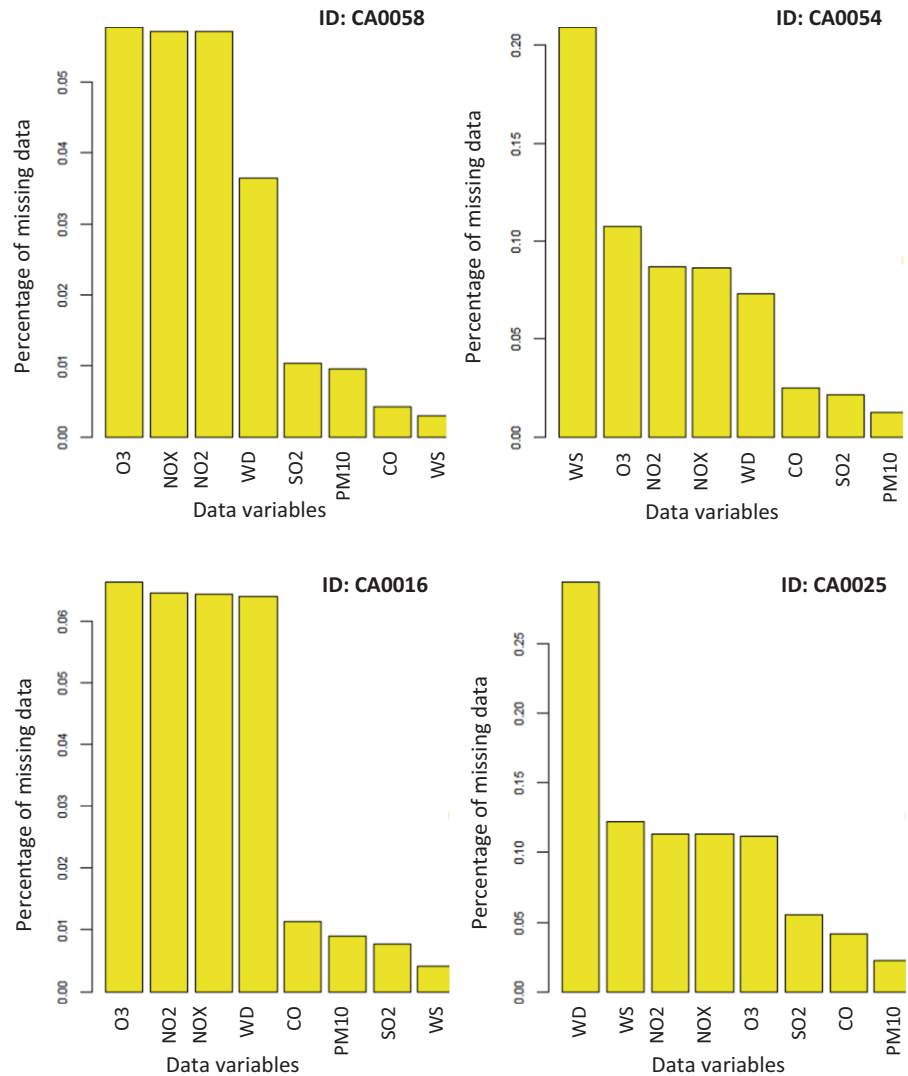
For a given wind sector, CPF estimates the probability that the measured concentration goes beyond the fixed threshold [14]. CPF was originally used to show the wind directions that dominate a specified high pollutant concentration, indicating the probability of such concentrations occurring due to wind direction [14].

$$CPF_{\Delta\theta} = \frac{m_{\Delta\theta|C \geq x}}{n_{\Delta\theta}} \quad (1)$$

Where  $m_{\Delta\theta}$  = number of samples in the wind sector  $\theta$ ,  $C$  = pollutants concentration,  $x$  = threshold value of high percentile of concentration e.g., 95th,  $n_{\Delta\theta}$  = total number of samples from wind sector  $\Delta\theta$

**Table 1** Minimum and maximum concentration of pollutants at the four stations.

Pollutant	Max. concentration	Station(s) with Max.	Min. concentration	Station(s) with Min.	Unit
SO <sub>2</sub>	0.025	CA0016	0	All	Parts per million by volume (ppm)
NO <sub>x</sub>	0.361	CA0016	0	All	Parts per million by volume (ppm)
CO	5.670	CA0025, CA0016	0	CA0025, CA0016, CA0054	Parts per million by volume (ppm)
PM <sub>10</sub>	426	CA0025	9	CA0058, CA0025	Gravimetric unit (μg/m <sup>3</sup> ),

**Fig. 2 Missing data.** Percentage of missing data for all variables in the 4 stations.


Conditional bivariate probability function (CBPF) combines CPF with the wind speed, which is considered the third variable in the equation. It assigns the pollutants' concentration to the cells that are defined by the wind direction and wind speed, having a higher reliability than the conventional methods that consider only wind direction [14]. CBPF is appropriate in areas characterized by high source complexity, with the potential of identifying more pollutants' sources in comparison with currently

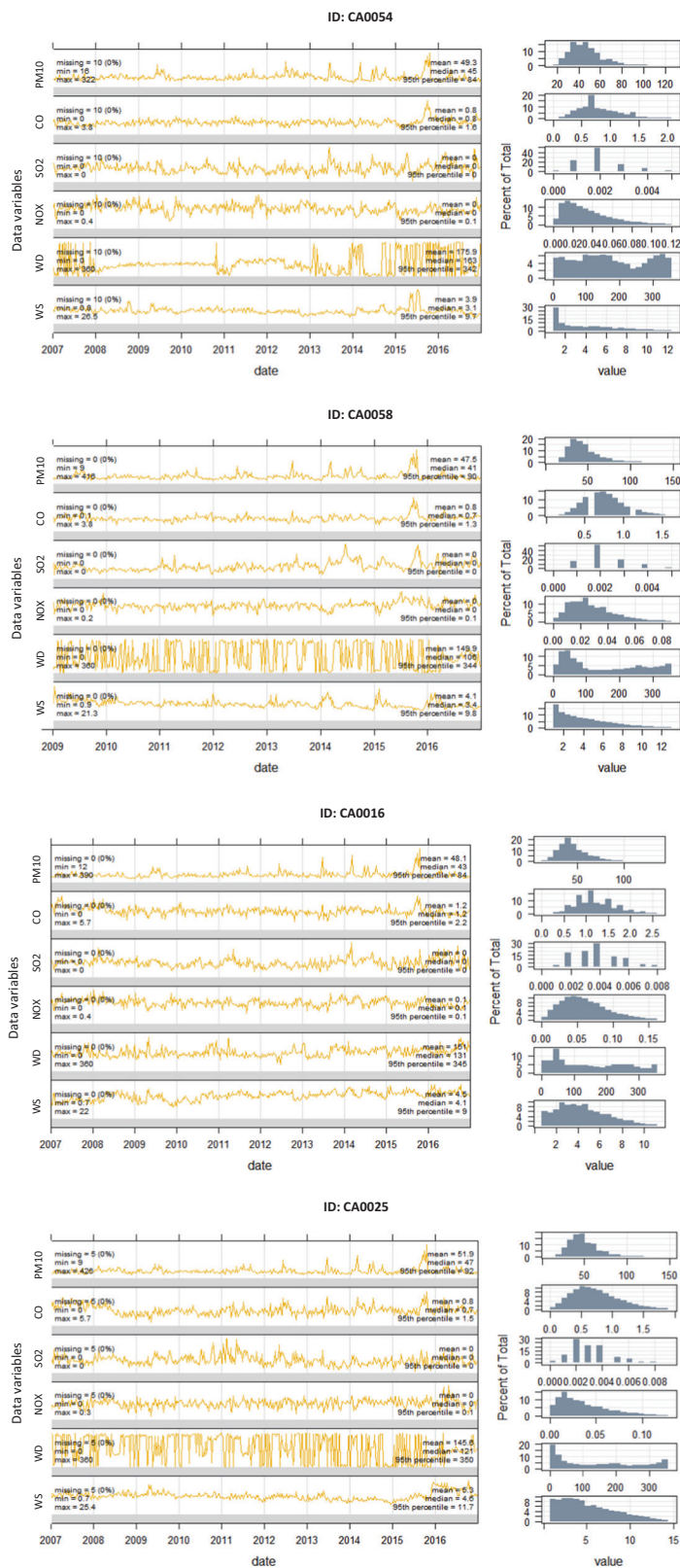
used techniques such as the CPF [17].

$$\text{Mathematically, } CBPF_{\Delta\theta, \Delta u} = \frac{m_{\Delta\theta, \Delta u | C \geq x}}{n_{\Delta\theta, \Delta u}} \quad (2)$$

Where  $m_{\Delta\theta, \Delta u}$  = number of samples in the wind sector  $\Delta\theta$ ,  $\Delta u$  = wind speed interval;  $C$  = pollutants concentration,  $x$  = threshold value of high percentile of concentration e.g., 95th,  $n_{\Delta\theta, \Delta u}$  = total number of samples in that wind direction-speed interval [17]



**Fig. 3 Summary plot.**  
 Summary plot and statistical summary of pollutants' concentration variables at the 4 stations.



**Table 2** Major pollutants' concentration in each station (Bolded values refer to maximum value for each station).

	CA0058	CA0054	CA0016	CA0025
SO <sub>2</sub> (ppm)	<b>0.198382</b>	0.110015	0.160800	0.099133
NO <sub>2</sub> (ppm)	0.171824	0.220054	0.103706	<b>0.247021</b>
CO (ppm)	<b>0.419687</b>	0.261765	0.248654	0.392757

The R programming package "OpenAir", which shows durability and efficiency in terms of data size and processing time, was used to implement the CBPF technique.

### Generalized linear model (GLM)

The generalized linear model (GLM) was generated for determining the correlation between PM<sub>10</sub> with other pollutant sources. Linear regression serves as a workhorse of statistics but cannot handle some types of complex data. A GLM expands upon linear regression to include non-normal distributions including binomial and count data.

Logistic regression is used to predict a class i.e., a probability, and it can predict a binary outcome accurately. Imagine you want to predict whether a loan is denied/accepted based on many attributes. The logistic regression is of the form  $0/1$ .  $y = 0$  if a loan is rejected,  $y = 1$  if accepted.

A logistic regression model differs from linear regression model in two ways.

- First, the logistic regression accepts only dichotomous (binary) input as a dependent variable (i.e., a vector of 0 and 1).
- Second, the outcome is measured by the following probabilistic link function called sigmoid due to its S-shape.:

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

The output of the function is always between 0 and 1

The sigmoid function returns values from 0 to 1. For the classification task, we need a discrete output of 0 or 1.

To convert a continuous flow into discrete value, we can set a decision bound at 0.5. All values above this threshold are classified as 1

$$\hat{y} = \begin{cases} 0 & \text{if } \hat{p} < .5 \\ 1 & \text{if } \hat{p} \geq .5 \end{cases}$$

### timetk: a tool kit for working with time series in R

For the future prediction of PM<sub>10</sub>, testing data is used in time series to analyze the residual and forecast PM<sub>10</sub> occurrence

for 3 months. Residual plot for the stations is prepared taking data for a year (1-1-2016 to 31-12-2016).

The time series signature is a collection of useful features that describe the time series index of a time-based data set. It entails a wealth of features that can be used to forecast time series containing patterns. In this vignette, the user implements advanced statistical analysis to predict future outcomes in a time-based data set. The *timetk* package, comprising tools to get the time series index, signature, and summary from time series objects and time-based tables, enables a user to work with time series objects more easily in R. The robust platform also supports inspecting and manipulating the time-based index, expanding the time features for data mining and machine learning, and converting time-based objects to and from the many time series classes.

## Results and discussion

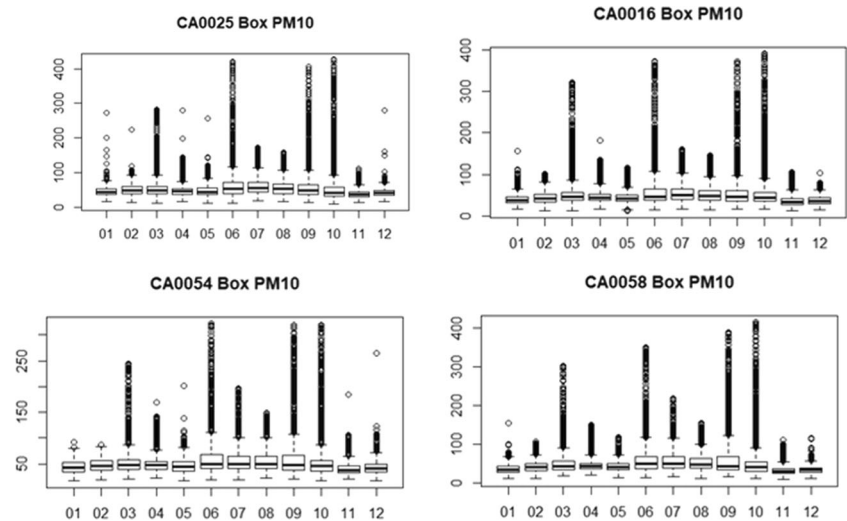
### Pollutants' concentration and sources

Figure 4 shows box plots of median concentration of PM<sub>10</sub> (monthly average) while Fig. 5 presents Whisker plots of PM<sub>10</sub> API concentration for the period of study. Box and whisker plots are effective tools for summarizing data over different time scales, including day of the week, day of the year, and month of the year. These show variations in pollutant concentrations over different time scales, providing valuable clues on pollutants' sources and their respective levels of significance. From Fig. 4, it is observed that the concentration of PM<sub>10</sub> at each of the station's peaks in March, June, May, September, and October. This trend occurs at all the four stations. However, the highest monthly concentrations are observed in August in station CA0025, presumably due to its proximity to Subang airport, which has significant chemical pollution episodes. In its investigation of the environmental and health impacts of proximity to airport infrastructure in Malaysia, Sahrir et al. [35] reported that aviation services, particularly airplane emissions, produce numerous hazardous materials such as CO, O<sub>3</sub>, SO<sub>2</sub>, NO<sub>2</sub>, and PM, confirming the likelihood of high emissions in areas close to the Subang airport.

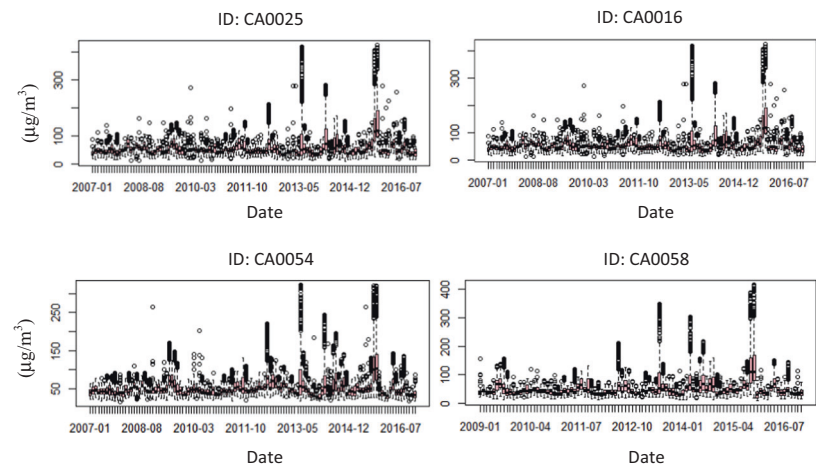
Analysis of Fig. 5 with respect to the Department of Environment's API limits (Low pollution (0–50), Moderate (51–100), Unhealthy (101–200), Very unhealthy (201–300), and Hazardous and risky (301–500) [36]), shows that the value of PM<sub>10</sub> did not exceeded 150 API (unhealthy), except in 2013 and 2015, and with maximum individual hour API values of 426 in Oct 2015, 416 in Oct 2015, 390 in Oct.2015, and 322 in June 2013 at stations ca0025, ca0058, ca0016 and ca0054 respectively.

Station CA0016 falls in the middle of a highly polluted area. This renders the conventional mean plot ineffective

**Fig. 4 Box plots.** Box plots of median concentration of PM10 (monthly average).



**Fig. 5 Whisker plots.** Whisker plots of PM10 API concentration for the period Jan.2007-July.2016, except ID CA0058 that started two years later (Jan.2009).



to observe the pollutants' sources, thus, underscoring the limitations of existing techniques in such instances. A similar scenario exists in station CA0058 due to the presence of some industrial areas in the north and north east directions. Also, station CA0025 falls in the middle of highly polluted area and an airport, which complicates the detection of pollutants' sources. Therefore, to overcome this limitation, CBPF is required to divide the mean values into 10 quantiles for further analysis. Figure 6 depicts the CBPF plot of mean values of main pollutants measured at the four stations. However, the absence of dominant pollutants around CA0054 enables a clear rendition of the effects of KL on the north and north west areas.

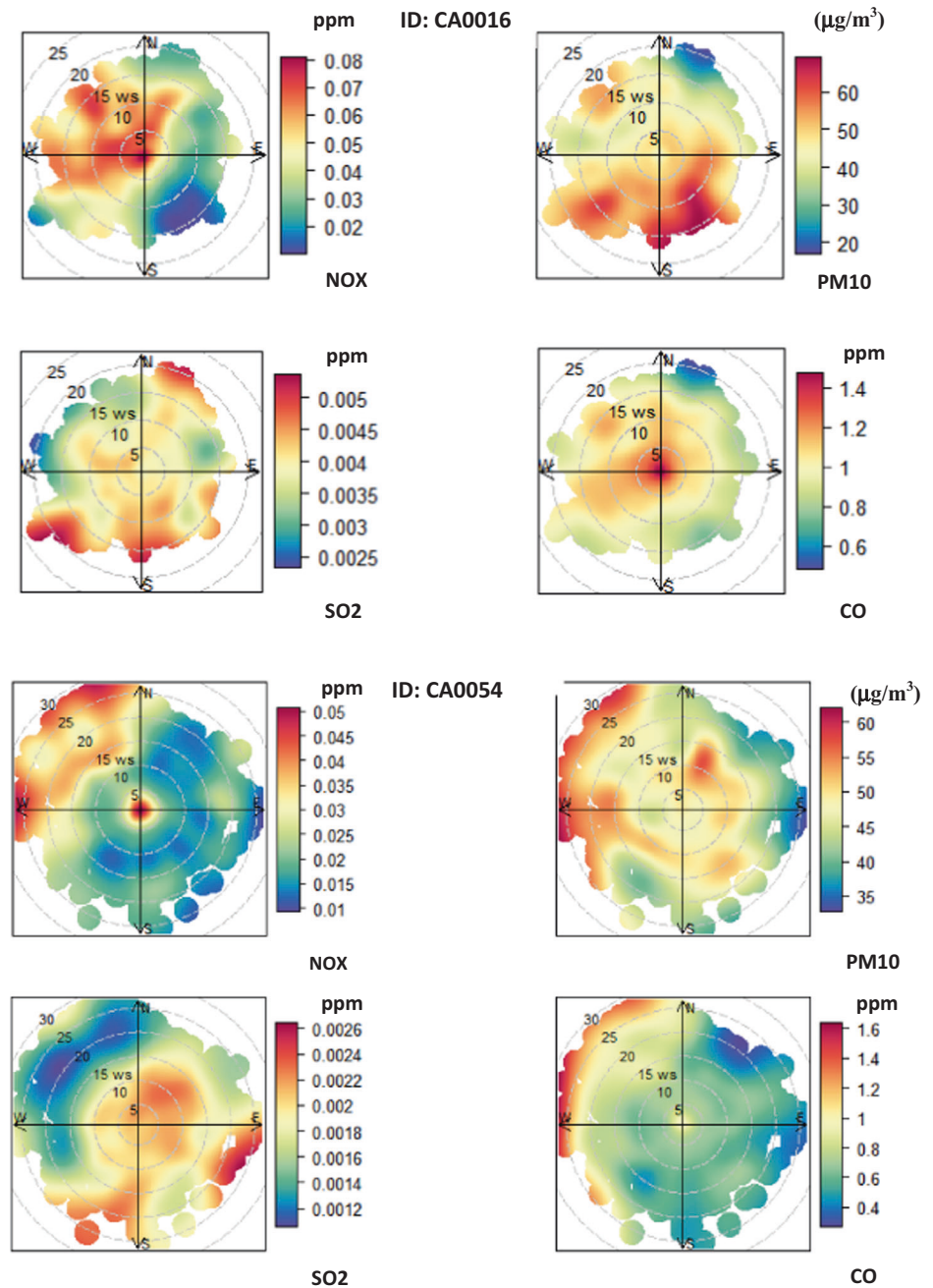
Although every pollutant has a specific emission range, this is usually 'washed out' in the normal conventional polar plot [37]. Therefore, considering the quantiles intervals using the CBPF's interval classification can reveal more information on pollutants' sources.

The CBPF was calculated for all the stations (Fig. 7) by taking the 10 quantiles of each pollutant for better visualization of different emission sources. This is particularly pertinent for stations CA0016 and CA0025 that did not reveal clear sources and directions for the pollutants as shown in Fig. 6. The CBPF plots reveal significant hidden information about different pollution sources which was not discernable in the ordinary plot, especially for the stations that are located in the middle of high-density industrial areas and pollutants. Subsequent analysis focused on the main pollutant in the two stations, PM10, since it always has the highest value in comparison to other pollutants, thus, determining the Air Pollution Index (API).

Nearer land use at station CA0016 have the highest percentile values while the distant ones recorded low percentiles. Analyzing Fig. 7 in tandem with Fig. 1, land uses in the NW direction were detected by two percentiles, 7th (47–52) and 8th (52–58), while the third zone of land use at the southern direction of the station is represented by SW



**Fig. 6 CBPF plot.** CBPF plot of mean values of main pollutants measured in four stations in volumetric units (ppm), except PM10 measured in Gravimetric unit ( $\mu\text{g}/\text{m}^3$ ).



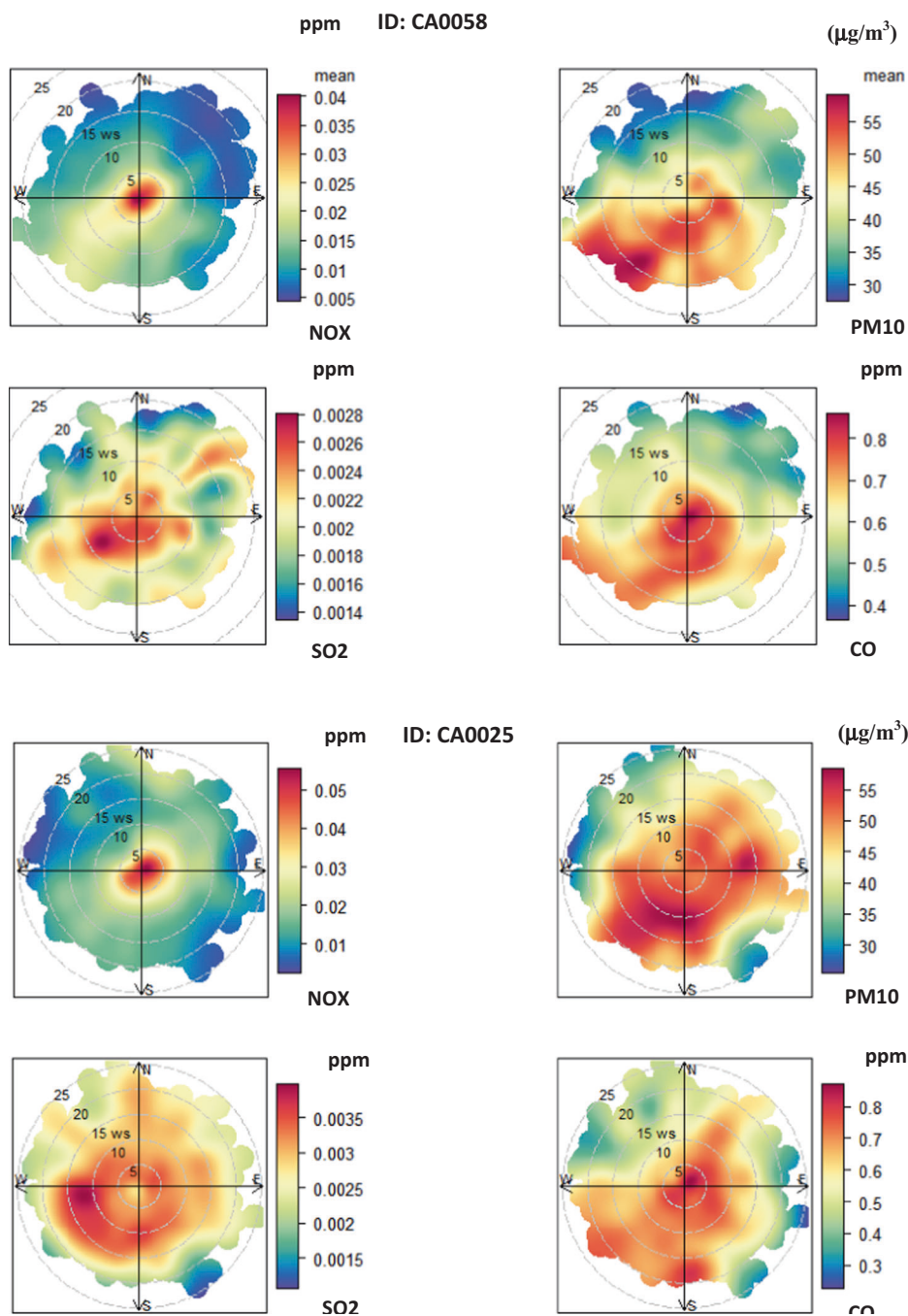
9th (58–70) and SE 10th (70–390) percentiles. Despite their low API values, CPBF successfully detected distant pollutants at the NE direction, in the 1st (12–29) and 2nd (29–33) percentiles.

The presence of many land use surrounding station CA0025 presents unique challenges in identifying sources of diverse pollutants, but the CBPF percentile plot presents some interesting findings. In the N and the NW direction, we observed the Subang airport and a huge industrial area, respectively. These were measured by stations on 2nd (30–35) and 3rd (35–39) percentiles, respectively. In S and SW directions, the effect of these two regions was higher

than the airport's location, and detected by 9th (64–76) and 10th (76–426) percentiles respectively, in addition to the 5th percentile (43–47) that detected the pollution originating from the center of the study area (eastern direction of the station). Other stations detected many sources that were not observed in close region to the station, but with low effect. The major effect originated from the close land use as shown in Fig. 7 for the two stations (ca0058 and ca0054).

Further insights can be gained by considering how percentile values vary in relation to other factors i.e., conditioning. For example, the plot shows the seasonal variation and whether it is nighttime or daytime. Also,

Fig. 6 (Continued)



PM10 concentration results in Fig. 7 indicate that identification of sources varies according to percentiles (high to low), while Fig. 8 highlights the impact of seasonal variation on identification of pollutants' sources. At station CA0016, PM10 is higher in September, October, and November, which refers to the SE direction that was indicated in Fig. 7. It is important to note that there are no significant differences between day and night concentrations. This suggests that PM10 is not affected by day or night temperature, but rather by the pollutants' active hours and seasons.

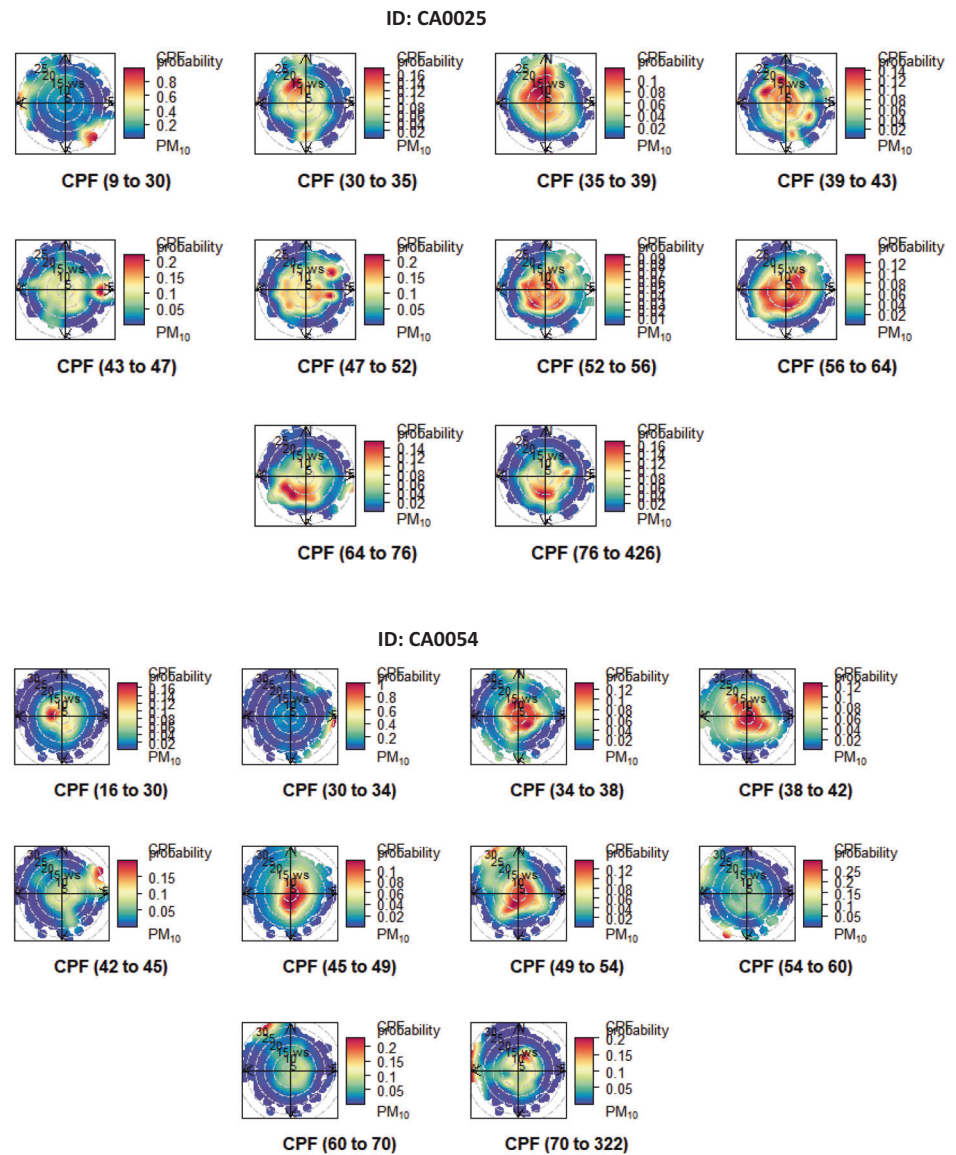
**Correlation between PM10 and other pollutants**

**Generalized linear model**

The three major pollutants, SO2, NO2, and CO, show positive correlation with PM10 in all stations based on outputs of generalized linear model using percentile value. CO has the highest correlation value in station CA0058, followed by CA0025.

A similar trend is observed in other cities, particularly in the Asian region. Table 3 documents selected studies on

**Fig. 7 Polar plots.** Polar plots of concentrations at 4 stations based on the CBPF function for a range of percentile intervals from 0–10, 10–20, ..., 90–100.



major pollutants and their concentration in different Asian cities. The data in the table shows the average concentration of the major pollutants, SO<sub>2</sub>, NO<sub>2</sub>, CO, and PM<sub>10</sub>. A correlation between PM<sub>10</sub> and the other pollutants is established in all the cities, which aligns with the findings of our present study. Vehicle emission and proximity to industrial areas is the major source of pollutants in these cities and our study area too.

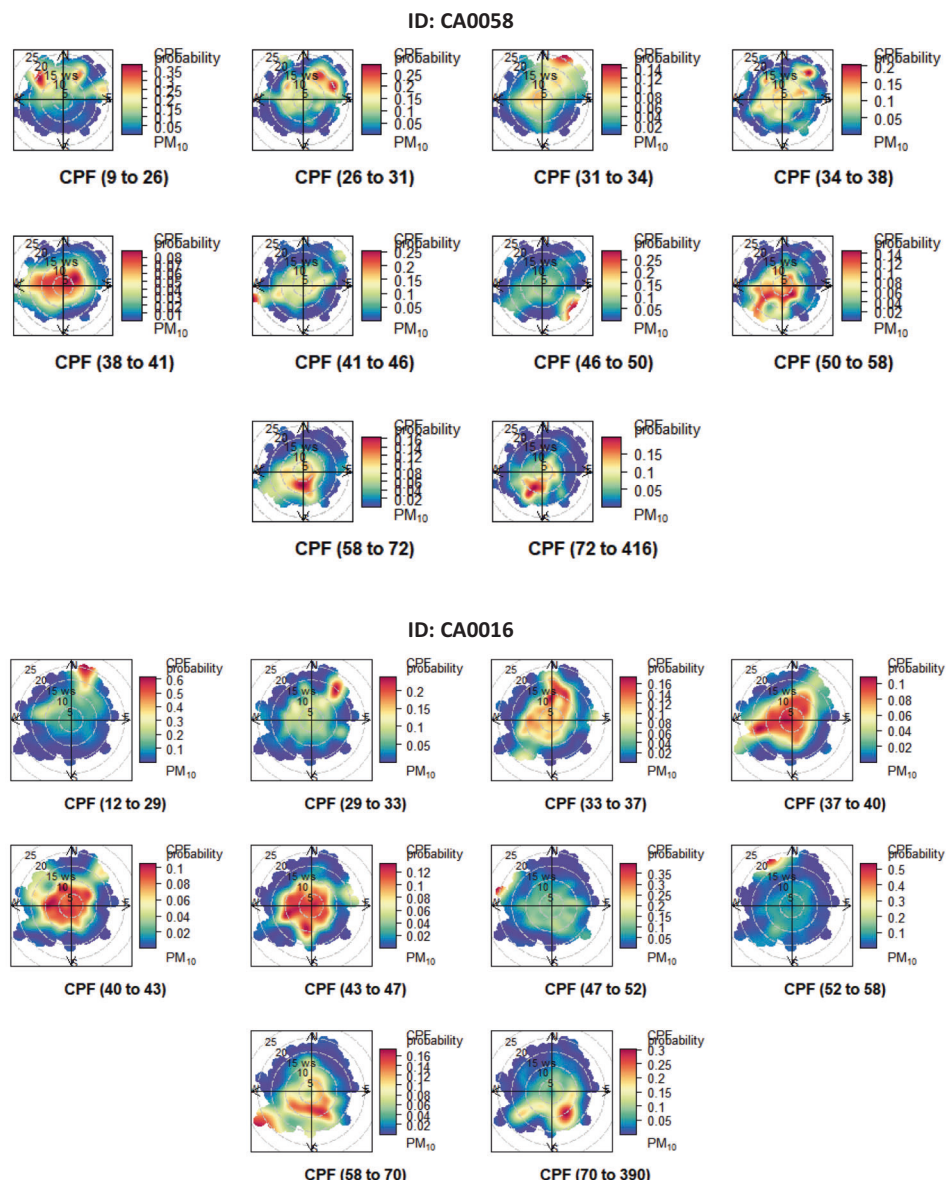
**Sensitivity analysis**

Sensitivity analysis was conducted at all stations to observe the relation between PM<sub>10</sub> and three major pollutants based on the correlation observed from outputs of generalized linear model using percentile value. CO, NO<sub>2</sub>, and SO<sub>2</sub> are

taken as explanatory variables whose values are kept constant at their minimum 20th percentile, 40th percentile, 60th percentile, 80th percentile and maximum percentile. At station CA0058, NO<sub>2</sub> and CO clearly displayed non-linear relation with PM<sub>10</sub>, showing a logarithmic growth and tend to level off with increasing value of the pollutants. SO<sub>2</sub> had almost negligible response. At station CA0054, the sensitivity model for SO<sub>2</sub> and NO<sub>2</sub> showed exponential growth trend in the first four quantiles and descended at maximum quantile. At station CA0025, the pattern of response by PM<sub>10</sub> varied with respect to quantiles. In the first four quantiles in CO and NO<sub>2</sub>, the graph shows slow logarithmic growth at the beginning, followed by a sharp increase. In contrast, for the maximum concentration quantile values, PM<sub>10</sub> exhibited an exponential growth with respect to both



Fig. 7 (Continued)



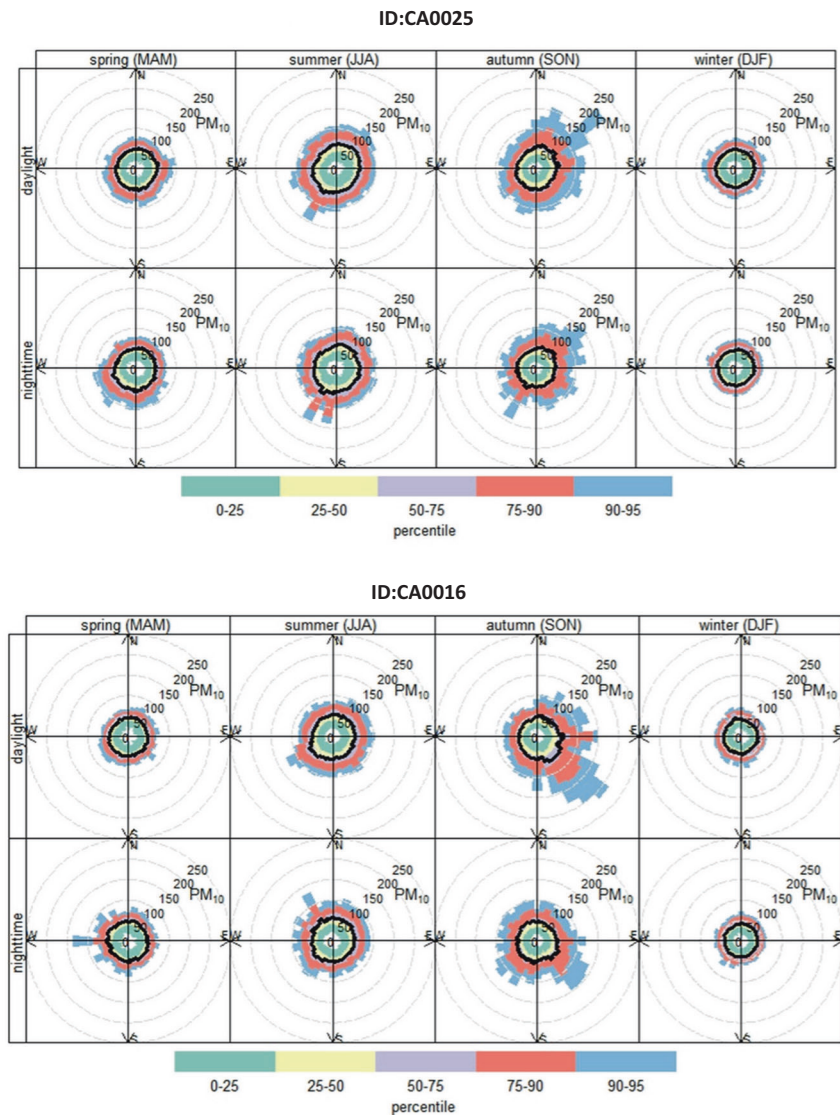
CO and NO<sub>2</sub>. However, SO<sub>2</sub> shows a linear relationship with the occurrence of PM<sub>10</sub>. At station CA0016, CO and PM<sub>10</sub> displayed a unimodal relation. The response pattern of PM<sub>10</sub> is non-linear with different patterns in different quantiles. PM<sub>10</sub> tends to show linear response with respect to SO<sub>2</sub>.

The sensitivity analysis at station CA0025 (Fig. 9) indicates that the current pollutants are not driven by daylight and night hours. This is evident in the concentration pattern from September to November, which increases especially towards the NE sources (Airport and city center). Between June and August, some pollutants started to emit PM<sub>10</sub> with lower values. However, these sources fall outside the data coverage, necessitating further investigation to understand the variance in seasonal concentration especially in PM<sub>10</sub>.

### Time series signature and forecast

By using tk functions in R, we obtain residual and forecasting plots. Figures S1a, S2a, S3a, and S4a (see Supplementary Material) show the mean residual of testing data for the period 1 Jan 2016–31 Dec 2016 for the stations CA0058, CA0054, CA0025, and CA0016, respectively. In these figures, the blue line shows the residual values and gray points represent tested data. Generally, the residual plots show an irregular pattern at all stations. Figures S1b, S2b, S3b, and S4b (see Supplementary Material) illustrate the 3-month forecast of PM<sub>10</sub> for the stations CA0058, CA0054, CA0025, and CA0016, respectively. The red points represent the modeled values of PM<sub>10</sub> while the white line represents the mean values of predicted concentration of PM<sub>10</sub>. The forecast plots reveal a cyclical pattern at all

**Fig. 8 Rose plot.** A percentile Rose plot of PM<sub>10</sub> concentration plotted for 3 months average (season) and relation to daylight and nighttime at two stations at the study area.



stations, indicating a seasonal behavior. It is also observed that the forecast concentration of PM<sub>10</sub> is very near and represent a continuous distribution to tested data.

## Conclusion

In the current article, we proposed an enhanced methodology to reliably identify the sources of PM<sub>10</sub>, SO<sub>2</sub>, CO, and NO<sub>2</sub> in high density urban areas of Kuala Lumpur, Malaysia, and other surrounding areas. We also investigated the correlation between PM<sub>10</sub> and the other pollutants and forecast the future occurrence and sources of PM<sub>10</sub> to aid air pollution mitigation and management strategies. To achieve the study's objectives, we selected 10 years' data from four randomly distributed air pollution stations-CA0016, CA0025, CA0058 and CA0054. The study

leveraged recent advances in statistical programming algorithms such as CPBF; polar plot; percentile seasonal and daily plot; and whisker plot embedded in OpenAir package within the R programming environment. These advanced algorithms have the capability to optimize the visualization mechanism of long time series air pollution data of large urban regions that are naturally correlated in time and spatially complicated to analyze, thereby elaborating different sources of pollution that were hitherto undetectable using ordinary plot models. Further, Generalized Linear model (GLM) and sensitivity analysis was applied to assess the relationship between Air Pollution Index (API) of PM<sub>10</sub> and other significant pollutants of GML outputs based on quantile values. The whisker time series plot showed that the value of PM<sub>10</sub> was within the city's healthy API range, with exceptions in 2013 and 2015. Results from the CBPF plot indicated that CA0058 and CA0054 enable easier



**Table 3** Characterization of pollutants in Asian cities.

Reference	Site	Pollutants	Season/Period	Pollutant sources	Method
Yang et al.	Western region, China	So <sub>2</sub> = 0.0203 ppm No <sub>2</sub> = 0.0243 ppm CO = 1.31 ppm PM <sub>10</sub> = 123.0 µg/m <sup>3</sup> So <sub>2</sub> = 0.038 ppm No <sub>2</sub> = 0.0424 ppm CO = 2.1 ppm PM <sub>10</sub> = 138.6 µg/m <sup>3</sup>	Spring    Winter	Industrial sectors, natural sources and sandstorms, meteorological factors   Coal consumption more in winter as compared to all other seasons	Weather Research and Forecasting model coupled with Chemistry (WRF-Chem)
Yousefian et al.	Tehran, Iran	So <sub>2</sub> = 0.00789 ppm No <sub>2</sub> = 0.05334 ppm CO = 2.55 ppm PM <sub>10</sub> = 83.26 µg/m <sup>3</sup>	Annual - 2017	Meteorological factors	Convergent cross mapping
Yin et al.	Shanghai, China	So <sub>2</sub> = 14.33 µg/m <sup>3</sup> No <sub>2</sub> = 58.00 µg/m <sup>3</sup> CO = 0.82 µg/m <sup>3</sup> PM <sub>10</sub> = 66.67 µg/m <sup>3</sup> So <sub>2</sub> = 12 µg/m <sup>3</sup> No <sub>2</sub> = 44.33 µg/m <sup>3</sup> CO = 0.76 µg/m <sup>3</sup> PM <sub>10</sub> = 62.67 µg/m <sup>3</sup>	Winter    Summer	Natural factors (such as air pressure, temperature, wind direction and speed, etc.), but human factors (such as industrial waste gas emissions, domestic coal combustion, automobile exhaust emissions, etc.)	Comprehensive Pollution Index Method, Improved Gray Relational Degree Method, and Euclid Approach Degree Method
Dumka et al.	Dehli, India	Maximum concentration: So <sub>2</sub> = NA No <sub>2</sub> = NA CO = 1.7 ± 0.5 ppm PM <sub>10</sub> = 245.5 ± 109.8 µg/m <sup>3</sup> Average during daytime: So <sub>2</sub> = NA No <sub>2</sub> = 13.25 (±1.90) ppb CO = 1.58 (±0.45) ppm PM <sub>10</sub> = 222.05 (±98.09) µg/m <sup>3</sup>		Fossil-fuel combustions from vehicles, industries, and domestic use of natural gas dominated when compared to burning of wood, waste material, dung cakes, and agricultural crop residue	Multi-instrument research campaign

Table 3 (continued)

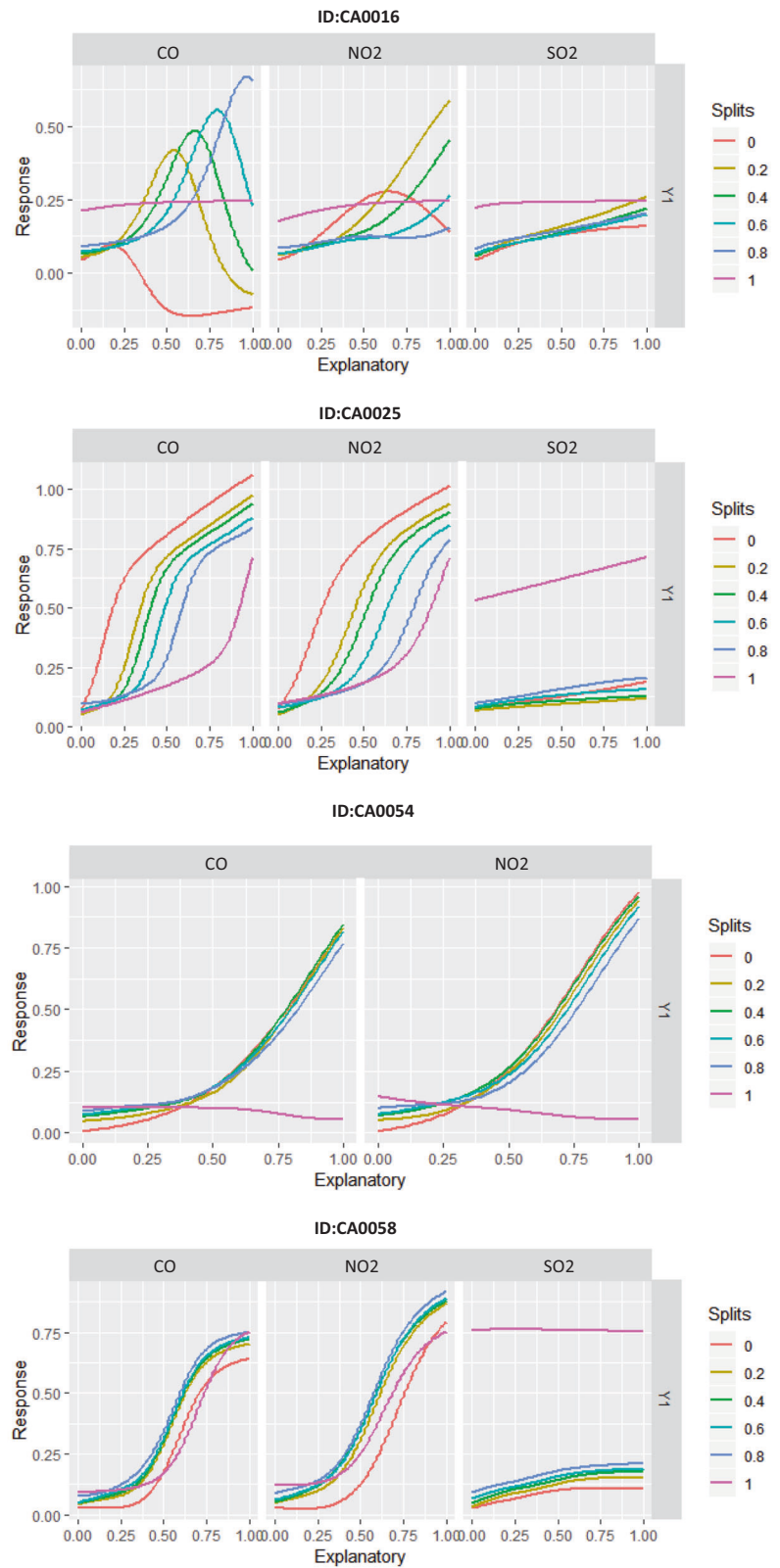
Reference	Site	Pollutants	Season/Period	Pollutant sources	Method
Ismail et al.	Northern region includes Perlis, Kedah, Pulau Pinang and Perak, Malaysia	PM10 = 673 µg/m <sup>3</sup> O <sub>3</sub> = 0.170 ppm  SO <sub>2</sub> = 0.170 ppm NO <sub>2</sub> = 0.1 ppm CO = 5.98 ppm	Seasonal	1. SO <sub>2</sub> , NO <sub>2</sub> and O <sub>3</sub> are the major pollutants 2. Due to the combustion process from vehicles and industries.	1. Cluster analysis for pollutants' volumes, followed by discriminant analysis (DA) to differentiate between these clustered classes. 2. Identify air pollution sources using principal component analysis (PCA)
Yoo et al.	Four land-use regional study of residence, commerce, industry and greenbelt, South Korea	Average annual in industry areas: O <sub>3</sub> = 23.2 ppb CO = 5.8 (0.1 ppm) NO <sub>2</sub> = 23.2 ppb SO <sub>2</sub> = 6.8 ppb PM10 = 56.4 µg/m <sup>3</sup>	Cycles (diurnal, weekly, and annual)	1. Vehicle emissions.  The highest concentrations were found in the industrial areas for SO <sub>2</sub> and PM10; in the commercial areas for NO <sub>2</sub> and CO; and in the greenbelt areas for O <sub>3</sub> , respectively. CO, NO <sub>2</sub> , SO <sub>2</sub> and PM10 pollutants, except for O <sub>3</sub> , were generally higher in the big cities during the weekdays, while the O <sub>3</sub> showed its highest values in the small cities during the weekends.	The variations were analyzed in terms of the cycles of the pollutants, their trends and inter-relationship.

detection of pollutants' sources in comparison to CA0016 and CA0025 due to their location near industrial areas. The highest monthly concentrations were observed in August at station CA0025, presumably due to its proximity to Subang airport, which has significant chemical pollution episodes. The CBPF plots revealed significant hidden information about different pollution sources in the study area, which was not discernable in ordinary plot, particularly for the stations that are located around high-density industrial areas. The sensitivity analysis at station CA0025 revealed that the pollutants are not affected by daylight and night hours and temperature. Rather, PM10 concentration is influenced by seasonal variations. Analysis of the Time series signature and forecast revealed that the residual plots generally indicate an irregular pattern at all stations while the forecasting plots reveal a cyclical pattern at all stations, which is indicative of a seasonal behavior. Based on these findings, the major conclusion from this study are summarized below:

- Despite missing data attributes and scarcity of air quality stations, the use of long-term data with advanced statistical and spatial analysis enhances the understanding of the mobility of pollutants along the study area, spatially and temporally.
- Proper distribution of air quality monitoring stations provides deep insights on temporal scenarios of pollution distribution, aided by detailed land use maps and Google earth services.
- CBPF plot can identify the sources of pollution in cases where the stations are located at the edge of pollutants' sources or there is a great variance in pollution concentration. Otherwise, CBPF's results will not be valid.
- Polar plot of concentrations based on the CBPF function for a range of percentile intervals from 0 to 100 successfully elaborated the different sources of pollution around the stations, which was not visible in the ordinary plot.
- Closer pollutants with high intensity have major impact on the stations' concentration reading, although small concentrations emanating from distant pollutants can still be detected.
- There are no significant differences between day and night concentrations in the study area. This offers interesting insights on the nature of PM10, which is not affected by temperature of daylights and night. However, further investigation of the impact of vehicular traffic is necessary.

The visual interpretation of results and analysis has impacts on comprehending the urban air pollution dynamics using advanced statistical models to convey vital information to the local community, decision makers and all stakeholders.

**Fig. 9 Sensitivity analysis.** Sensitivity analysis at the study area showing the effect of variable values' splits with response.



For future work, it is imperative to get a detailed land use data within the industrial land use for the purpose of comparing the outcome with acceptable or dominant range of pollution concentration for each industry.

**Acknowledgements** The authors gratefully acknowledge the financial support from the University Teknologi PETRONAS (UTP) STIRF research grant [0153AA-F83] for this project. Also, we are very grateful to Department of Environment (DoE), Malaysia, for providing the air quality data used in this study and the Federal Department of Town and Country Planning (PLANMalaysia) for providing spatial and attribute data of the study area.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## References

- Seinfeld JH, Pandis SN. Atmospheric chemistry and physics: from air pollution to climate change. New York: John Wiley & Sons; 2016.
- Jiang D, Zhang Y, Hu X, Zeng Y, Tan J, Shao D. Progress in developing an ANN model for air pollution index forecast. *Atmos Environ*. 2004;38:7055–64.
- Murena F. Measuring air quality over large urban areas: development and application of an air pollution index at the urban area of Naples. *Atmos Environ*. 2004;38:6195–202.
- Jeričević A, Gašparac G, Mikulec MM, Kumar P, Prtenjak MT. Identification of diverse air pollution sources in a complex urban area of Croatia. *J Environ Manag*. 2019;243:67–77.
- Huang Z, Yu Q, Ma W, Chen L. Surveillance efficiency evaluation of air quality monitoring networks for air pollution episodes in industrial parks: pollution detection and source identification. *Atmos Environ*. 2019;215:116874.
- Salim I, Sajjad RU, Paule-Mercado MC, Memon SA, Lee B-Y, Sukhbaatar C, et al. Comparison of two receptor models PCA-MLR and PMF for source identification and apportionment of pollution carried by runoff from catchment and sub-watershed areas with mixed land cover in South Korea. *Sci Total Environ*. 2019;663:764–75.
- Potier E, Waked A, Bourin A, Minvielle F, Péré J, Perdrix E, et al. Characterizing the regional contribution to PM10 pollution over northern France using two complementary approaches: Chemistry transport and trajectory-based receptor models. *Atmos Res*. 2019;223:1–14.
- Tiwari A, Kumar P, Baldauf R, Zhang KM, Pilla F, Di Sabatino S, et al. 2441. Considerations for evaluating green infrastructure impacts in microscale and macroscale air pollution dispersion models. *Sci total Environ*. 2019;672:410–26.
- Pokorná P, Hovorka J, Hopke PK. Elemental composition and source identification of very fine aerosol particles in a European air pollution hot-spot. *Atmos Pollut Res*. 2016;7: 671–9.
- Holmes NS, Morawska L. A review of dispersion modelling and its application to the dispersion of particles: an overview of different dispersion models available. *Atmos Environ*. 2006; 40:5902–28.
- Qin Y, Oduyemi K. Atmospheric aerosol source identification and estimates of source contributions to air pollution in Dundee, UK. *Atmos Environ*. 2003;37:1799–809.
- Kim E, Hopke PK. Comparison between conditional probability function and nonparametric regression for fine particle source directions. *Atmos Environ*. 2004;38:4667–73.
- Malby AR, Whyatt JD, Timmis RJ. Conditional extraction of air-pollutant source signals from air-quality monitoring. *Atmos Environ*. 2013;74:112–22.
- Ashbaugh LL, Malm WC, Sadeh WZ. A residence time probability analysis of sulfur concentrations at Grand Canyon National Park. *Atmos Environ*. 1967;19:1263–70. 1985
- Henry R, Norris GA, Vedantham R, Turner JR. Source region identification using kernel smoothing. *Environ Sci Technol*. 2009;43:4090–7.
- Bae M-S, Schwab JJ, Chen W-N, Lin C-Y, Rattigan OV, Demerjian KL. Identifying pollutant source directions using multiple analysis methods at a rural location in New York. *Atmos Environ*. 2011;45:2531–40.
- Uria-Tellaetxe I, Carslaw DC. Conditional bivariate probability function for source identification. *Environ Model Softw*. 2014;59:1–9.
- Marmur A, Park S-K, Mulholland JA, Tolbert PE, Russell AG. Source apportionment of PM<sub>2.5</sub> in the southeastern United States using receptor and emissions-based models: conceptual differences and implications for time-series health studies. *Atmos Environ*. 2006;40:2533–51.
- Elangasinghe M, Singhal N, Dirks K, Salmund J, Samarasinghe S. Complex time series analysis of PM10 and PM<sub>2.5</sub> for a coastal site using artificial neural network modelling and k-means clustering. *Atmos Environ*. 2014;94:106–16.
- Khan J, Kakosimos K, Raaschou-Nielsen O, Brandt J, Jensen SS, Ellermann T, et al. Development and performance evaluation of new AirGIS-A GIS based air pollution and human exposure modelling system. *Atmos Environ*. 2019;198:102–21.
- Gulliver J, Briggs D. STEMS-Air: a simple GIS-based air pollution dispersion model for city-wide exposure assessment. *Sci Total Environ*. 2011;409:2419–29.
- Wang Y, Zhang X, Draxler RR. TrajStat: GIS-based software that uses various trajectory statistical analysis methods to identify potential sources from long-term air pollution measurement data. *Environ Model Softw*. 2009;24:938–9.
- Mukherjee A, Agrawal M. Assessment of local and distant sources of urban PM<sub>2.5</sub> in middle Indo-Gangetic plain of India using statistical modeling. *Atmos Res*. 2018;213:275–87.
- Ding H, Kumar KR, Boiyo R, Zhao T. The relationships between surface-column aerosol concentrations and meteorological factors observed at major cities in the Yangtze River Delta, China. *Environ Sci Pollut Res*. 2019;26:36568–88.
- Rana MM, Khan MH. Trend characteristics of atmospheric particulate matters in major urban areas of Bangladesh. *Asian J Atmos Environ*. 2020;14:47–61.
- Kang N, Deng F, Khan R, Kumar KR, Hu K, Yu X, et al. Temporal variations of PM concentrations, and its association with AOD and meteorology observed in Nanjing during the autumn and winter seasons of 2014–2017. *J Atmos Solar Terrestrial Physics*. 2020;203:105273.
- Jain S, Sharma S, Vijayan N, Mandal T. Seasonal characteristics of aerosols (PM<sub>2.5</sub> and PM10) and their source apportionment using PMF: a four year study over Delhi, India. *Environ Pollut*. 2020;262:114337.
- Althwaynee OF, Pradhan B. Semi-quantitative landslide risk assessment using GIS-based exposure analysis in Kuala Lumpur City. *Geomatics Nat Hazards Risk*. 2017;8:706–32.
- Sanusi M, Ramli A, Hassan W, Lee M, Izham A, Said M, et al. Assessment of impact of urbanisation on background radiation

- exposure and human health risk estimation in Kuala Lumpur, Malaysia. *Environ Int.* 2017;104:91–101.
30. Shakir SK, Azizullah A, Murad W, Daud MK, Nabeela F, Rahman H, et al. Toxic metal pollution in Pakistan and its possible risks to public health. *Rev Environ Contam Toxicol.* 2016;242:1–60.
  31. Sulong NA, Latif MT, Khan MF, Amil N, Ashfold MJ, Wahab MIA, et al. Source apportionment and health risk assessment among specific age groups during haze and non-haze episodes in Kuala Lumpur, Malaysia. *Sci Total Environ.* 2017;601:556–70.
  32. Khan MF, Hamid AH, Bari MA, Tajudin ABA, Latif MT, Nadzir MSM, et al. Airborne particles in the city center of Kuala Lumpur: origin, potential driving factors, and deposition flux in human respiratory airways. *Sci Total Environ.* 2019;650:1195–206.
  33. Tajudin MABA, Khan MF, Mahiyuddin WRW, Hod R, Latif MT, Hamid AH, et al. Risk of concentrations of major air pollutants on the prevalence of cardiovascular and respiratory diseases in urbanized area of Kuala Lumpur, Malaysia. *Ecotoxicol Environ Saf.* 2019;171:290–300.
  34. Shah AD, Bartlett JW, Carpenter J, Nicholas O, Hemingway H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol.* 2014;179:764–74.
  35. Sahrir S, Bachok S, Osman MM. Environmental and health impacts of airport infrastructure upgrading: Kuala Lumpur International Airport 2. *Procedia-Soc Behav Sci.* 2014;153:520–30.
  36. Awang MB, Jaafar AB, Abdullah AM, Ismail MB, Hassan MN, Abdullah R, et al. Air quality in Malaysia: impacts, management issues and future challenges. *Respirology.* 2000;5:183–96.
  37. Carslaw DC, Ropkins K. Openair—an R package for air quality data analysis. *Environ Model Softw.* 2012;27:52–61.