Data Analytics for Visualizing Air Pollutant Curves: Application of Functional Data Analysis

Dr. Norshahida Shaadan
Center of Statistical and Decision Science Studies
Faculty of Computer & Mathematical Sciences,
UiTM, 40450, Shah Alam, Selangor

Presentation Outline

- **I**ntroduction
- **P**roblem Statement
- **R**esearch objectives
- **M**ethodology
- **R**esults of Analysis
- **C**onclusion

Introduction

- Air quality monitoring has become part of the initial strategy in the pollution prevention program in Malaysia.
- The government has established the Malaysian Air Quality Guidelines (MAAQG) to set up standard and the implementation of Air Pollutant Index (API).
- API is developed based on five important air pollutants namely, Ozone (O3), suspended particulate matter of less than 10 microns in size (PM10), Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2) and Carbon Monoxide (CO).
- Alam Sekitar Malaysia Sdn Bhd (ASMA) has been given the responsibility and has been awarded a concession (by Malaysian Government) for the monitoring since 1995.
- ASMA provides fully integrated air monitoring systems which continuously recorded the pollutant data in the air.
- The reasons or aims for such data collection are to assess the extent of pollution, provide air quality information to the public, support implementation of air quality goals or standards, evaluate the effectiveness of emissions control strategies and etc.
- Hence in achieving the aims, this is where " Data Analytics" play the role and become important.

Introduction

- **What is Data Analytics?** - refers to qualitative and quantitative techniques and processes used to enhance data evolvement and exploration to facilitate decision-making.
- **Common type and level of data analytics**:

✓ Descriptive analytics
✓ Inference analytics
✓ Predictive analytics
✓ Prescriptive analytics

- **Guidelines for employing data analytics**: The correctness and success is depending on:
  - the research objectives
  - the level of data measurement: nominal, ordinal, interval or ratio
  - the type of data –quantitative of qualitative
  - the dimension of attributes (variables) – univariate, multivariate

Problem Statement

- The variation of air pollutants level often due to different background environment, the variation of meteorological variables, the interaction of the factors as well as the pattern of the emission sources.
- Understanding the temporal variation of air pollutants is important because it would provide information on the critical time and sources of air pollution as well as providing new insights on the complex underlying of a particular pollutant process such as its formation and destruction.
- **However, despite of the availability of large amounts of data, a gap persists in terms of the kind of analytics to highlight this continuous and dynamics information by means of a *continuous statistics*. This is because the conventional statistical analysis namely the univariate and multivariate analysis have limitation due to the *static nature* of the conventional statistics.**
- Aiming to increase understanding on the air pollutants behavior, thus, this research will investigate the whole fluctuation of the recorded data within the continuum of a specific time period in the form of curves. Instead of using the discrete recorded values, researcher will analyze the data as a function of time. Hence, this approach would offer wider kind of data analytics to be employed and developed.
- This paper highlights on the analytics on how to visualize temporal behavior of PM10 and Ozone pollutant. by means of statistical analytics using curves data.
- By using this approach, it is hoped that the temporal and dynamic pattern of pollutant concentration levels can be explored and visualized.

Research Objective

The main objective is to visualize the temporal (diurnal) behavior of selected air pollutants (PM10 and Ozone) at several selected locations in the Selangor state of Malaysia which consist of visualizing the:

i. average diurnal cycle

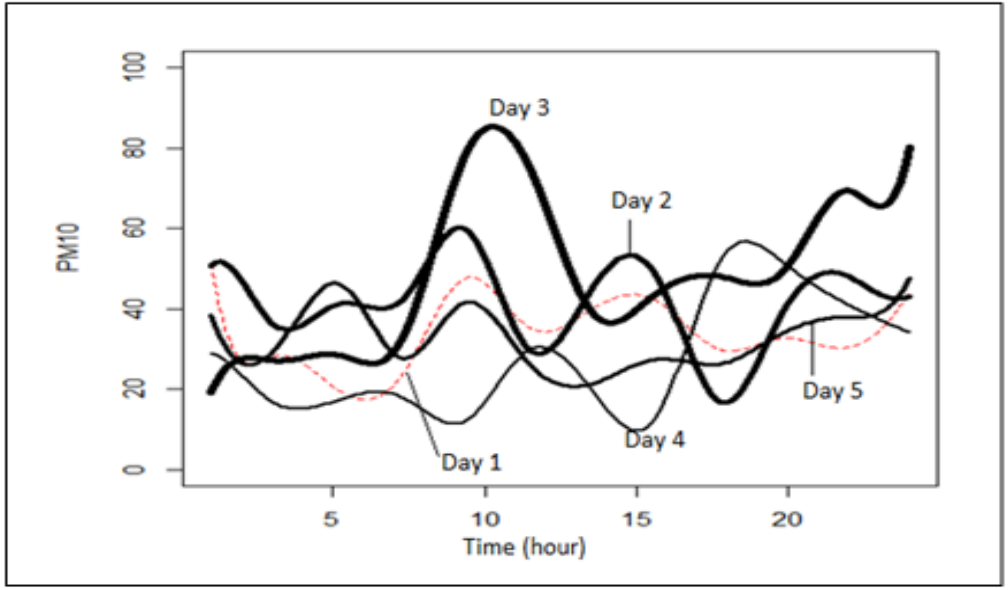
ii. mode of diurnal variation

iii. dynamics pattern: rate of change and the acceleration

What is Functional Data Analysis (FDA)?

- **Functional Data Analysis (FDA)** is branch of statistics that analyzes curves data or functional data (Ramsay and Silverman, 2006).
- **Curves also refers to Functional Data -**is defined as recorded data that arise in a continuum such as time or space.
- FDA involves transforming data points to continuous functions, which is basically done using Fourier function (basis) or spline functions.
- FDA are the analysis that providing information about curves, surfaces or anything else that varies over time.
- FDA analysis is done in terms of functions instead of single data point.
- The physical forms of the data are curves or surfaces.
- These data are represented by functions x(t).

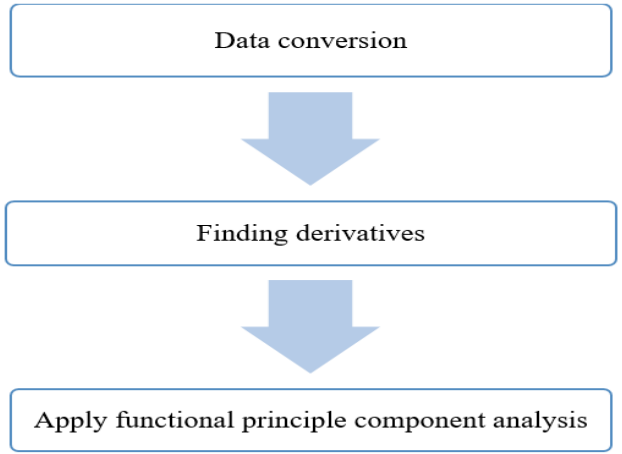
Example of Curves Data



Methodology: Curves Analytics –the procedure

FUNCTIONAL DATA ANALYSIS (FDA)

Framework:
To achieve the study objective



The **analytics** consist of the procedure and techniques for data conversion from discrete to curves, the application of functional descriptive statistics and functional principal component analysis (FPCA).

Curves Analytics....

- Converting raw data into curves.

-Method: Basis function expansion

- Formula:

$$x_i(t) = \sum_{k=1}^{K} C_k \phi_k(t)$$

Daily curve data were modeled using a ***set of spline functions*** given as $x(t)$ where $t \in [1,24]$, which is a continuous time function of Ozone and PM10 levels over 24 hours period

Describing Curve Data

The formulae for functional descriptive statistics are as follows:

$$\text{Functional mean, } \bar{x}_i(t) = \frac{\sum x_i(t)}{n} \quad (2)$$

$$\text{Functional rate of change, } \Delta_1 = x_i'(t) \quad (3)$$

$$\text{Functional acceleration rate, } \delta_1 = x_i''(t) \quad (4)$$

FPCA: Analytics for Investigating Dominant Pattern (Mode)

- ✓ One of the most important exploratory tools in FDA.
- ✓ Used to search for several important (principal) components that can describe major variations in curves data.
- ✓ The components are independent and account for different proportion of variability –given by eigenvalue.
- ✓ The first contribute the largest amount of variation, the second contribute the second largest amount and so on – which independently indicating different info.

Each principal component is specified by an eigen weight function $\xi(t)$ that represent the dominant features of variation in the functional data (curves).

THEORY AND CONCEPT OF FPCA

- ✓ The methods use an eigenvalue decomposition of the covariance matrix to find direction in the observation space along with the data have the highest variability.
- ✓ The direction of variation in the functional context, for each principal component, it is specified by a principal component weight function -

Computational aspect:

Solving the eigen equation problem- given by:

$$\int v(s,t)\xi(t)dt = \lambda\xi(s)$$

FDA Application (Case Study): Research Data and Location

Type of Data: Hourly Recorded Data

Source of Data: Department of Environment Malaysia (DOE)

Pollutant 1: Ozone (O3)

Duration: 2013-2015

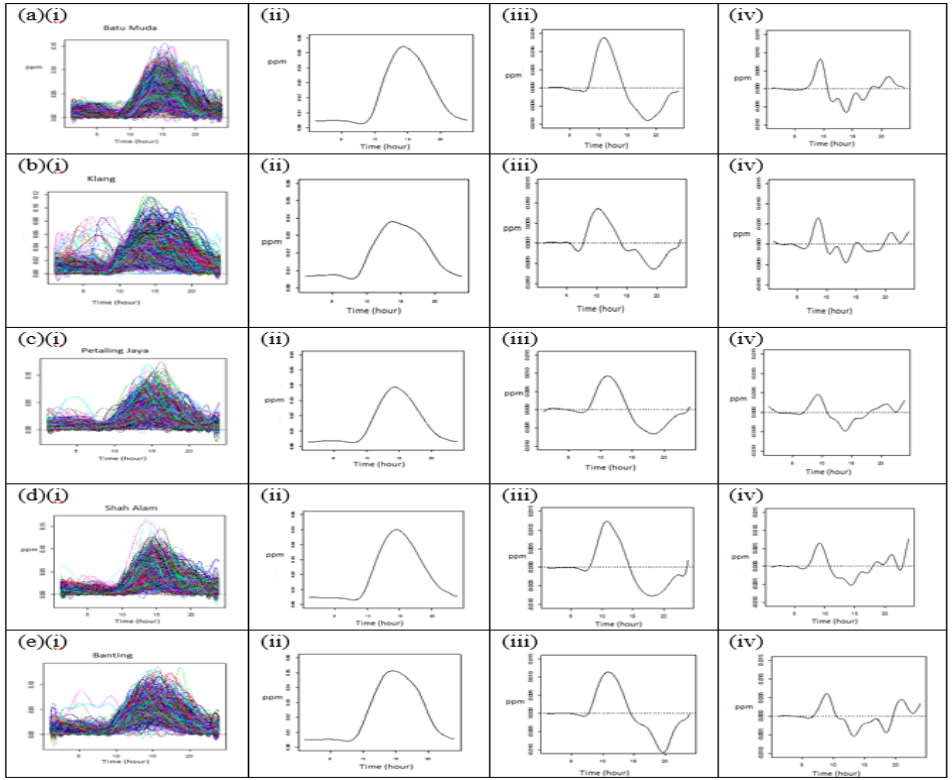
Location: Batu Muda, Klang, Petaling Jaya, Shah Alam, Banting

Pollutant 2: PM10

Duration:2001-2010

Location: 30 air quality monitoring sites in Peninsular Malaysia

Figure: (i) 1641 Ozone Curves (ii) Average (iii) Rate of change (iv) Acceleration rate



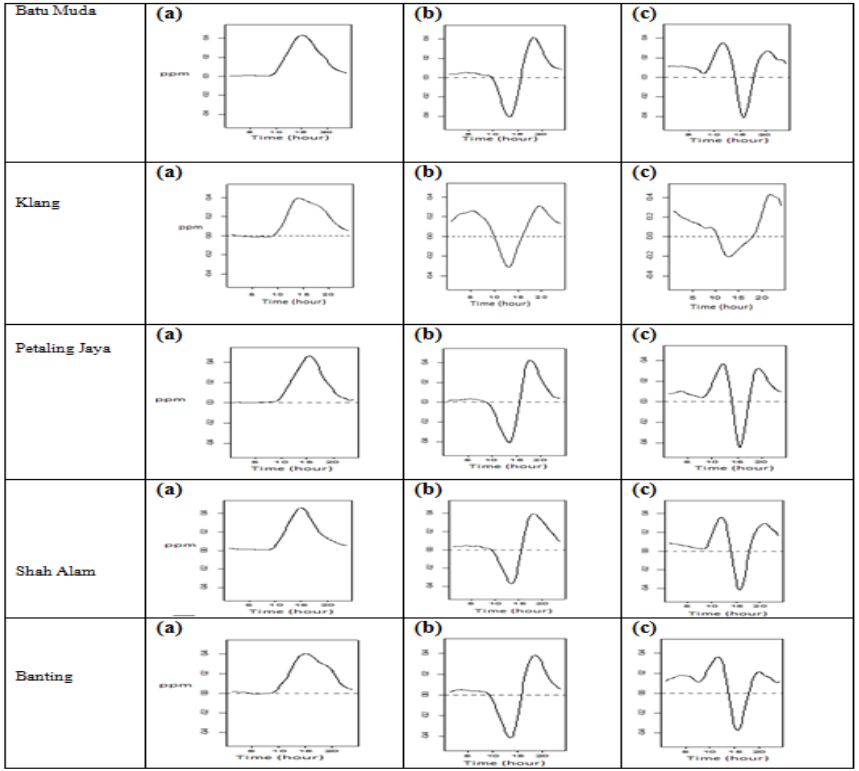
Results of Analysis

Findings:

- ✓ Ozone curves over 24 hours period for all sites is about the same with the **peak** level at 3 pm in the **evening** and **lower** during **night-time**
- ✓ More Ozone **anomalies** occurred in Klang while Banting and Petaling Jaya are the two stations with higher Ozone intensity.
- ✓ Batu Muda and Banting have recorded more number of **exceedance days** compared to other stations since the stations having more curves with peak exceeding the standard (>0.1 ppm)
- ✓ The average day to day Ozone level at all sites having similar **bimodal shape**.
- ✓ On the average, the **critical exposure hours** with high health risk have occurred during the day period which lies between **12 pm till 5 pm** at all sites with the highest concentration level occurred around 3 pm
- ✓ **Ozone rate of change** is similar with bimodal curves for all stations. Evaluation on the acceleration of average Ozone shows that at all sites Ozone change rates were **nearly constant after late midnight**, **positive after sunset** and during the day before around 3 pm and **negative after 3 pm**. Shah Alam and Batu Muda have recorded the highest peak change rate while Banting has the lowest.
- ✓ **the change rate at Batu Muda is the fastest compared to other stations.**
- ✓ **Ozone acceleration helps to identify Ozone critical spurt-the - point in the day where Ozone starts to form at 9 am at all stations.**

RESULTS OF THE ANALYSIS

Application of FPCA: Three Important mode of pattern of Ozone Diurnal Cycles



Findings:

✓Most of the days in Batu Muda, Klang, Petaling Jaya, Shah Alam and Banting having a bimodal shape of Ozone cycles exhibited by PC1.

✓It is also shown that all sites also having similar pattern of PC2 variations, the second major pattern of Ozone cycle. Ozone concentration during the day reached the peak level at night time around 8 pm and the valley around 12 pm. Thus, this valley shows that the concentration level had the minimum drop at midnight hour.

✓The third mode of variation explained by PC3 explains that some of the days having two major peak of Ozone levels at 10 am and 8 pm and experiencing maximum drop at 3 pm in the evening. This behavior occurred at all study locations except Klang.

Results of analysis

FPCA Application: Features of eigen weight functions **(PM10)** $\xi(t)$

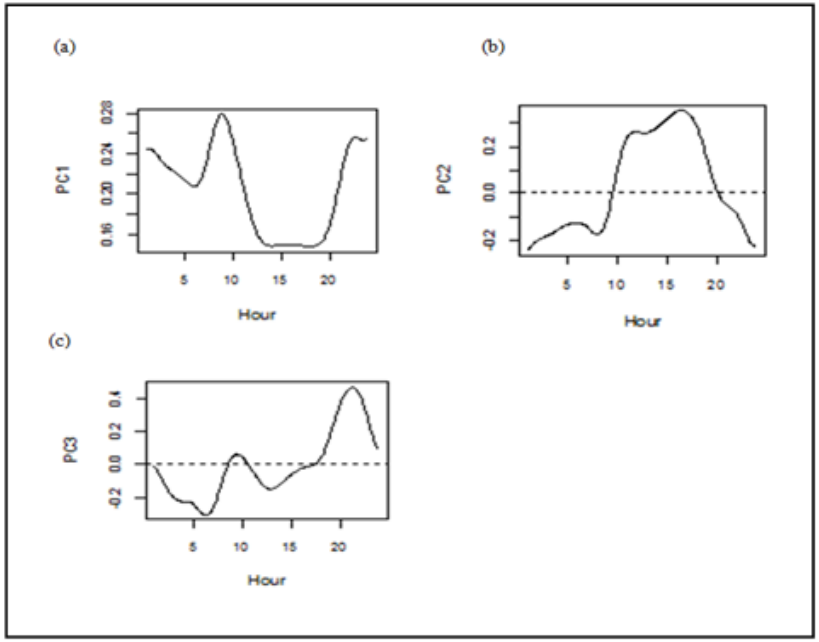


Fig. 3: Contribution of time (hour) to the diurnal variation shown by (a) PC1, (b) PC2, and (c) PC3 eigen-weight function curve.

- Eigen weight functions explains the contribution hours to the PM10 variations

- PC1 -defines the highest mode of variation – positive through out the day-producing trend of a peak during busy hours – heavily weighting morning and late evening rush time hours – this pattern indicates contribution by vehicular activity (Chang and Lee, 2007)
- PC2- shows positive contribution for the day time hours –starts after 10.00 am, end around 5.00 pm –negative at other time.
  -The trend oh high during daylight hours coincide with period of hours for photochemical activity (Chang and Lee,2007).
  -PC3- the pattern of variation described the contribution of mixing factors (Morawska et al. 2007).

Results of analysis

- To help interpreting the variation use plot of PC as perturbation to the mean curves

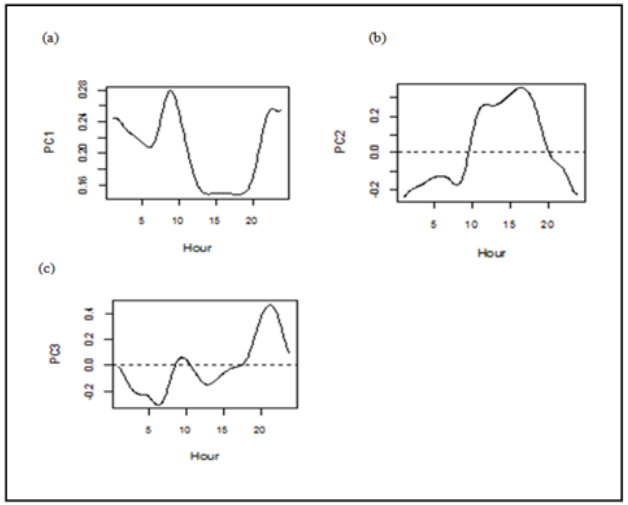
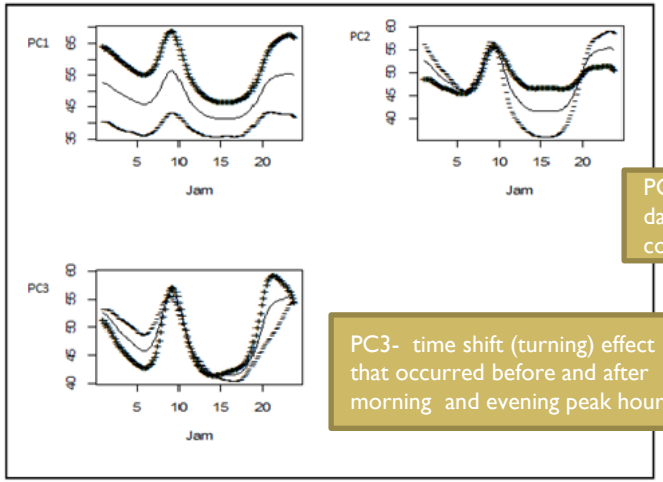


Fig. 3: Contribution of time (hour) to the diurnal variation shown by (a) PC1, (b) PC2, and (c) PC3 eigen-weight function curve.

PC1 - Vertical shift in the level



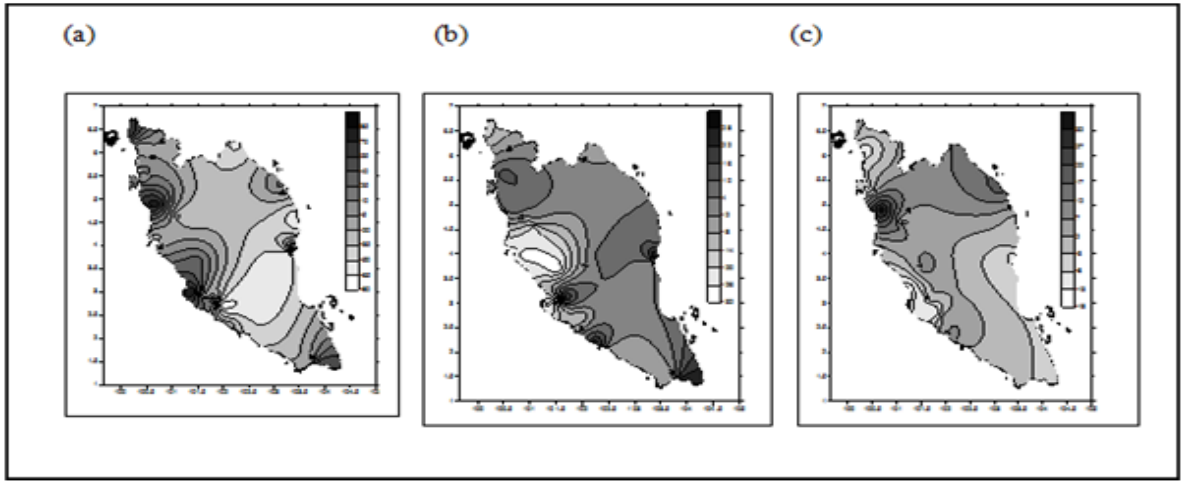
PC2-Time shift before and after daytime hours –capture daylight contrast in the level

PC3- time shift (turning) effect that occurred before and after morning and evening peak hours

Fig. 4: The average PM10 diurnal curves (middle line) and the effect of adding (+) and subtracting (-) a suitable multiple of each PC curve.

Results of analysis

- Spatial pattern of PM10 diurnal variations according to PCs



•The first mode of variation PC1- dominantly occupied the western coastal region.
•The second mode PC2- dominant in the northwest (Kedah state) and east (Terengganu) coastal region.
•The third mode PC3 - delineates the northern part of the Malaysian Peninsular.

CONCLUSION

The application of FPCA gives several advantages: it provides the ability to visualize, evaluate, and describe continuous variation of air pollutants; Ozone and PM10 concentration level over a day period.

The results provide the information such as:

1. The average day to day diurnal Ozone cycles in Selangor is regional but the magnitude level and other kind of dynamics such as the rate of change and acceleration is non-regional.
2. Batu Muda is shown to be the most potential location with higher risk of day to day (diurnal) Ozone pollution and has been identified to experience the most active ozone formation and Banting has the slowest Ozone destruction.
3. There exist three major component patterns of daily Ozone cycles -The largest component shows similar pattern from site to site. However, for the second and third components, the pattern is different in Klang.
4. The diurnal Ozone and PM10 variations provide meaningful insight on the emission sources and the contributing factors.
5. Peninsular Malaysia experiences three major patterns of PM10 diurnal variation.
6. The results have also provided evidence that, vehicular emission is the primary source of PM10 pollution in Peninsular Malaysia. Industrial activity and mixing factors are also shown to be the second and third major contributing factors towards PM10 variation.

Acknowledgement:

Special thanks to the Department of Environment (DOE), Putrajaya, Malaysia for providing the data . The work on Ozone (O3) pollution profile using FDA is supported by the UiTM Research University Grant

[600-IRMI/DANA5/3/LESTARI (0124/2016)]