Effect of Monsoonal Clustering for PM_{10} Concentration Prediction in Keningau, Sabah using Principal Component Analysis

To cite this article: Muhammad Izzuddin Rumaling *et al* 2022 *IOP Conf. Ser.: Earth Environ. Sci.* **1103** 012003

Effect of Monsoonal Clustering for PM_{10} Concentration Prediction in Keningau, Sabah using Principal Component Analysis

Muhammad Izzuddin Rumaling[1], F P Chee[1*], J H W Chang[2], J Sentian[3]

[1]Faculty of Science and Natural Resources, Universiti Malaysia Sabah (UMS), Kota Kinabalu, Sabah, Malaysia
[2]Preparatory Centre for Science and Technology, Universiti Malaysia Sabah (UMS), Kota Kinabalu, Sabah, Malaysia
[3]Climate Change Research Group, Faculty of Science and Natural Resources, Universiti Malaysia Sabah (UMS), Kota Kinabalu, Sabah, Malaysia

Email: fpchee06@ums.edu.my

Abstract. Particulate matter (PM) has caught scientific attention in scientific research due to its harmful effect on human health. While prediction is essential for future development in Keningau, temporal clustering in Keningau has yet to be studied. Thus, this research aims to determine whether monsoonal clustering is required for meteorological and pollutant concentration data collected in Keningau. Missing data is first imputed using Nearest Neighbour Method (NNM). Then, wind direction and wind speed are converted into northern (W_y) and eastern (W_x) component of wind speed. Data is then temporal clustered based on monsoonal season (NEM, IM_4, SWM, IM_{10}). Both clustered and unclustered data are analysed using principal component (PC) analysis (PCA). The findings revealed that humidity in PC_1 with average EV (explained variation) of 93.92 ± 0.52 contribute the most variation of PM_{10}, followed by W_x in PC_2 with average EV of 3.51 ± 0.48. Regression analysis shows that humidity and PM10 are negatively moderate to strongly correlated except for IM_4 (intermonsoon April), which may be due to dry climate during the season. As for W_x, it has weak correlation with PM_{10}. This may be due to location of Keningau at western part of Crocker range. However, the spread of PM_{10} due to eastern wind causes weak to zero correlation. Due to consideration of dry climate as revealed by the findings from IM_4 cluster, there is need for data collected by Keningau to be clustered by monsoon.
Keywords: PM_{10}, nearest neighbour method, monsoonal cluster, principal component analysis, regression analysis

1. Introduction

Particulate matter is ambient particle that are respirable and suspended in air [1, 2]. Due to its effect on human health, it has attracted scientific attention in conducting research on particulate matter [3]. PM_{10} (particulate matter with aerodynamic size less than 10 microns) is considered to be one of major air pollutants, and thus is continuously monitored in stations around Malaysia. Major sources of PM_{10} includes biomass burning, motor vehicles, and industrial activities [4]. Depending on its composition, PM10 poses more hazard to human health compared to other pollutants especially on children. It has caused various lung diseases and reduces body's ability to fight infections [1].

Due to hazardous effect of PM_{10}, short-term and long-term prediction is essential for early preventive measures and reduces casualties due to high concentration in ambient PM_{10} [3]. As a developing city such as Keningau in Sabah, long-term prediction is essential in proper development planning [2, 3]. The prediction becomes more accurate when significant factors affecting PM_{10} are considered. However, the effect of monsoonal season in Keningau on variation of meteorological and gaseous factors have not yet been studied. Thus, this research aims to study the significance of monsoonal season in variation of meteorological and gaseous factors that affect PM_{10} concentration in Keningau.

2. Data and Methods

2.1. Study area and data

Keningau (5.34° N, 116.15° E, altitude: 288 m) is the largest district located in interior part of Sabah. It is located in a valley in the western part of Crocker Range, as shown in Figure 1. The climate in Keningau is the driest in Sabah, with average annual rainfall between 1100 – 1500 mm [5].

Figure 1. Location of Keningau (5.34° N, 116.15° E)

The monitoring station in Keningau (CA0049) is located at SMK Gunsanad and is at relatively high altitude compared to other monitoring stations found in Sabah. CA0049 is operated by a private company known as Alam Sekitar Sdn. Bhd. (ASMA) under Department of Environment (DOE). CA0049 is operated at temporal resolution of 1 h. Meteorological data (W_s, W_d, Hum, Temp) and pollutant concentration (CO, NO_2, O_3, SO_2, PM_{10}) are recorded at every hour and are made available by Air Quality Division under DOE. In this research, data from 2003 to 2012 is studied.

Angular quantity such as wind direction Wd is not usually used in principal component analysis (PCA) or regression analysis [2, 6]. Equations (1) and (2) are employed to convert W_d into linear quantities W_x and W_y. Positive values of W_x and W_y indicate that the wind is blowing eastward and northward respectively.

$$W_x = W_s \sin W_d \quad (1)$$

$$W_y = W_s \cos W_d \quad (2)$$

2.2. Nearest Neighbour Method (NNM)

One of the most widely used method for missing data imputation is known as Nearest Neighbour Method (NNM) [7]. The missing data y at corresponding time x contained in a stream of missing data with known borders (x_1, y_1) and (x_2, y_2) can be imputed using (3) and (4) [8]. In order to reduce further data loss, W_s and W_d are imputed using Nearest Neighbour Method before converted into W_x and W_y.

$$y = \begin{cases} y_1, \text{ when } x < x_1 + \bar{x} \\ y_2, \text{ when } x \geq x_1 + \bar{x} \end{cases} \tag{3}$$

$$\bar{x} = (x_2 - x_1)/2 \tag{4}$$

2.3. Monsoonal Clustered and Unclustered Data

Data grouped into a set based on similar characteristics is known as clustered data. Meanwhile, ungrouped data is known as unclustered data. In this research, data is clustered in terms of monsoonal season (NEM, IM_4, SWM, IM_{10}). This is because Keningau experiences monsoonal season due to its geographic location. Thus, Keningau experiences Northeast Monsoon (NEM) from November to March and Southwest Monsoon (SWM) from May to September. Meanwhile other season is known as intermonsoon, occurred in April (IM_4) and October (IM_{10}).

2.4. Principal Component Analysis (PCA)

In order to determine the most significant variable in contributing PM10 variation, Principal Component Analysis (PCA) is used [7]. This method transforms a set of variables to n number of principal components (PC). Compared to original variable X, PC accounts for more significant variation. Equation (5) explains the ith principal component, PC_i, with l as factor loading of a variable [6].

$$PC_i = l_{1i}X_1 + l_{2i}X_2 + \ldots + l_{ni}X_n \tag{5}$$

3. Results and Discussion

Unclustered and clustered data at CA0049 are analysed using PCA and the factor loadings are tabulated in Table 1 along with explained variation (EV) and cumulative explained variation (CEV). The most significant variable contributing to variation of each PC is highlighted. Two first PCs are usually chosen because both PCs take account into 90% of variation [6]. However, Table 1 shows that for both clustered and unclustered data, the PC_1 have taken more than 90% of variation into account. Therefore, both PCs are considered for factor loading plot [6].

Table 1. PCA of monsoonal clustered and unclustered data in Keningau

Component	Unclustered		Clustered							
			NEM		***IM₄***		***SWM***		***IM₁₀***	
PC	PC_1	PC_2	PC_1	PC_2	PC_1	PC_2	PC_1	PC_2	PC_1	PC_2
W_x	–0.023	**0.976**	–0.025	**0.979**	–0.047	**0.973**	–0.019	**0.997**	–0.015	**0.851**
W_y	–0.003	–0.213	0.011	–0.200	0.017	–0.224	–0.015	–0.072	–0.018	–0.524
Hum	**0.979**	0.027	**0.979**	0.032	**0.977**	0.050	**0.978**	0.023	**0.978**	0.011
Temp	–0.205	0.023	–0.202	0.018	–0.209	–0.005	–0.206	0.022	–0.207	0.037
CO	0.001	–0.003	0.002	–0.005	0.001	–0.005	0.001	-	0.001	–0.004
NO_2	-[a]	-	-	-	-	-	-	-	-	-
O_3	-	-	-	-	-	-	-	-	-	-
SO_2	-	-	-	-	-	-	-	-	-	-
EV (%)	93.81	3.52	93.40	4.14	94.56	3.23	94.11	3.06	93.60	3.63
CEV (%)	93.81	97.33	93.40	97.54	94.50	97.79	94.11	97.17	93.60	97.23

[a]Magnitude smaller than 0.001 are not shown

Based on Table 1, it is shown that PC_1 and PC_2 for both clustered and unclustered data are mostly contributed by humidity and east-west-component wind speed respectively. The mean factor loading for humidity (PC_1) in monsoonal clustered data is 0.978 ± 0.001, showing that the most significance in variation for humidity for all clusters. As for east-west-component (W_x) wind speed (PC_2), the mean factor loading is 0.950 ± 0.067. W_x is not as significant as humidity because the high factor loading is only reflected in PC_2. Furthermore, EV in PC_2 (3.51 ± 0.48) is drastically lower compared to PC_1 (93.92 ± 0.52). This further reduces the significance of W_x. Based on factor loading plot as shown in Figure 2, the factor loading for humidity and east-west-component wind speed does not show significant difference between clusters.

According to Figure 3, PM_{10} concentration and W_x shows weak correlation for all clusters. Positive correlation may be attributed to geographic location of CA0049 station, which is at western part of Crocker range. High mountain of Crocker range blocks PM_{10} blown by eastern wind. However, it is possible that PM_{10} spreads away towards north, which may cause weak to no correlation.

Furthermore, it is shown that PM_{10} concentration for all clusters (except IM_4) has negative moderate to strong correlation with humidity. Moderate to strong negative correlation is related to high humidity level. PM_{10} absorbs water vapour and may become too heavy to stay suspended in the air and thus deposited [9]. Due to minimal rainfall during intermonsoon in April, low level of humidity causes rise in PM_{10} concentration due to dust emission [8]. This contributes to weak negative correlation ($R^2 = –0.2699$) for IM_4 cluster.

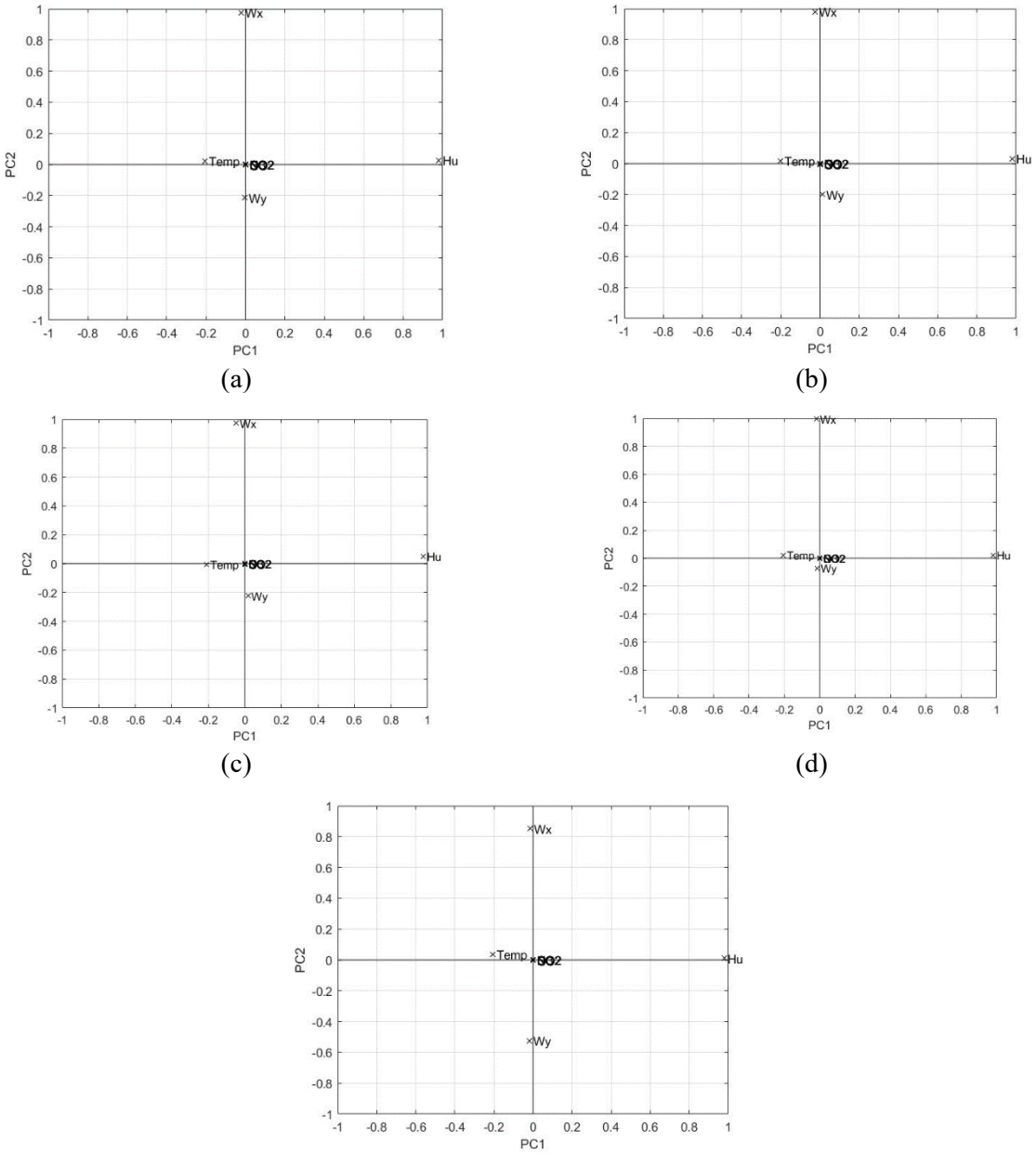

Figure 2. Factor loading plot for (a) unclustered, (b) Northeast Monsoon (NEM), (c) Intermonsoon April (IM_4), (d) Southwest Monsoon (SWM), and (e) Intermonsoon October (IM_{10})

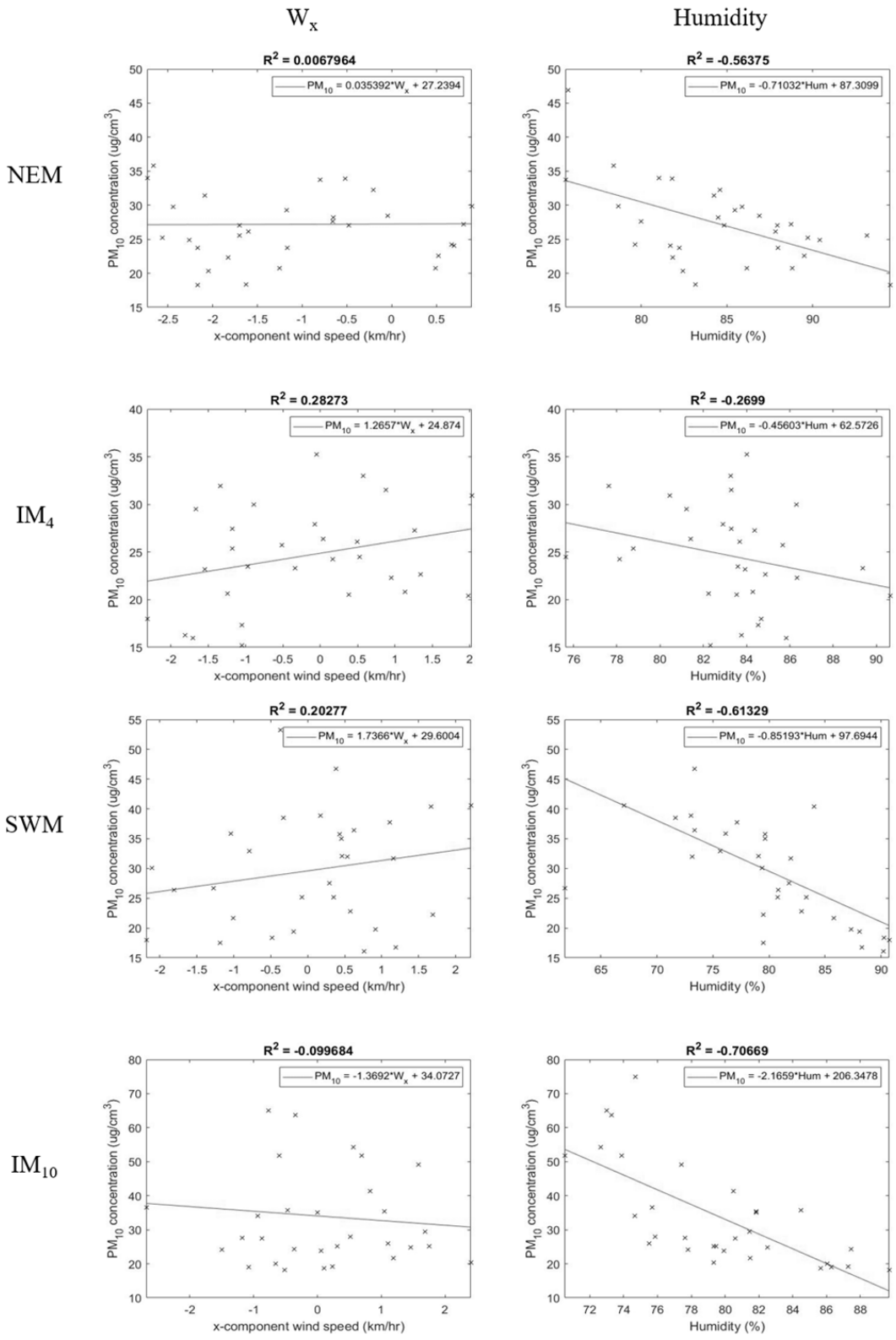

Figure 3. Regression analysis between PM_{10} concentration against both east-west component wind speed and humidity for monsoonal cluster

4. Conclusion

Although PCA shows that factor loading for humidity (PC_1) is almost similar for all monsoonal cluster, regression analysis shows that data collected in CA0049 still needs clustering by monsoon season because the correlation between PM_{10} and humidity during intermonsoon in April is negatively weak compared to other clusters. This ensures that dry climate during intermoonson in April is considered. Therefore, by clustering the data from CA0049 can reduce PM_{10} variation which is crucial for improving the accuracy of long-term prediction and modelling.

Acknowledgment

The authors of this paper would like to thank Universiti Malaysia Sabah for supporting this research by providing research grants (SBK0324-2018, SGI0054-2018 and GUG0378-2019) and Department of Environment Malaysia for providing meteorological and pollutant data from CA0049 for research purpose.

References

[1] Chang, H W J, Chee, F P, Kong, S K S and Sentian J 2018 Variability of the PM10 concentration in the urban atmosphere of Sabah and its responses to diurnal and weekly changes of CO, NO2, SO2 and Ozone *Asian J. Atmos. Environ.* **12**(2) 109–26

[2] Rumaling, M I, Chee, F P, Chang, H W J, Payus, C M, Kong, S K, Dayou, J and Sentian, J 2021 Forecasting particulate matter concentration using nonlinear autoregression with exogenous input model *Global J. Environ. Sci. Manage.* **8**(1) 27–44

[3] Shahraiyni, H T and Sodoudi, S 2016 Statistical modelling approaches for PM10 prediction in urban areas; a review of 21-st century studies *Atmos.* **2**(15) 1–24

[4] Shaadan N, Jemain A A, Latif M T and Deni S M 2014 Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia *Atmos. Pollut. Res.* **6** 365–75

[5] Muhammad M, Abdullah M, Singh M J, Suparta M, Islam M T and Tangang F 2013 Characterization of GPS PWV during flooding event over Keningau, Sabah *International Conference on Space Science and Communication, IconSpace* 429–33

[6] Gvozdić V, Kovač-Andrić E and Brana J 2011 Influence of Meteorological Factors NO2, SO2, CO and PM10 on the Concentration of O3 in the Urban Atmosphere of Eastern Croatia *Environ. Model. Assess.* **16**(5) 491–501

[7] Dominick D, Juahir H, Latif M T, Zain S M and Aris A Z 2012 Spatial assessment of air quality patterns in Malaysia using multivariate analysis *Atmos. Environ.* **60** 172–81

[8] Li L and Liu D J 2014 Study on an air quality evaluation model for Beijing City under haze-fog pollution based on new ambient air quality standards *Int. J. Environ. Res. Public Health* 11(9) 8909–23

[9] Lou C, Liu H, Li Y, Peng Y, Wang J and Dai L 2017 Relationships of relative humidity with PM2.5 and PM10 in the Yangtze River Delta, China *Environ. Monit. Assess.* 189(11) 1–6