Utility of synthetic datasets to refine the particular matter source profiles for chemical mass balance analysis

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

[1]APJ Abdul Kalam Technological University, Thiruvananthapuram, 695 016, India

[2]Department of Civil Engineering, Rajiv Gandhi Institute of Technology, Kottayam, 686 501, India

Source apportionment of particulate matter samples using the chemical mass balance (CMB) approach is undertaken from a known set of information regarding chemical species present in the sources. The chemical species data pertaining to a particular source often would contain interfering chemical species originating from other sources due to the mixed state of its presence in the actual state. Such mixed source profiles, when used in the CMB analysis, result in collinearity and induce several errors in source apportionment analysis. To ensure the reliability of the source apportionment process, collinearity is to be eliminated by refining the source profiles used in the source apportionment process. The utility of synthetic datasets to refine the source profile in a source apportionment process is demonstrated for the particulate matter sampled from a non-urban region in the South Indian state of Kerala. To bring out cleaner source profiles, a large sample of receptor datasets that closely resemble the field data is generated using the artificial neural network module conditional tabular generative adversarial network(CT-GAN) and principal component analysis/multi linear regression(PCA/MLR) approaches are further applied to isolate various combined sources, and the source profile information is refined using published information. The effectiveness of the refined source profile information in a CMB analysis is demonstrated using the field receptor data.

Keywords: Chemical mass balance, collinearity, conditional generative adversarial network, principal component analysis.

random proportions, and it would not be easy to physically separate mixed components of such sources from arriving at cleaner source profiles. This natural mixing of various sources results in strong correlations across a few species. When such source profiles are used in the source apportionment process, they generate negative values of source contribution due to the collinearity existing among the sources. The presence of collinearity among the sources is mathematically estimated using statistical parameters of (MLR), such as variance inflation factor (VIF), tolerance value and condition index variables[3,4,5]. When the parametric values for the coefficient of the regression equation for the species, measured as the coefficient of determination (R^2), are obtained to be above 0.8, the VIF is above five, and the tolerance value is below 0.2 would indicate strong collinearity among the species considered. On the other hand, if the estimates for the coefficient of determination (R^2) and VIF are obtained as 0 and 1, respectively, then the collinearity across the chemical species can be ruled out. The implication of collinearity in the source apportionment process was demonstrated in the CMB model analysis undertaken on a particulate matter sample ($PM_{2.5}$). Sampled along a semi-urban road in the South Indian state of Kerala. In addition, specific source sampling exercises were also undertaken to create the source profiles for soil, paved road dust, wood fines, and traffic emissions, the sources that are identified to contribute to particulate matter in the air. Elemental species in all the sample sources of particulate matter were analysed using inductively coupled plasma mass spectrometry. The CMB analysis undertaken using the source profiles of soil, paved road dust, biomass, and traffic generated results having negative source contributions from paved road dust. Thus, it has become essential to employ appropriate corrections to eliminate the collinearity in the source profiles (Table 1).

The research is undertaken to generate source profile information that is free from any intervening species based on the detailed examination of large-scale synthetic receptor datasets that are generated artificially to represent the field data. Synthetic receptor datasets closely match both the behaviour and characteristics of the real dataset. The synthetic generation of receptor data is undertaken using a conditional generative adversarial network (CT-GAN), which is mathematically represented as[6,7].

$$V(D,G) = E_{x\sim p_{data}}[logD(x)+E_{Z\sim p_z}(Z)[log(1-D(G(z)))]]$$