Determination of Collinearity Developed in the CMB Model with the Concepts of Multi Linear Regression Analysis

Rejivas V. A.[1]*, Praveen A.[2], Ajitha T.[3]

[1]Department of Civil Engineering, Kerala APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India

[2]Department of Civil Engineering, Kerala APJ Abdul Kalam Technological University, Thiruvananthapuram, Kerala, India

[3]Department of Civil Engineering, Rajiv Gandhi Institute Technology, Kottayam, Kerala, India

*Corresponding author e-mail id: rejivas@gmail.com

doi: https://doi.org/10.21467/proceedings.160.12

ABSTRACT

The Chemical Mass Balance model (CMB) gives an accurate source apportionment for the contribution of the sources with the input data of the source profile and receptor data collected. The source profiles for different sources should have a unique and specific species characterization for getting accurate source apportionment results. But due to the mixing of sources, the species characterization source profile may not have unique and specific species characterization due to the non-availability of the exact representation of particular sources and culminates collinearity of species during the CMB analysis. It leads to negative source apportionment results in the CMB analysis. Multi Linear Regression analysis that addresses in the study can effectively be used to identify the collinearity contributing sources. The Multi Linear Regression parameters such as tolerance, variance inflation factor (VIF), condition index, and variance decomposition proportions developed with the source profile variables (source profiles for soil, paved road dust, biomass, and traffic) are used for identifying the collinear sources. The tolerance value for the soil and paved road dust sources are obtained as 0.001 each and the variance inflation factor (VIF) for both are obtained as 204.2 and 208.8 respectively. It indicates the collinearity between soil and paved road dust. Collinearity diagnostics of the regression equations showed that the condition index and the variance decomposition proportion obtained for the soil and paved road dust were greater than 30 (104.09) and 90% (100%) respectively. Therefore, the presence of strong collinearity between soil and paved road dust can be understood.

Keywords: CMB, Collinearity, Variance inflation factor

1 Introduction

Particulate matter pollution is considered as a major air quality problem that is faced across globally. Several studies have been conducted all over the world for characterizing the concentration of the sources and the effect of Particulate Matter in the ambient air. India is reported to occupy 4th position among the world's largest automobile market which is showing a steadily increasing trend in vehicle count over the period [1]. Hence the concentration of particulate matter from traffic sources are usually much more severe than any other polluting sources [2]- [4]. $PM_{2.5}$ and PM_{10} significant increase can adversely affect human health [5], [6] and also lead to severe cardiac and respiratory diseases such as Asthma, bronchitis, etc. culminating in higher mortality rates [7]- [9].

The source apportionment approach is usually used for calculating the traffic and non-traffic source's contributions to the receptor. The receptor model approach of source apportionment is done by determining the source contributionsat the receptor location by analysing its elemental concentration at the receptor and the mass fractions of elemental species characterization of source profiles. In addition to traffic sources, the non-traffic source contributions such as soil, pavedroad dust, and biomass, etc. are also would be determined by source apportionment. The source apportionment technique using Chemical mass

Proceedings DOI: 10.21467/proceedings.160; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-965621-9-9

balance (CMB) is a commonly used receptor model approach [10]- [12] and the major advantageis that it does not require large sets of sample data for the analysis [13].

As stated earlier, the source contribution towards each receptor would be calculated based on the source profile input and elemental species concentration at the receptor in the CMB model. The CMB model gives an accurate and complete source apportionment result only if the source profile of each source has a unique or specific species characterization. Each source profile from traffic sources and non-traffic sources has its own identical elemental species characterizations and each source profile would have unique species mass proportions [14], [15].

However, the species characterization of source profiles for the sources collected may not be specific always due to the mixing of different sources and it culminates the collinearity of species between the sources. When two or more source profiles are strongly correlated, the collinearity would affect the source apportionment outcome [16]. Due to the influence of collinearity, the CMB model gives source contribution for collinear sources as negative [17]- [19]. The source apportionment for all sources would be executed only by eliminating the negative source contributions [20]. Therefore, the collinearity assessment is a major task for getting an accurate source apportionment result for traffic and non-traffic sources.

Multi-linear regression equations addressed in the studyhave been reported to be effectively used for determining collinearity contributing sources [21]- [23]. The collinearitymeasures of the regression equation such as variance inflation factor (VIF), tolerance, condition index, and variance decomposition proportions are used for the determination of collinear sources [24]. The regression equation parameters developed with the source profile variables that help to identify the collinear sources [25]. It is reported from the study that thevariance inflation factor (VIF) of the regression equation is used as a reference for the collinearity analysis and it can be considered as a simple way to eliminate collinearity for getting a better regression result [23], [26]. The variance inflation factor(VIF) is determined from the following equation i.e. $VIF=1/1-R^2$. The R^2 value indicates the coefficient of determination for each regression equation developed for source profile variables which gives the significant correlation between the variables. When $R^2=0$ means no significant correlation exists between thevariables and hence VIF is unity which indicates no collinearity between the sources. When $R^2=0.8$ or more, the VIF becomes greater than 5 which indicates the collinearity.

The objective of this paper is to address the collinearity issue that leads to inconsistency of source apportionment in Chemical Mass Balance analysis (CMB) of particulate matter collected along a semi-urban road segment. It also explains the sources that contribute to the collinearity in the CMB model with the concepts of the Multi Linear Regression approach.

2 Methodology

2.1 Site Selection

Nedumkuzhy junction is a semi-urban road having an average daily traffic (ADT) of 9500 veh/day which is 13 km away from Kottayam town which is shown in figure 1 (9.34 °N, 76.37°E). The traffic and Particulate Matter (PM) data were collected from the site continuously over a period of 1 week. The traffic data was collected by video graphic method and PM data was collected by a mini volume sampler.

Figure 1: *Nedumkuzhy, Kottayam*

2.2 Sampling procedure

The Mini volume sampler is a handheld sampler occupied with Teflon filter paper ($PM_{2.5}$) which can receive the $PM_{2.5}$ exposure at a flow rate of 5 lit/hr. The filter paper kept at the mini volume sampler was usually changed at every 12-hr interval. The filter paper was weighted by gravimetric method and species analysis for determining the trace element's elemental concentration. The collected sample weight was taken by gravimetric weighing method and later it is analyzed speciously with the help of inductively coupled plasma mass spectrometry (ICP-MS). There are 22 trace element species concentrations in the filter paper that were extracted by ICP- MS. For determining the source profile data, the soil, paved road dust, and wood fine samples were also taken from the site. The samples collected from the site are sieved to $PM_{2.5}$ after drying. It was placed in a resuspension chamber occupied with a mini volume sampler with $PM_{2.5}$ filter paper in it. The $PM_{2.5}$ collected in the filter paper was speciously analyzed and source profiles for soil, paved road dust, biomass, and traffic were determined. The traffic source profile was taken from the available literature.

2.3 Principles of Chemical Mass Balance model

The chemical mass balance model equation is termed as follows.

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij}$$

Where x_{ij} is the concentration measured for the j^{th} species in the i^{th} sample, f_{kj} is the fractional mass proportion of the j^{th} species of material emitted by the source k, g_{ik} is the contribution from the k^{th} source to the i^{th} sample, and e_{ij} is the measurement that cannot be fitted by the model.

2.4 Multi Linear Regression model used for assessing thecollinearity

The influence of collinearity in the elemental mass fractionsof the source profile can be analyzed by the application of the Multi Linear Regression model. The Multi Linear Regression model concepts are explained based on the equation as follows.

$$y = b_1x_1 + b_2x_2 + \cdots . + b_kx_k$$

The Multi Linear Regression analysis is usually executed in such a way that $PM_{2.5}$ emission is considered as the dependent variable (y) and source profiles collected from the site are independent variables (x_1, x_2). For the collinearity analysis, each explanatory variable (x_1, x_2, x_3) are again considered as the dependent variable, and a set of regression equations are formulated with the rest of the independent variables. The statistical performance measures of regression equations of each variable are variance inflation factor (VIF), tolerance value, condition index, and variance decomposition proportion are measured which help to identify the

Proceedings DOI: 10.21467/proceedings.160
ISBN: 978-81-965621-9-9

Series: AIJR Proceedings
ISSN: 2582-3922

collinear sources. As stated earlier, the variance inflation factor (VIF) above 5 indicates the collinearity (i.e.R^2 is above 0.8), and the reciprocal of VIF is termed as tolerance value. If the tolerance value is less than 0.2 means the collinearity exists. The collinearity diagnostics of the regression equation explains the intensity of collinearity with condition index and variance decomposition proportion. By developing a matrix with the explanatory variable, the eigen values are calculated and from that the condition index can be formulated. The number of explanatory variables is the same as that of the sum of eigen values and the average of the eigen values is 1. As the total sum of eigen values is constant, the maximum value indicates that the other eigenvalues are low compared to the maximum (λmax). For the indication of collinearity, the average Eigen values are close to 0.

All the explanatory variable are intercorrelated and very short changes that leads to significant changes in the regression coefficient estimates. The higher value of the condition index is called the condition number. The condition number values between 10 and 30 show the presence of collinearity and higher collinearity indicates if it is above 30. Every explanatory variable has a variance decomposition related to each condition index. When two or more variance decomposition proportions related to the condition number above 10 to 30 or greater than 30 exceed 80 % to 90%, indicate the presence of collinearity in the explanatory variables [25].

3 Result and Discussion

3.1 CMB analysis of source apportionment

The Environmental Protection Agency (EPA) developedCMB 8.2 version was used for CMB analysis. The CMB analyses were conducted by providing source profile data and receptor data (speciously analyzed) as data input. The source profile data for soil, paved road dust, biomass, and traffic data, etc. were collected from the site are used as source profile data. And the elemental mass concentration which was speciously analyzed is used as receptor data. Table 1 shows the CMB analysis result of the source contributions for the soil, paved road dust, biomass, and traffic sources to the receptor sample. It was found from the analysis that the paved road dust contribution was obtained as a negative value compared to other source contributions. As stated earlier it is due to the influence of collinearity that exists in the elemental mass proportion of source profiles.

Table 1: *Source apportionment result of CMB analysis*

Species	**Measured PM (µg/m^3)**	**Calculated PM (µg/m^3)**	**Soil (µg/ m^3)**	**Paved road dust (µg/ m^3)**	**Biomass (µg/m^3)**	**Traffic (µg/ m^3)**
Total weight	32.34	25.78	2.02	-1.48	0.19	9.11
Na	0.36	0.26	0.02	-0.03	0	1.92
Mg	0	0.02	57.79	0	192.4	6.47
Al	0.25	0.24	7.42	-4.92	0.1	0.01
K	0.04	0.01	0.1	-0.42	1.15	0.08
Ca	0.05	0.01	0.18	-0.72	0.58	0.7
Ti	0	0	11.47	-2.01	1.33	0
V	0	0	3.25	-2.34	0.02	3.93
Fe	0.18	0.18	5.73	-3.78	0.03	0.68
Co	0	0	0.35	-0.71	0.06	0.69
Ni	0.01	0	0.07	-0.05	0.05	0.54
Cu	0	0.01	0.65	-3.46	0.44	19.22

Zn	0.05	0.14	0.02	-0.1	0.03	7.76
As	0	0	0.02	-0.02	0	0
Se	0	0.07	0	-0.01	0	43.79
Sr	0	0	0.1	-0.4	0.41	0
Mo	0	0	0.07	-0.08	0.01	0.46
Ag	0	0	1.45	-0.59	0.14	0
Cd	0	0	0.02	-0.15	0.04	1.46
Sn	0	0	0.23	-0.7	0.14	0
Sb	0	0	0	-0.05	0.26	23.43
Ba	0.05	0	0.05	-0.19	0.05	0.26
Pb	0	0.1	0.88	-1.23	0.14	0

From Table 1, it is found that the negative value for the paved road dust contributions affected the whole source proportion result and it was not given an accurate source apportionment result. Due to this negative source contribution,the whole source contributions such as paved, biomass and traffic seemed as not accurate. This collinearity of source contributions cannot be solved by the CMB analysis. However,the CMB result could give source contributions for the other sources only if the negative source profiles are eliminated fromthe CMB data input. But it is an incomplete source apportionment process and the information about the collinearity contributing sources would be still unknown. Therefore, the measures of the Multi Linear Regression modelcan be used for determining the collinearity contributing sources.

3.2 Multi Linear Regression model analysis for collinearity assessment

The Multi Linear Regression analysis with the regression equations developed by considering the source profile variablesuch as soil, paved road dust, biomass, and traffic are illustratedin Table 2 and Table 3.

Table 2*: Coefficient of Multi Linear Regression*

Source profile variables	Tolerance value	Variable Inflation Factor (VIF)
Soil	0.001	204.2
Paved road dust	0.001	208.8
Biomass	0.12	19.47
Traffic	0.941	1.06

Table 3: *Collinearity diagnostics*

Dimension	Eigen values	Condition index	Variance decomposition proportions			
			Soil	Paved road dust	Biomass	Traffic
1	2.35	1	0.01	0.01	0.01	0.02
2	1.25	1.37	0.01	0.01	0.01	0.28
3	0.94	1.58	0.01	0.01	0.03	0.18
4	0.45	2.28	0.01	0.01	0.02	0.49
5	0.01	104.09	1.00	1.00	0.88	0.04

From Table 2, it is implicated that the tolerance value for soil and paved road dust is less than 0.2 (0.001) and the Variance Inflation Factor (VIF) is greater than 5 (204.2,208.8 respectively). It indicates the presence of collinearity between the sources of soil and paved road dust. The collinearity diagnostics of the regression

Proceedings DOI: 10.21467/proceedings.160
ISBN: 978-81-965621-9-9

Series: AIJR Proceedings
ISSN: 2582-3922

equation are shown in Table 3. In the 5th dimension of Table 3, the condition index is shown as 104.098 (greater than 30 indicates strong collinearity) and it hasvariance decomposition proportion for soil and paved road dustabove 90% (100%). Hence it can be found that the collinearity that exists between soil and paved road dust is strong as it exceeds 90%. The whole source contribution to the receptor canbe determined only if the collinearity is eliminated. The importance of synthetic receptor data sets is now being taken asa vital measure that can be used for eliminating collinearity.

3.3 Future scope of the work

The present work addresses the limitation of CMB model analysis as it fails due to collinearity. The present work can be extended by developing a synthetic receptor data set. The collinearity of the species of source profile can be eliminated and source contributions to the receptor can be clearly identified by evolving synthetic receptor data sets. Since the traffic deals a significant role in the emission contributions to the receptor, the synthetic receptor data sets would be generated by considering the traffic flow information in addition to source profile variability as future work. These synthetic receptor data sets and CMB model analysis can be utilized by the Ministry of earth science, the pollution control board, other government agencies, and researchers.

4 Conclusions

The CMB model is a widely used effective source apportionment approach for determining the source contributions to the receptor. The CMB model input for the source profile should be unique and specific which is needed for the CMB model to determine accurate source apportionment. But the elemental species mass proportions may not be unique and specific always due to the mixing of sources that leads to the collinearity of sources. Due to this collinearity issue, the CMB model gives only negative source apportionment results. The CMB model only gives the accurate source apportionment as this negative contributing source is eliminated.

Soil, paved road dust, biomass, and traffic, etc. were the source profile data collected from the site used as source data for the CMB analysis. However, the paved road dust source contributions were obtained as negative due to collinearity. For eliminating this negative contributing source and improving thesource apportionment process, collinearity must be eliminated.The principles of the Multi Linear Regression model can effectively be used to determine the collinearity contributing sources. By using the Multi Linear Regression equation parameters such as tolerance, variance inflation factor, condition index, and variance decomposition proportions, etc. of variables (soil, paved Road dust, biomass, and traffic), the collinearity contributing sources were determined. The tolerance value for the soil and paved road dust was obtained as 0.001 each and the variance inflation factor (VIF) for the same was obtained as 204.2 and 208.8 respectively. Since the tolerance value for soil and paved road dust are less than 0.2 and the variance inflation factor (VIF) are greater than 5 respectively shows a clear indication of collinearity. The collinearity diagnostics of the regression equations showed thatthe condition index obtained was greater than 30 (104.09) and the variance decomposition proportion for the same was greater than 90% (100%). Therefore, it can be attributed to the presence of strong collinearity between the soil and paved road dust sources.

5 Declarations

5.1 Acknowledgment

The authors gratefully acknowledge Sri. Prasanth Hedge, Space Physics Laboratory, VSSC, Trivandrum, Kerala for his extensive support in analysing the receptor sample's elemental characterization.

5.2 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Rejivas *et al.* (2023). Determination of Collinearity Developed in the CMB Model with the Concepts of Multi Linear Regression Analysis. *AIJR Proceedings*, 102-109. https://doi.org/10.21467/proceedings.160.12

References

[1] K. Liu and U. S. Racherla, Innovation, Economic Development, and Intellectual Property in India andChina. 2019.

[2] H. Bogo, D. R. Gómez, S. L. Reich, R. M. Negri, and E. San Román, "Traffic pollution in a downtown siteof Buenos Aires City," Atmos. Environ., vol. 35, no. 10, pp. 1717–1727, 2001, doi: 10.1016/S1352- 2310(00)00555-0.

[3] R. J. Laumbach and H. M. Kipen, "Respiratory healtheffects of air pollution: Update on biomass smoke andtraffic pollution," J. Allergy Clin. Immunol., vol. 129, no. 1, pp. 3–11, 2012, doi: 10.1016/j.jaci.2011.11.021.

[4] Y. Hao, S. Deng, Y. Yang, W. Song, H. Tong, and Z.Qiu, "Chemical composition of particulate matter from traffic emissions in a road tunnel in Xi'an, China," Aerosol Air Qual. Res., vol. 19, no. 2, pp. 234–246, 2019, doi: 10.4209/aaqr.2018.04.0131.

[5] K. H. Kim, E. Kabir, and S. Kabir, "A review on thehuman health impact of airborne particulate matter,"Environ. Int., vol. 74, pp. 136–143, 2015, doi: 10.1016/j.envint.2014.10.005.

[6] J. Rovira, J. L. Domingo, and M. Schuhmacher, "Air quality, health impacts and burden of disease due toair pollution (PM10, PM2.5, NO2 and O3): Application of AirQ+ model to the Camp de Tarragona County (Catalonia, Spain)," Sci. Total Environ., vol. 703, no. xxxx, 2020, doi: 10.1016/j.scitotenv.2019.135538.

[7] S. C. Anenberg et al., "Estimates of the global burdenof ambient PM2:5, ozone, and NO2 on asthma incidence and emergency room visits," Environ. Health Perspect., vol. 126, no. 10, pp. 1–14, 2018, doi: 10.1289/EHP3766.

[8] A. Haikerwal et al., "Impact of fine particulate matter(PM2.5) exposure during wildfires on cardiovascular health outcomes," J. Am. Heart Assoc., vol. 4, no. 7, pp. 1–11, 2015, doi: 10.1161/JAHA.114.001653.

[9] T. Maté, R. Guaita, M. Pichiule, C. Linares, and J. Díaz, "Short-term effect of fine particulate matter (PM2.5) on daily mortality due to diseases of the circulatory system in Madrid (Spain)," Sci. Total Environ., vol. 408, no. 23, pp. 5750–5757, 2010, doi: 10.1016/j.scitotenv.2010.07.083.

[10] I. M. Al-naiema, S. Yoon, Y. Wang, Y. Zhang, R. J. Sheesley, and E. A. Stone, "Source apportionment offi ne particulate matter organic carbon in Shenzhen , China by chemical mass balance and radiocarbon," vol. 240, pp. 34–43, 2018.

[11] N. Cheng, C. Zhang, D. Jing, W. Li, T. Guo, and Q. Wang, "Science Direct An integrated chemical mass balance and source emission inventory model for thesource apportionment of PM 2 . 5 in typical coastal areas," no. February, pp. 1–11, 2020.

[12] A. M. Villalobos et al., "A tm spheric P ollution," vol.6, pp. 398–405, 2015, doi: 10.5094/APR.2015.044.

[13] J. Feng, N. Song, and Y. Li, "Source apportionment ofpahs in road sediments by CMB models: Considering migration loss process," Desalin. Water Treat., vol. 200, pp. 422–431, 2020, doi: 10.5004/dwt.2020.26124.

[14] E. G. Group and A. Sciences, "MATTER USING ORGANIC COMPOUNDS AS TRACERS," vol.2310, no. 22, pp. 3837–3855, 1996.

[15] J. G. Watson, T. Zhu, J. C. Chow, J. Engelbrecht, E.M. Fujita, and W. E. Wilson, "Receptor modelingapplication framework for particle source apportionment," Chemosphere, vol. 49, no. 9, pp.1093–1136, 2002, doi: 10.1016/S0045- 6535(02)00243-6.

[16] X. Zhu, E. Blanco, M. Bhatti, and A. Borrion, "Jo urna l P re of," Sci. Total Environ., p. 143747, 2020, [Online]. Available: https://doi.org/10.1016/j.scitotenv.2020.143747.

[17] L. W. Antony Chen and J. Cao, "PM2.5 Source Apportionment Using a Hybrid Environmental Receptor Model," Environ. Sci. Technol., vol. 52, no.11, pp. 6357–6369, 2018, doi: 10.1021/acs.est.8b00131.

[18] G. L. Shi et al., "Combined source apportionment, using positive matrix factorization-chemical mass balance and principal component analysis/multiple linear regression-chemical mass balance models," Atmos. Environ., vol. 43, no. 18, pp. 2929–2937, 2009, doi: 10.1016/j.atmosenv.2009.02.054.

[19] G. L. Shi et al., "Estimated contributions and uncertainties of PCA/MLR-CMB results: Source apportionment for synthetic and ambient datasets,"Atmos. Environ., vol. 45, no. 17, pp. 2811–2819, 2011, doi: 10.1016/j.atmosenv.2011.03.007.

[20] USEPA, "EPA-CMB8.2 Users Manual," Off. Air Qual. Plan. Stand. Emiss. Monit. Anal. Div. Air Qual.Model. Group, US. Environ. Prot. Agency, p. 123, 2004.

[21] C. H. Mason and W. D. Perreault, "Collinearity IPower I and Interpretation of," vol. XXVIII, no. August, pp. 268–280, 1991.

[22] J. I. Daoud, "Multicollinearity and Regression Analysis," J. Phys. Conf. Ser., vol. 949, no. 1, 2018, doi: 10.1088/1742-6596/949/1/012009.

[23] D. H. Vu, K. M. Muttaqi, and A. P. Agalgaonkar, "A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables," Appl. Energy, vol. 140, pp. 385–394, 2015, doi: 10.1016/j.apenergy.2014.12.011.

[24] D. H. Lowenthal, R. C. Hanumara, K. A. Rahn, and L.A. Currie, "Effects of systematic error, estimates anduncertainties in chemical mass balance apportionments: Quail Roost II revisited," Atmos. Environ., vol. 21, no. 3, pp. 501–510, 1987, doi: 10.1016/0004-6981(87)90033-3.

[25] J. H. Kim, "Multicollinearity and misleading statistical results," Korean J. Anesthesiol., vol. 72, no. 6, pp. 558–569, 2019, doi:

10.4097/kja.19087.

[26] C. G. Thompson, R. S. Kim, A. M. Aloe, and B. J. Becker, “Extracting the Variance In flation Factor andOther Multicollinearity Diagnostics from Typical Regression Results,” Basic Appl. Soc. Psych., vol. 39, no. 2, pp. 81–90, 2017, doi: 10.1080/01973533.2016.1277529.