



This is the accepted version of this journal article:

Mostert, Maria and Ayoko, Godwin and Kokot, Serge (2010)
Application of chemometrics to analysis of soil pollutants. Trends in
Analytical Chemistry, 29(5). pp. 430-445.

© Copyright 2010 Elsevier BV

Application of chemometrics to analysis of soil pollutants

Maria M.R. Mostert^{a‡}, Godwin A. Ayoko^{a,+} Serge Kokot^b

^aInternational Laboratory for Air Quality and Health, School of Physical and Chemical Sciences,, Queensland University of Technology, Brisbane, QLD 4001, Australia

^bInorganic Materials Research Program, School of Physical and Chemical Sciences. Queensland University of Technology, Brisbane, QLD 4001, Australia

[‡]Present address: Age Compilation Project, School of Earth Sciences, University of Queensland, Brisbane, QLD 4072, Australia

⁺Correspondence author: G.A.Ayoko, email: g.ayoko@qut.edu.au; Ph: +617 3138 2586; Fax:+61 7 3138 1804

Abstract

This overview focuses on the application of chemometrics techniques for the investigation of soils contaminated by polycyclic aromatic hydrocarbons (PAHs) and metals because these two important and very diverse groups of pollutants are ubiquitous in soils. The salient features of various studies carried out in the micro- and recreational environments of humans, are highlighted in the context of the various multivariate statistical techniques available across discipline boundaries that have been effectively used in soil studies. Particular attention is paid to techniques employed in the geosciences that may be effectively utilized for environmental soil studies; classical multivariate approaches that may be used in isolation or as complementary methods to these are also discussed. Chemometrics techniques widely applied in atmospheric studies for identifying sources of pollutants or for determining the importance of contaminant source contributions to a particular site, have seen little use in soil studies, but may be effectively employed in such investigations. Suitable programs are also available for suggesting mitigating measures in cases of soil contamination, and these are also considered.

Specific techniques reviewed include pattern recognition techniques such as Principal Components Analysis (PCA), Fuzzy Clustering (FC) and Cluster Analysis (CA); geostatistical tools include variograms, Geographical Information Systems (GIS), contour mapping and kriging; source identification and contribution estimation methods reviewed include Positive Matrix Factorisation (PMF), and Principal Component Analysis on Absolute Principal Component Scores (PCA/APCS). Mitigating measures to limit or eliminate pollutant sources may be suggested through the use of ranking analysis and multi criteria decision making methods (MCDM). These methods are mainly represented in this review by

studies employing the Preference Ranking Organisation Method for Enrichment Evaluation (PROMETHEE) and its associated graphic output, Geometrical Analysis for Interactive Aid (GAIA).

Keywords: Soil studies; chemometrics; multivariate analysis; geostatistics; source identification; source contributions; PCA/APCS; PROMETHEE; GAIA.

Abbreviations and acronyms

CA	Cluster Analysis
FC	Fuzzy Clustering
FKA	Factorial Kriging Analysis
GAIA	Geometrical Analysis for Interactive Aid
GIS	Geographical Information Systems
HMW	High Molecular Weight
LDA	Linear Discriminant Analysis
LMW	Low Molecular Weight
MCDM	Multi Criteria Decision Making
PAH	Polycyclic Aromatic Hydrocarbons
PCA	Principal Component Analysis
PCA/APCS	Principal Component Analysis/Absolute Principal Component Scores
PMF	Positive Matrix Factorization
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluation

1. Introduction.

Pollutants reach surface soils through atmospheric deposition of suspended particles [1-8], deposition of contaminated sediments [9-13] and the circulation of groundwater [13-17]. In general, such substances can be inhaled either as re-suspended particulates in their natural state or as adsorbates on dust or soil particles [18].

In recent years, there has been a number of investigations of soil contamination within cities, recreational spaces and in children's environments [3, 5, 19-25]. Many of these studies have been undertaken partly because soil ingestion has been recognised as an important exposure route for contaminants to humans,

especially infants [18, 26, 27]. For example, some toxic pollutants such as lead [26] and arsenic [28] are known to be more likely to be ingested than inhaled. Therefore, soil pollution continues to be an important area of study for environmental scientists. However, apart from a few early studies [29, 30] the application of multivariate data analysis in such studies has been fairly limited to date. Many investigators preferred classical univariate statistics for processing data [19, 20, 25, 31, 32]. Chemometrics techniques such as Principal Components Analysis (PCA), Linear Discriminant Analysis (LDA) and the various source apportionment techniques have not found significant usage [7, 8, 12, 33, 34]. However, cluster analysis [7, 22, 23, 26, 34] has been employed with some effect. Recently, a few studies have appeared in which some apportionment techniques have been explored [18, 33, 35-37], but source contributions are rarely studied [36, 38], and multicriteria decision making methods (MCDM) are used even more rarely [38]. Because many parameters are usually measured during soil analysis, multivariate data analysis techniques such as those involved in the chemometrics procedures reviewed in this paper are more likely than univariate methods to explain the underlying structures of the data.

On the other hand, specialised geostatistical tools for the geosciences do exist and are widely employed in the mining industry [39]. These have found some application in the study of soils [18, 35, 40-46]. However, many environmental scientists appear to be less familiar with such tools and their applications.

This overview focuses on the application of chemometrics techniques for the investigation of soils contaminated by polycyclic aromatic hydrocarbons (PAHs) and metals because these two important and very diverse groups of pollutants [31, 47], [48, 49] are ubiquitous in soils [29, 37]. The approach employed in the review is to highlight the salient features of studies carried out in the micro- and recreational environments of humans, in the context of the various multivariate statistical techniques available across discipline boundaries that have been effectively used in soil studies.

2. Pollutants and Related Factors

2.1. Occurrence and Impact

Although the focus of this review is the application of chemometrics to soil pollutants, it is important to provide contextual framework for the ensuing discussion. Therefore, we have outlined the occurrence of the two broad classes of pollutants targeted in the review as well as their human effects and issues such as sampling, which are important when one is considering how to generate meaningful data from soil analysis.

2.1.1. Metals

Metals are found in all soils. It is the concentration levels which present the risk, and it is the mechanism of enrichment that defines the enriched environment as “polluted”. The earth’s crust consists of metals

combined with various non-metallic elements, mainly oxygen and silicon and they are not equally distributed over the surface of the earth. The unequal distribution of metals in the crust gives rise to exceedingly high levels in some localities, which can be redistributed by stormwater run-off [10, 50] and circulating groundwater [13, 14, 17, 28] with dire consequences for humans. Soil contamination by metals may occur as a result of various mechanisms. Worldwide, the most important sources of metal pollution are mine tailings, smelter emissions, waste incineration and atmospheric deposition [4, 5, 9]. But the main contribution to the trace element load in urban topsoils and dust is almost unanimously considered to be traffic emissions [25] derived from atmospheric deposition.

Interest in the metal content of soils stems from the fact that certain metals are essential for human health, for example, Ca, Fe, Mg, Zn and Cu [51]. However, some of the elements, which are essential for human health, may be injurious when present in excess. For instance, excessive amounts of Fe, Cu, Zn, Co, Mg, Cr and Se are known to have adverse human effects [17]. Furthermore, many metals such as Cd, Pb and Hg [15] have no known biological importance in human biochemistry and physiology and can have adverse human effect even at low concentrations.

2.1.2 PAHs

PAHs are ubiquitous contaminants in both the general environment and in certain working environments [49] and are deposited on soils through atmospheric fallout. They are generally produced in incomplete combustion processes [52-54], and the composition of PAHs emitted from such processes is dependent on a variety of factors, such as the fuel and its properties, as well as atmospheric oxidation and degradation [49]. Although PAHs are produced in volcanic activity and in forest fires [3, 54, 55], and are also present in petroleum and coke products, their primary source is anthropogenic activities [3, 49, 53, 54].

PAHs constitute a wide class of compounds consisting of two or more fused benzene rings. The conceivable number of different PAHs with eight aromatic rings is more than one thousand [56]. A large number of structural isomers exist, as well as nitro- and other substituted compounds. PAHs are highly lipophilic compounds with low chemical reactivity [49]. They occur in the environment as complex mixtures of many components with widely varying toxic potencies [49, 52, 56]. PAHs bind strongly to soils, especially the organic components and are difficult to dissipate, due to their octanol-air fraction coefficient [3, 8]. Soils thus act as sinks for PAHs, where they tend to persist for a long time [3, 8, 55]. They are specifically strongly adsorbed to the organic fractions of sediments and soil [3, 52]. Consequently, soil is one of the main reservoirs of PAHs in the environment [8, 57].

There have been increased interest in the profiles of PAHs in soils because these compounds have mutagenic, carcinogenic and immunotoxic properties [8, 58-60]. Animal studies have shown that orally

ingested PAHs cause cancers of the stomach and gastrointestinal tract [60]. The most active compounds are those composed of four to seven rings [49, 54]. The USEPA has classified seven PAHs as probable human carcinogens: benzo[*a*]pyrene (BAP), benz[*a*]anthracene (BAA), chrysene (CHR), benzo[*b*]fluoranthene (BBF), benzo[*k*]fluoranthene (BBK), dibenz[*a,h*]anthracene (DBA), and indeno[1,2,3-*cd*]pyrene (IND) [61]. In addition to these, a further 9 are regarded as priority PAHs: naphthalene (NAP), anthracene (ANT), fluorene (FLU), fluoranthene (FLT), acenaphthylene (ACY), acenaphthene (ACE), phenanthrene (PHE), pyrene (PYR) and benzo[*g,h,i*]perylene (BGP) [3].

Studies concerned with the presence of these PAHs will feature in this review. The best-characterised source for PAHs is vehicle emissions followed by wood burning [62, 63]. The composition of PAHs in combustion products depends on the combustion temperature. At low to moderate temperature, as in wood or coal stoves, low molecular weight PAH compounds are abundant, whereas at higher temperatures, such as those in vehicular emissions, higher molecular weight compounds abound [63].

It has been difficult to make accurate estimations of PAH emissions because it is not easy to quantitatively determine how much of any particular PAH has been contributed by any specific source. Residential heating, automobile exhaust, industrial power generation, incinerators and production of coal, coke and asphalt produce complex mixtures of PAHs. Some PAHs are common to a number of these sources [49, 62]. Source apportionment is therefore an important consideration in the study of PAHs in the environment in general and in soils in particular.

2.1.2. Sampling issues.

A number of issues impact on meaningful data analysis of soil samples. Of these, sampling depth, sampling density [35, 39, 57] and the analytical techniques employed [57, 64] are the most important in ensuring that the results are reliable and amenable to multivariate analysis. Soil characteristics such as organic content and stratification and external factors such as climate [65] further complicate investigations and obscure observable trends and comparison of different studies [57]. Necessitated by the requirement for economic feasibility and accurate projections of mine lifetimes and productive capacity [39], the earth sciences have developed rigorous sampling procedures to ensure that accurate and representative images of the sampled body are obtained [13, 39]. Exploration and sample preparation for metals is thus fairly well understood and universally applied in the geosciences. However, these techniques are not widely employed by environmental scientists seeking to determine soil pollution by metal species [34, 35].

Sampling and sample preparation for the determination of organic soil pollutants is confounded by various factors [57]. Sampling site, sampling time, sampling depth, sampling support, extraction

techniques and the sum calculation of organic species all hamper the comparability between different studies. A multitude of analytical techniques and the effect of soil stratification, land use and climatic conditions further complicate matters [57]. To date, the problem of how to quantify reliably the bias of soil sampling is not considered really solved [57]. It has been recognised that there is an urgent need for harmonisation of soil monitoring within the European Community, since standardisation is dependent on harmonisation, and biased standardisation is misleading not only for sampling, but for the whole measurement chain [57] and consequently also for the subsequent data analysis.

From the above, it is evident that sampling and sample preparation procedures are vitally important for the derivation of results amenable to data interpretation. Environmental studies may greatly benefit by applying soil sampling techniques as developed and used in the geosciences; additionally, great care must be taken with sample preparation, especially with extraction techniques [57].

3. Soil pollutants and multivariate techniques

In recent decades, it has become routine for large amounts of data to be generated with the use of automated analysis. Univariate data analysis is no longer sufficient for exploring fully the complexity of the vast amounts of information available [66]. Therefore, it has become general practice to employ various multivariate techniques for pattern recognition, classification and prediction for analysis of large datasets [66-69]. The common methods are summarised in Table 1 and a number of these are outlined in the Appendix.

It is usually important for environmental investigators to determine one or more of the following parameters when studying pollutant species:

1. Distribution of pollutants and associated factors, such as background values and physico-chemical factors (for instance pH); the extent of affected areas; the nature of the distributing agent, whether by atmospheric wet (rain) or dry (dust) deposition and the nature of transportation – whether in vapour or particulate phase, circulation of groundwater, or stormwater run-off; and, the distinction between different populations of pollutants.
2. Identification and apportionment of sources – where do the contaminants originate?; and, determining source contributions – how much does each source contribute to the contamination?
3. Mitigating measures and their effectiveness.

This overview is not intended to be exhaustive because its intent is to show the reader how chemometrics can enhance the interpretation of data from soil pollutants to improve knowledge outcomes. Although the review is not chronological, it does acknowledge the researchers who applied chemometrics to soil

science. The studies referred to in this overview are therefore used to focus on and illustrate the application of various chemometrics techniques as applied to soil studies.

3.1. Pollutant distribution and chemometrics

3.1.1. Geostatistics, GIS mapping and chemometrics

Depending on the aims of a particular study there are a number of approaches that are available for the interpretation of pollutant distribution data. Perhaps the most visually satisfying and direct method is by generating a contour map [6, 18, 33, 35-37, 41] displaying spatial variability of a particular chemical species. This type of approach is usually suitable for regional studies where a large quantity of soil data has been gathered in a systematic way. In this context, geostatistical tools such as inverse-distance squared interpolations, variograms and kriging are the most applicable [18, 33, 35, 36].

Soil data are gathered from predetermined sampling points situated on a map grid, and values between points and missing values are interpolated. The contour diagram can then be generated using Geographical Information Systems (GIS) [18, 35, 70]. The reader is referred to Wellmer [39] and Goovaerts [41-43] for a discussion on the theory and practice of geostatistical interpolatory methods and to studies by Mielke et al [31, 32], Zhou et al [12], Zheng, Chen and He [40], Zhang et al [33] and Facchinelli, Sacchi and Mallen [45] for examples where this technique was used to study metal enrichment, and Wang et al [6] where it was used for studying PAH spatial distribution.

Mielke and co-workers [31, 32] used kriging (geostatistics) to plot a contour diagram of multiple metal accumulation in central New Orleans, United States, and compared the results with Louisiana Educational Assessment Program (LEAP) scores (scholastic achievement). With the use of correlation analysis, they were able to show that there was a strong inverse relationship between learning achievement of students attending schools in a particular community and the quantity of multiple metals accumulated in the soil environments of that community. This was particularly valid for Pb [32]. It was also demonstrated that the most disadvantaged communities were more adversely affected by soil Pb than other communities, since, according to census data, most of the members of those communities were living in areas with elevated soil Pb values [32]. The study did not attempt to identify pollutant sources.

Facchinelli, Sacchi and Mallen [45] performed a classic study in the Piemonte region of Italy. Using geostatistics and GIS software, they plotted various contour maps of metal occurrences on a regional scale. They also indicated highways, rock exposures of various types and cultivated areas. With the aid of Principal Components Analysis (PCA) and matrix rotation they were able to identify the main metal contributors to both the upper and lower soil levels as natural parent rocks, agricultural activity (mainly

viticulture) and transportation/industry, in that order. Cluster Analysis (CA) confirmed these findings. Cluster Analysis is often used to confirm the results of PCA in other studies [7, 45].

Wang, Qin and Sang [7] studied the accumulation and sources of heavy metals in the industrial city of Xuzhou in China. Three main sources were indicated by PCA after rotation of matrices. These were the natural background, traffic, and coal-burning activities, in that order. Again, the results from PCA were confirmed by CA [7]. Contour maps were also used to effectively display the pollution levels derived from two possible sources. Similarly, Zheng, Chen and He [40] used semivariograms (a geostatistical technique, which is more recently known simply as variograms) to derive maps depicting spatial variability of different metals in topsoils from Beijing, and PCA and partial correlation analysis to attempt to identify sources.

Van Meirvenne and Goovaerts [44] proposed a scheme for estimating the probability that a particular site may exceed the regulatory limit for any contaminant, using soft indicator kriging, a specialised application of kriging analysis. As they pointed out, sampling campaigns may be very costly [35, 44] and the proposed technique can be used as a tool for improving the cost-effectiveness of a sampling regime. Other researchers, too, saw the need for such techniques. For example, Tao Chen et al attempted to develop a methodology for limiting the number of repeat samplings necessary for adequate environmental monitoring using co-kriging and ArcGIS [35].

Geostatistical tools have also been used successfully for the interpretation of spatial distribution of PAHs. Wang et al [6] studied distribution patterns of PAHs in topsoils of the Tianjin area of China using a specialised kriging procedure, factorial kriging analysis (FKA), to study various soil parameters, including total organic carbon (TOC) and pH, at varying spatial scales. FKA allowed the successful interpretation of the distribution patterns of individual PAHs by the construction of contour maps, but failed to deliver clear spatial distribution patterns for the combined group of the 16 high-priority PAHs (Σ PAH16) as defined by USEPA. To enable the investigators to study the distribution of the total PAHs, the results were subjected to PCA and three latent variables extracted; high molecular weight (HMW) PAHs, low molecular weight (LMW) PAHs and anthracene on PC1, PC2 and PC3 respectively. Using these three variables, successful contour maps could again be constructed, leading to useful conclusions concerning possible sources and pollution patterns arising as a consequence.

Yu et al [71] and Zhang et al [33] studied PAHs and metals respectively, using GIS techniques to draw contour maps, but employed multivariate statistics (PCA with Varimax rotation) to identify sources. These two studies and that of Wang et al [6] mentioned above, show that, although geostatistical techniques can be usefully applied in soil studies, classical multivariate techniques are still often needed for a more focused approach.

Yu et al [71] investigated PAH pollution in Guiyu city, Guangdong. The aim of the study was to investigate the distribution, concentration, profile and source of PAHs in the Guiyu soil affected by E-waste recycling. The investigators constructed a contour map using GIS technology, but the most useful results were obtained by the application of multivariate techniques (PCA, variance analysis and Pearson correlation) on the total PAH content for each sampling site. After PCA, the Σ PAHs reverted to three latent variables related to molecular weight, which, again, could be related to particular sources [71]. Zhang et al [33] performed a similar study in the Fuyang valley of Zhejiang Province in East China. This area has had intensive copper and zinc smelting activity for over fifty years and previous geostatistical studies failed to distinguish between pollutant sources. Using PCA, they isolated four principal components (factors) which explained 90% of the variance in the data. Three of these could be related to smelting, coal combustion, and the bedrock underlying the soil; a fourth was less clearly defined. Soil physicochemical properties was thought to be an important factor in obscuring clear identification of pollutant sources [33]. Most of the above studies sought to identify the most important sources of pollution, but did not attempt to calculate the actual pollutant contributions from each source. Such an analysis could be indicative of the actual impact of each source on pollutant concentrations at the receptor site and could provide very useful information when considering mitigating measures. Other techniques which would have been extremely beneficial for evaluating the impact of pollutants and their sources are the various multicriteria decision making methods (MCDM). These are discussed in greater detail below (see Section 3.3).

3.1.2. Distribution studies without geostatistics.

As explained above, geostatistical methods require extensive sampling and are usually applied on a regional scale [34]. This can be a very costly exercise [35, 44]. Furthermore, in some cases, it may not be possible or useful to construct maps for a particular study [38]. Multivariate analytical techniques are designed for the interpretation of multivariate datasets on soils even without the assistance of geostatistics.

In general, PCA is the workhorse for the multivariate analysis of complex soil information. Almost all of the examples already cited in Section 3.1.1 employed this basic tool to expand on results obtained from various geostatistical techniques. PCA can be combined with other multivariate techniques as well. CA is often used to confirm findings by PCA [1, 7, 34, 45]. PROMETHEE and GAIA are specialized forms of PCA which have also been used for evaluating pollutants in soils and sediments [10, 38, 50]. It has been shown that the use of one chemometrics techniques alone on soil data can still lead to extremely useful information on the quality of the soil [1, 22, 23, 34, 66].

Soil contamination does not occur only on the surface but also at depth [33, 34]. Metals are mobilised and transported through soil profiles by different mechanisms [17] and consequently contamination profiles may be quite complex. Sielaff and Einax [34] used a number of multivariate strategies to study three-dimensional relationships between elements, soil features and parent materials in a study of the polluted surroundings of a former cement and phosphate fertilizer factory. They used, inter alia, cluster analysis (CA), linear discriminant analysis (LDA) and PCA with varimax rotation to evaluate a high-dimensional data set containing the results of three-dimensional soil sampling. Three soil profiles were sampled in the vicinity of a former fertilizer plant, with the view to investigate whether previously identified pollutants were mobilized and moved vertically within the soil profile. To this end, a few lateral sampling locations were chosen, and many vertical samples were taken from these locations. Univariate analysis merely revealed that the Cd content exceeded the limiting value. With cluster analysis (CA), however, it was possible to detect enrichment in Ca, Cd, Cu, F, Na, P, Pb and Zn some 10 years after cessation of activities at the fertilizer plant. CA combined with linear discriminant analysis (LDA) differentiated between soil samples according to parent material and pollution state, and PCA with varimax rotation could discriminate between lateral sample location and vertical depth. The results of the multivariate analysis were confirmed by the application of sequential extraction [34].

An interesting and quite unique three-dimensional study of geological cores [66] using chemometrics techniques involved cores taken from six playa* salt lakes in central Australia and analysed for eight different elements using the XRF technique. The dataset was subjected to PCA and Fuzzy Clustering. Fuzzy Clustering is an unsupervised technique for classification of objects; that is, it separates objects in a dataset into a number of classes nominated by the user, but without prior training of the model. The resulting distribution pattern of the data objects from the playa lakes clearly indicated the presence of three classes of objects, which could be interpreted sensibly both chemically and geologically as having been derived from sand, clay and evaporites [66]. This study did not involve anthropogenic contamination by heavy metals, but is mentioned here for the singular use of a particular chemometrics technique as applied to three-dimensional soil samples, the application of which may be useful in environmental soil studies.

Chen et al [22] studied soil quality as it relates to trace metal content in the surface soils of Hong Kong with the help of CA. They showed that urban and orchard soils were most enriched in anthropogenic trace metals while rural and forest soils were least enriched. The high content of anthropogenic trace metals in orchard soils were ascribed to agricultural applications of pesticides, fertilizers and animal manure. This finding is similar to that of Facchinelli et al [45] and Navarro et al [11]. The latter evaluated historical

* A playa lake or pan is a dry lake bed consisting of alkali salts or common salt intermingled with fine sediments.

soil and sediment data gathered from 1996 to 2003 from the Ebro river basin in northern Spain. The chemical species under investigation included PAHs and various organochlorines (OCs). With the help of PCA and correlation analysis, three contamination profiles were identified which persisted over a number of years. One of these appeared to be associated with chemicals used in agriculture [11].

Conversely, Carlosena et al [1] confirmed Pb as the most distinctive metal arising from road traffic pollution in La Coruña, NW Spain. This early study made extensive use of CA and PCA. In a later study, Ljung et al [26] mainly applied CA when evaluating multiple metal accumulations (MMA) in children's playgrounds in Uppsala, Sweden. Chen et al [23] also investigated heavy metal pollution in urban parks of Beijing employing PCA and CA and demonstrated that historic parks (those older than 100 years) exhibited greater accumulation of anthropogenic heavy metals than younger parks. Wang et al [55] found support for the theory that PAHs may be distributed by a distillation-type process whereby lower molecular weight varieties were transported atmospherically over greater distances than higher molecular weight PAHs in Dalian, China. They primarily made use of CA to study the distribution and extent of these pollutants [55].

The studies cited above were all performed without the aid of geostatistics but still provided greatly expanded information on the nature of various forms of contamination, in a variety of scenarios. The expensive and exact sampling regimes and map construction required for the application of geostatistics was thus obviated. Most of these studies would, however, have benefited from source contribution investigations and MCDM analysis to estimate the detrimental impact of specific sources and suggest possible mitigating measures.

3.2. Identification of pollutant sources, their contributions, and chemometrics

Inherent in the study of pollutants is the attempt to identify the sources responsible for certain distribution patterns. To date, PCA has been the most commonly used multivariate tool for this purpose in soil studies [2, 6, 7, 11, 34, 45, 66, 71] and has been mainly applied for the purpose of identification. However, it is also possible to determine quantitatively the loading of each variable on each source, and the contribution of that source to the total pollutant concentration.

Factor Analysis (FA) has been applied as a statistical technique to identify a relatively small number of factors (latent variables) that can be used to represent the actual sources of contamination [14]. Three common methods source apportionment techniques that are based on factor analysis are: Positive Matrix Factorisation (PMF) [72-74], UNMIX [73], and PCA with Multiple Linear Regression Analysis (PCA/MLR), which is also known as Principal Component Analysis/Absolute Principal Component

Scores (PCA/APCS) [14, 63, 73, 75, 76]. Of these methods, PCA/APCS has been the technique most widely used for soil studies [2, 12, 14, 33, 36, 38, 74, 75].

PCA/APCS requires an initial PCA to be performed on the standardized data [73, 74, 76]. The PC matrix is then rotated using varimax rotation, which retains the orthogonality of the axes and hence the independence of the latent variables (or sources) [2, 63, 74, 76]. The rotation of the matrix realigns the matrix axes with the adjusted model parameters after the removal of the non-significant variables and thus clarifies the variable loading on each source [33]. Once this step has been completed, it is possible to identify the individual sources based on their chemical profiles. These chemical profiles may be defined variously by mathematical procedures [77], literature values [2, 62, 75, 77, 78] or diagnostic ratios [2, 3, 55, 62, 75, 77, 79]. Topographical [11, 45] or underlying rock features [14, 34, 38, 45, 66] are also strong indicators of possible sources, as are various human activities [11, 14, 22, 45, 75, 79, 80].

The application of Principal Component Analysis (PCA) to environmental data is associated with significant setbacks because its outcomes are correlated with but not proportional to source contributions Anderson et al (2002) [81]. Consequently, PCA results cannot be used directly for source apportionment. However, when PCA is coupled with absolute principal component scores or multi-linear regression, it becomes a powerful tool for source identification and source apportionment [12, 14, 38, 63, 74, 75]. Therefore, source apportionment studies in soil analysis utilizing conventional PCA have been quite limited.

Pujari and Deshpande [14] conducted a study on groundwater contamination in the vicinity of a landfill site in Nagpur, India. Strictly speaking, the study was concerned with water quality, but since the source of water contamination was directly dependent on the soil quality, this study may be used as an illustration of the application of PCA/APCS. The authors identified five factors contributing to chemical species in the groundwater. The interplay of the landfill, sanitary practices and parent rocks on the water quality was illustrated and the danger of the contamination of groundwater leaching through an unconfined aquifer in such an environment was demonstrated [14]. Other examples of the application of PCA/APCS are discussed below.

Zhou, Guo and Liu [12] conducted a large study on the marine sediments of Hong Kong harbours, incorporating many important procedures for the evaluation of the anthropogenic impact on sediment quality. Particular care was taken to distinguish between the natural contribution from parent materials and the man-made enrichment of heavy metals by appropriate pretreatment of data and the application of suitable Enrichment Factors (EFs) [12]. Three pollutant sources were identified by PCA with varimax rotation. The most important factor was derived from industrial runoff mainly caused by the electronics industries, riparian runoff and vehicle exhaust, it affected the most confined waterways; the second factor

appeared to be derived from a far wider area and was considered to arise from the weathering of underlying rock. The third factor could be linked to effluent discharge from textile factories and paint, and affected the confined southern harbours [12]. The contributions from each of these sources were then calculated using PCA/APCS. Estimates were cross-checked using observed values and were found to have uncertainties of generally less than 8%. This study addresses many of the issues and difficulties inherent in the statistical analysis of heavy metal pollution and is a valuable background source for similar studies.

Mostert [38] investigated the quality of children's playgrounds in south-east Queensland, Australia, using PCA with Varimax rotation for identifying sources of metals and PAHs, and estimating the contribution of each source with PCA/APCS. It was found that the main source of the metal content in playground surfaces was the natural background derived from underlying rock formations and from dissolved salts in the marine air associated with the proximally-situated ocean, while vehicular emissions contributed the majority of the PAHs.

Wang et al [8] used a slightly different procedure, FA with non-negative constraints, for determining sources of PAH pollution in Dalian, China, followed by MLR. Conventional FA can return negative values after matrix factorization, which, in practice, loses physical meaning. FA with non-negative constraints avoids this problem, but suffers from the disadvantage that factor axes are no longer orthogonal [8]; thus, the sources can no longer be regarded as independent.

Zhang et al [33] carried out a study of an area in eastern China with a long history of metallurgical activity, containing a number of clearly identifiable potential sources of soil pollutants. The aim of the study was to examine the possible problems involved in source identification of soil inorganic pollutants using a multivariate statistical approach. PCA with Varimax rotation identified four main sources of heavy metal pollution; metallurgy, coal combustion, "chemical mobility" and parent materials, in that order. Unfortunately, perhaps because of the exploratory nature of the study, the authors did not attempt to calculate the source contributions.

It is evident from the above discussion that source apportionment and receptor modelling is an underutilised technique in soil studies. It is therefore suggested that this methodology be far more widely applied as has already been the case in atmospheric studies [63, 72, 73, 75, 76, 78, 82]. In this context, it is important to note that, in recent years, the PMF receptor model, which was developed by Paatero and his group (Paatero and Tapper, 1993)[83] and used frequently by Hopke and his co-workers (Hopke (2003), Kim et al. (2001), Kim, Hopke and Edgerton (2004) and Ramadan et al. (2003) [68, 82, 84, 85] to identify and quantify sources of air pollutants have found increasing application in soil analysis. For example, Wang et al. (2009)[86] applied the technique to PAH data derived from soil analysis. The PAH

sources showed markedly different source contributions during summer and winter. Thus the three dominant sources, with percentage contribution in parenthesis during the winter were: coal-fired engine (72%), traffic average (20%) and gasoline (8%) while the corresponding summer sources were: coal combustion (46%), diesel engine (30%) and gasoline (24%). (The dominant PAH in each of the summer sources was: retene, phenanthrene and benzo (g,h,i) perylene respectively.

Vaccaro et al (2007) [87] measured pH and a suite of major and minor elements including, Hg, C, H, N, Si, Ca, K, Fe, Mg, Ti, S, P, Pb, Zn, Cu, Ni, Mn, Cr, Na, Al, V, Co, As, Cd, and used PMF to identify their sources in soil samples. The identified sources were: a contamination source (indicated by the presence of Mn, As, V, Fe, Ni and Co), a polluting source rich in Hg, Pb, Zn, Cu, P and Cd, and an organic/biogenic source, evident from the predominance of C, H, N, S in the source profile and a source consisting of Si, Ca, K, Fe, Mg, Ti and Al, which reflected the local geo-morphology and typical minerals in the area. By combining the results from the PMF analyses with GIS-based geo-statistical analyses, the authors showed the potential use of this approach in the interpretation of soil polluting sources.

Lu et al (2008) [88] also applied PMF to compositional data matrix consisting of 24 soil properties- 8 physical, 13 chemical and 3 biological properties. Two factors were required to model the original data satisfactorily. The first was characterised by coarse-textured slightly acidic and physically loose soil while the second consisted of fine-textured, slightly acidic and physically compact soil. The authors showed that the results obtained with PMF for this dataset was physically more meaningful than those obtained with comparable models such as PCA and Target transform analysis (TFA).

Apart from PCA/APCS and PMF, another receptor model used in literature to source identification and source apportionments of elements involved the application of target transform factor analysis method to urban road dust. Since the majority of the elements measured originated from soil, the study is reviewed briefly in the current work. Hopke et al (1980) [29] measured 35 elements in 30 subsamples of road dust and analysed the results using target transformation factor analysis. The elements include: Sb, As, Ba, Br, Ca, Cd, Ce, Cs, Cr, Co, Dy, Eu, Ga, Hf, Fe, La, Pb, Lu, Mn, Hg, Ni, K, Rb, Sm, Sc, Se, Ag, Na, Sr, Tb, Th, U, Yb, Zn and Zr, and the primary sources and source contributions of the elements in the dust samples are: soil (76%), cement (5%), rust (7.7%), tire wear (7.2%), automobile exhaust emissions (1.5%) and salt (0.3%).

3.3. Mitigating measures and chemometrics.

Much environmental work in soils is primarily involved with monitoring the environment [3, 11, 26, 31, 32, 45, 57, 67, 77, 89-91]. The ultimate aim, however, is to alleviate or eliminate the effects of pollution on the environment and on human health. When the sources have been identified and their contributions

estimated, it may be possible to either eliminate or modify the source so that the emission of pollutants is minimised. This has been the approach of environmental agencies and the intent of regulatory measures all over the world [52, 54, 61, 92-98].

Multivariate tools may, however, be used to craft a more focused strategy, especially if the pollutant sources are dispersed and pervasive [11, 45]. Consider, for instance, the problem presented to municipalities with cities established on parent rock that are enriched in metals, and the necessity of providing safe environments for human recreation [25, 26, 28, 38]. Municipalities have approached this issue in different ways. Some do not attempt to control the soil quality by replacement or other means [26], others cover recreational areas with sand, which is regularly replaced [25] and others still, dig up underlying soils and fill the space with bark chips [38]. These are situations in which municipalities and city councils as the decision makers could well use multivariate analysis to assist them in effective planning. Ranking Analysis (also known as Multi Criteria Decision Making, or MCDM) is a suitable tool for such a purpose. MCDM allows the user to model data by selecting, weighting and optimizing parameters in order to reach appropriate outcomes for addressing difficult scenarios and to present a variety of possible solutions. By using MCDM in the above example, it could be shown that the procedure of removing soil and replacing it by bark reduced the metal content of recreational areas in the most effective way of all, and that increasing the moisture content of the surface by irrigation or similar measures would be highly beneficial as well [38].

MCDM is as yet virtually unknown in soil studies, although it has been used for atmospheric studies [50, 67, 80, 99] and studies in fuel efficiency and exhaust emissions [99-101]. It has been applied, for instance, to the study of urban sediments resulting from stormwater runoff [10, 50], and also, to highlight existing good practices in the management of recreational spaces [38].

A number of MCDM models are available. Some common methods include *elimination et choix traduisant la réalité* (ELECTRE), simple multiattribute ranking technique (SMART), and preference ranking organisation method for enrichment evaluation (PROMETHEE), Hasse diagram technique (HDT), multiple attribute utility theory (MAUT), and analytical hierarchy process (AHP) [69]. The PROMETHEE program and GAIA, its graphic, PCA-type output program, is considered to be one of the most robust and generally applicable [69, 102, 103] MCDM methods. This model has been used in studies specifically involving soils, sediments and groundcover in urban environments [10, 38, 50].

Herngren and his co-workers [10, 50] used the PROMETHEE and GAIA methods to study pollutant loading on urban sediments in south-east Queensland, Australia. They were able to show correlations between particle size, metal accumulation and organic content and propose methods for dealing with the build-up of pollutants in urban environments., Using the same methods, Mostert [38], was able to show

that measures enacted by the owners of public recreational areas were effective since treated surfaces showed a much diminished metal content in comparison to the metal content of the natural background.

MCDM methods are still very limited in their use, and this is perhaps because, as multivariate techniques, they are still fairly unknown. They are, however, powerful tools that can provide practical guidance for compromise solutions where there are many considerations, which may involve ethical and subjective issues as well as scientific findings [69]. This is well demonstrated by results from the use of the PROMETHEE and GAIA combination [10, 38, 50].

PROMETHEE is a non-parametric method of data analysis whereas PCA is a parametric method, which is therefore constrained by rules of normal distribution. The principal differences between PROMETHEE/GAIA and PCA are illustrated in Table 1 and 2 and in Figures 1 - 3. Table 3 is the data matrix used for generating Figures 1 – 3. Table 3 shows the PAH concentrations of soil samples from sites S1-S15, while Figures 1, 2 and 3 are respectively, the PCA biplot obtained from the standardised data, the PROMETHEE ranking result obtained without data pre-treatment and the GAIA biplot. As can be seen from Figure 1, the PCA shows that there are broadly two clusters of site, which are separated from each other on PC1. One set (e.g.S11 and S5) has high positive PC1 scores, while the other, generally has low positive or negative PC1 scores. In addition, the vectors for BgP, BAP, and Chr are oriented in the same direction suggesting that they provide similar information about the sites. Similar conclusions can be drawn from the GAIA biplot (Figure 3). However, the result of the GAIA plot differs from that of the PCA biplot in two major ways. Firstly, S11 and S5 have negative PC1 scores because the variables used for the plot were minimised (i.e.sites with lower concentrations of PAHs were preferred over those with higher concentrations). Secondly, the GAIA biplot had an additional vector, π - the decision axis, which is oriented in the direction of the best ranked sites (S6 and S8) and opposite to the least ranked sites (S11 and S5). It is also noteworthy that the PROMETHEE results (Figure 2) reinforced this observation (sites S5 and S11 ranked last with almost identical Φ values while S13 and S6 ranked first with almost similar Φ values). Thus, PROMETHEE and GAIA provided information that can be used for pattern recognition and ranking analysis. Such information aids the decision making processes.

4. Concluding Remarks.

Soil studies have hitherto made only limited use of various chemometrics techniques to enhance results obtained by classical univariate analysis. Where multivariate techniques have been employed, geostatistical tools have been preferred, although good results have been obtained using common methods, of which PCA and CA proved to be the most widely used. Generally, chemometrics methods

have been used to study distribution of pollutants and identification of sources; little has been done on the quantification of source contributions, although it has been proven to be entirely possible to make such estimates after PCA analysis and matrix rotation or PMF analysis [12, 14, 38, 73]. Other versatile methods are also available [73, 84, 85] and could be explored with valuable outcomes. As clearly demonstrated by the few soil related studies PROMETHEE and GAIA are very useful and informative techniques; as exemplified by the ever increasing applications in the allied fields of water and air quality investigations. The software is readily available (<http://www.visualdecision.com>) and would be of considerable benefit for modelling solutions to complex problems where a variety of considerations impact on the ultimate decision. These are methods that are well suited for application in environmental studies of soil pollutants. Methods discussed briefly in the review are more fully explained in the Appendix below.

APPENDIX

Geostatistics.

Geostatistics considers the spatial dependence of samples or analytical measurements on each other. All geostatistical tools such as variograms, inverse-square interpolation, kriging and GIS technology are designed to construct maps. In the context of environmental analysis, these maps are contour maps of contaminant concentration. Viable maps for such a purpose can only be created if a large quantity of soil data is available, for which the geographical position of each data point is accurately known. To obtain such information, a paper map must either be constructed and each data point correctly represented thereon (a tedious and specialised task which requires surveying equipment that is seldom used outside the discipline of geology or surveying), or else, the data point coordinates (latitude/longitude) can be located using modern global positioning instrumentation. The coordinates must then be plotted onto a digital map, and the concentration contours of the chemical species being investigated may then be generated by suitable software (MapInfo or ArcGIS). The software is designed using the principles inherent in the geostatistical techniques mentioned above.

The latter is the methodology widely applied in the environmental studies discussed in section 3.1.1 above. If the money for a large sampling program is available, as well as the necessary equipment and software, this is a most effective visual tool and therefore very popular. There are, however, drawbacks to the methodology quite apart from the cost and required equipment. Contour maps can only model one chemical species at a time, or a summed group of compounds. It is therefore often still necessary to use some form of multivariate analysis to discover inverse relationships or latent variables and to distinguish

between pollutant sources or evaluate the impacts of those sources. On the other hand, the techniques discussed in section 3.1.2 above did not require the intensive sampling regime or the mapping framework at all. It is therefore entirely possible to obtain valid scientific analysis using only classical multivariate techniques.

Variograms

The basic means of quantifying this spatial dependence, is the variogram (formerly known as the semivariogram). The variogram describes the average of the squared differences of two values that are calculated as a function of the distance between them, and this is the basis of all geostatistical calculations [39].

The values for the variogram are calculated experimentally from the following formula [39]:

$$\gamma(h) = \frac{1}{2n_{(h)}} \times \sum_{i=1}^n (x_i - x_{i+h})^2 \quad \text{Equation 1}$$

where γ is the variogram function, x_i are the data values, x_{i+h} is the data value at distance h from x_i and $n_{(h)}$ is the total number of value pairs that are included in the comparison.

The theoretical relationships between the γ -value, the variance σ^2 and the so-called covariance are given by the equation [39]

$$\frac{1}{2} \sigma^2 x_i - x(i+h) = \gamma(h) = \sigma^2 x_i - Cov [X_i, X(i+h)] \quad \text{Equation 2}$$

The derivation of the above equation is explained in detail by Wellmer [39] and interested readers are referred to his work, which also elaborates on the different types of variograms used for different applications.

Variograms are a robust statistical technique and small deviations do not significantly affect final results, but, generally, the requirement is that data distribution be relatively dense and, for geological applications, at least 30 drill holes are required in order to calculate meaningful variograms in several orientations. This will, of course provide considerably more data points than 30, for drill cores can be evaluated at any number of sections; note, however, that a variogram is calculated on the *difference* between data points.

Inverse Squared Distance Weighting Method

Strictly speaking, Inverse Squared Distance cannot be considered a geostatistical method, as it exploits the relationship between decreasing influence with increasing distance [39]. This is an empirical

technique and consists of multiplying the various data values with a weighting factor, adding the products, and then dividing this sum with the sum of the weighting factors. The grades are weighted by the inverse value of the square of the distance between them. The resulting equation is written [39]:

$$\bar{G} = \frac{\sum_{i=1}^n G_i * \frac{1}{d_i^2}}{\sum_{i=1}^n \frac{1}{d_i^2}} \quad \text{Equation 3}$$

\bar{G} is the average grade or value, G_i are the individual values and d_i is the distance between the values.

Kriging

The kriging method is named after the South African geostatistician Daniel J. Krige. It is a technique for determining the best linear unbiased estimator with minimal estimation variance. It can be used on a point as well as on a block. A block is simulated by numerous points that are then integrated. In the point kriging method, a reference point P is selected in a block, and the grade of the block is estimated by the kriging technique. In block kriging, the size of the block is taken into consideration. Only point kriging will be discussed here and further details can be obtained elsewhere [39]. Point kriging can be performed with a mean value (if the area being studied exhibits no obvious trends) or without a mean value (where clear trends are identified or there are clearly definable low- and high-grade zones).

Kriging uses a linear equation system that can be written in matrix notation (**K**), which contains all the variances σ_{ii} and covariances σ_{ij} of all the points x_i ($i = 1,2,3...n$) around the reference point P that are included in the weighting. σ_{ij} is the covariance between the values x_i and x_j and σ_{ii} are the covariances of a point with respect to itself (where $h=0$) and thus identical with the variances. The kriging matrix (**K**) is then multiplied with a vector λ , which includes all the weighting factors, which are derived from either the Inverse Squared Distance Method or a variogram. The vector also contains the so-called Lagrange multiplier for cases where kriging is performed without a mean.

The following equation then holds [39]:

$$\mathbf{K} \times \boldsymbol{\lambda} = \mathbf{D} \quad \text{Equation 4}$$

where **D** is a vector which contains the covariances of the reference point P to all the other points that are being taken into consideration. From this equation the accumulation value for point P at the centre of the block can then be estimated employing linear algebra.

Principal component analysis (PCA)

Principal Component Analysis (PCA) is a well-known multivariate data display method [104]. It effects data reduction by transforming an original data matrix into a set of linear, orthogonal latent variables referred to as Principal Components (PCs). Each PC produces for each object a characteristic value called a score, and for each variable a characteristic value called a loading. Both scores loadings may have positive and negative values for objects and variables respectively. The PCs are chosen such that the first PC explains most data variance and so on sequentially with the following PCs. Depending on the preference of the user, the objects and variables may be displayed together (biplot) or separately (scores or loadings plots). Such plots indicate relationships between objects or loadings or between objects and loadings as the case may be. Thus, it is a valuable tool for the recognition and elucidation of patterns in physical and chemical characteristics of soil samples.

Fuzzy clustering analysis

Fuzzy Clustering (FC) is a non-hierarchical cluster method, i.e., clusters are formed either by merging small groupings into larger ones or, conversely, by subdividing large clusters. Traditionally, classification implies that an object has a unique membership of a class i.e., the object's membership of any other class is zero. However, the fuzzy clustering models attempt to assign a degree of class membership for an object over a number of classes. The data matrix is submitted to processing by a membership function and the objects are classified into the number of classes selected by the user. The membership (m) values are displayed on a scale of 0 (no membership) to 1. In addition, the method is able to test the robustness of an object in a class by modifying the membership function to accommodate the multiple class membership. This method is particularly useful for the classification and selection of samples with multiple class membership i.e. fuzzy samples.

Principal Components Analysis /Absolute Principal Component Scores (PCA/ACPS)

Factor analysis in conjunction with varimax rotation and multilinear regression are mathematical algorithms employed for source apportionment [14, 63, 76, 105].

The main aim of applying PCA to datasets is to derive a limited number of components which explain a majority of the total variance in the original variables, and thus reduce the dimensionality of the data [76]. In order for this reduction to be useful, the new variables (PCs) must have substantive interpretations. However, very often, the components resulting from the initial PCA do not have unique or straightforward interpretations. For this reason, it is useful to rotate the PCs resulting from the PCA after discarding of non-significant PC's. The remaining few significant PCs [78], are then rotated to maximise they interpretability [76].

Many methods are available for the rotation of PCs. These may be classed as either orthogonal rotations, where the components remain uncorrelated with one another, or, oblique rotations, where PC's are allowed to be intercorrelated [76]. Orthogonally rotated PCs may be employed in subsequent regression or correlation analysis without concern that multicollinearities will confound the results, which is not the case for oblique component scores. Techniques such as varimax rotation, which results in orthogonal rotations, are therefore typically used with multilinear analysis for the purpose of calculating source contributions. Oblique component scores cannot be used for this purpose, since intercorrelation between the PCs may confound results, but they may yield more readily interpretable components in cases where independent PCs cannot [76].

The PCA/APCS method calculates the contribution of the different sources by performing multiple linear regression on the scaled scores after rotation [105]. For scaling, rotated scores of an artificial sample with concentrations equal to zero for all variables, are subtracted from the real sample scores for all sources. This transformation returns positive scores for all sources. Thus, the absolute factor scores are used as independent variables and multiple linear regressions [76, 105] are performed using the concentration of chemical species as dependent variables in accordance with the following equation [106].

$$C_i = bi_0 + \sum_{p=1}^p b_{pi} (APCS)_p \quad \text{Equation 5}$$

Where $p = 1, 2, \dots, n$, bi_0 the constant term of the multiple regression for chemical species i , represents the average contribution of the species from sources that were not determined by the principal component analysis, b_{pi} is the multiple linear regression coefficient for the source p and chemical species i , $APCS_p$ is the absolute principal component score of the rotated factor p for a sample, $APCS_p \times b_{pi}$ gives the contribution of source p to C_i . The values of C_i , bi_0 and b_{pi} should have the same dimensions as the original concentration of the chemical species.

Ranking analysis – PROMETHEE and GAIA

Preference Ranking Organization Method for Enrichment Evaluation (PROMETHEE) and Graphical Analysis for Interactive Assistance (GAIA) are the techniques discussed below. PROMETHEE, an outranking method, ranks objects (actions) on the basis of a set of criteria (variables) and is often coupled to the PCA display tool, GAIA [69, 102, 103]. Each variable must be modelled separately: the objects are ranked top-down or bottom up. It is also necessary to choose a preference function for ranking the objects and a weighting (default value = 1) that must be applied to the variables by the user. The preference is selected according to a “preference function”, six of which are available in the procedure [67]. The preference functions are described in terms of their shape for instance U-shaped, V-shaped, etc., which

reflects a particular mathematical treatment of individual measurements and maxima and minima to determine their relative importance. For the V-shaped preference function, for instance, all measurements are regarded as equally important up to a cut-off value (the “indifference threshold”). The preference functions are crucial because they define how much one object is to be preferred to others [102]. Initially, the objects for each variable in the data matrix are compared by subtraction in all possible ways to produce a difference matrix. Then, the steps in the following ranking procedure are similar to those discussed fully elsewhere [69, 103]. The PROMETHEE procedure can produce either PROMETHEE 1 outranking flows, which include the possibility where an object, *a*, cannot be compared with, *b*, but both have the same rank order; or, the more familiar rank order PROMETHEE II where this option is excluded and a rank order index, Φ , is obtained for each object. To produce a GAIA PCA biplot the rank order of the object from PROMETHEE is decomposed into a matrix form (not the same as the original data matrix), and this is then submitted to PCA, which produces the GAIA PC1 versus PC2 biplot.

GAIA thus displays the PROMETHEE results as a simple principal component analysis biplot that also incorporates a decision vector, $\Pi (pi)$, which appears in the biplot as a highlighted axis, pointing in the direction of the most preferred decision. P_i is determined on the basis of the chosen parameters and the weightings allocated. This is a useful device, since objects, which are the most preferred are indicated more clearly than by the PROMETHEE display. Also, objects can be classified according to categories, which simplify the display and enable the user to view different scenarios as they would develop using different categories of objects or variables.

This method is particularly useful for the selection of sites, ranking of sites and prioritization of remedial actions.

Positive Matrix Factorisation (PMF)

Various authors have explained PMF modelling in details (see, e.g. [72, 74])

In its simple form, the PMF equation can be written as:

$$X = GF + E$$

Where X = concentration data matrix for n number of samples and m chemical species; E = matrix of residuals; G = source contribution matrix for p sources and n number of samples, and

F = source profile matrix for p sources and m chemical species.

PMF resolves the receptor modelling problem by minimising an object function Q so that:

$$Q = \sum_{i=1}^n \sum_{j=1}^m (e_{ij} / s_{ij})^2$$

where s_{ij} is the uncertainty in the j th chemical species in the i th sample and,

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

is the part of the data variance that was not explained by the model. In addition to weighting the data points individually, PMF constrains the results to be always non-negative. This constrain reduces the rotational ambiguity in the factor analysis problem with the view to obtaining physically realistic solutions. It also ensures that the outcomes are positive since the concentrations of chemical species in environmental data cannot be negative.

The principal objective of the application of PMF analysis to a data matrix with n number of samples and m number of chemical species is to resolve the number of p independent sources as well as the values of g_{ik} (source contribution) and f_{kj} (source profile) that best fit the concentration data, χ_{ij} . Thus the number of pollutant sources and the contribution of each source to each sample obtained from a sampling site can be evaluated. This method is particularly useful for source identification and source apportionment of environmental pollutants.

5. References.

1. Carlosena, A., J.M. Andrade, and D. Prada, *Searching for heavy metals grouping roadside soils as a function of motorized traffic influence*. Talanta, 1998. **47**: p. 753-767.
2. Fang, G.C., C.N. Chang, Y.S. Wu, P.P.C. Fu, I.L. Yang and M.H. Chen, *Characterization, identification of ambient air and road dust polycyclic aromatic hydrocarbons in central Taiwan, Taichung*. Science of the Total Environment, 2004. **327**: p. 135-146.
3. Morillo, E., A.S Romero, C. Maqueda, L. Madrid, F. Ajmone-Marsan, H.Grcman, C.M. Davidson, A.S. Hursthouse and J. Villaverde, *Soil pollution by PAH's in urban soils: a comparison of three European cities*. Journal of Environmental Monitoring, 2007. **9**: p. 1001-1008.
4. Rowat, S.C., *Incinerator toxic emissions: a brief summary of human health effects with a note on regulatory control*. Medical Hypotheses, 1999. **52**(5): p. 389-396.
5. Vidovic, M., A. Sadibasic, S. Cupic and M. Lausevic, *Cd and Zn in atmospheric deposit, soil, wheat, and milk*. Environmental Research, 2005. **97**(1): p. 26-31.
6. Wang, X.J., R.M. Liu, K.Y Wang, J.D. Hu, Y.B. Ye, S.C. Zhang, F.L. Xu and S. Tao, *Application of multivariate spatial analysis in scale-based distribution and source study of PAHs in the topsoil: an example from Tianjin, China*. Environmental Geology, 2006. **49**: p. 1208-1216.
7. Wang, X.S., Y. Qin, and S.X. Sang, *Accumulation and sources of heavy metals in urban topsoils: a case study from the city of Xuzhou, China*. Environmental Geology, 2005. **48**(1): p. 101-107.

8. Wang, Z., J. Chen, X. Qiao, P. Yang, F. Tian and L. Huang, *Distribution and sources of polycyclic aromatic hydrocarbons from urban to rural soil: a case study in Dalian, China*. *Chemosphere*, 2007. **68**: p. 965-971.
9. Carrizales, L., I. Razo, J.I. Tellez-Hernandez, R. Torres-Nerio, A. Torres, L.E. Batres, A. Cubillas and F. Diaz-Barriga, *Exposure to arsenic and lead of children living near a copper-smelter in San Luis Potosi, Mexico: Importance of soil contamination for exposure of children*. *Environmental Research*, 2006. **101**(1): p. 1-10.
10. Hergren, L., A. Goonetilleke, and G.A. Ayoko, *Understanding heavy metal and suspended solids relationships in urban stormwater using simulated rainfall*. *Journal of Environmental Management*, 2005. **76**: p. 149-158.
11. Navarro, A., R. Tauler, S. Lacorte and D. Barcelo, *Chemometrical investigation of the presence and distribution of organochlorine and polyaromatic compounds in sediments of the Ebro River Basin*. *Analytical and Bioanalytical Chemistry*, 2006. **385**(6): p. 1020-1030.
12. Zhou, F., H. Guo, and L. Liu, *Quantitative identification and source apportionment of anthropogenic heavy metals in marine sediment of Hong Kong*. *Environmental Geology*, 2007. **53**(2): p. 295-305.
13. Levinson, A.A., *Introduction to Exploration Geochemistry*. 1974, Calgary: Applied Publishing.
14. Pujari, P.R. and V. Deshpande, *Source apportionment of groundwater pollution around landfill site in Nagpur, India*. *Environmental Monitoring and Assessment*, 2005. **111**, p. 43-54
15. Duruibe, J.O., M.O.C. Ogwuegbu, and J.N. Egwurugwu, *Heavy metal pollution and human biotoxic effects*. *International Journal of Physical Sciences*, 2007. **2**(5): p. 112-118.
16. Arakel, A.V. and T. Hongjun, *Heavy metal geochemistry and dispersion pattern in coastal sediments, soil and water of Kedron valley floodplain area, Brisbane, Australia*. *Environmental Geology*, 1992. **20**(3). p. 219-231
17. Selinus, O., *Medical Geology: an emerging speciality*. *Terrae*, 2004. **1**(1): p. 8-15.
18. Shi, G., Z. Chen, S. Xu, J. Zhang, L. Wang, C. Bi and J. Teng, *Potentially toxic metal contamination of urban soils and roadside dust in Shanghai, China*. *Environmental Pollution*, 2008. **156**: p. 251-260.
19. Ng, S.L., L.S. Chan, K.C. Lam and W.K. Chan, *Heavy metal contents and magnetic properties of playground dust in Hong Kong*. *Environmental Monitoring and Assessment*, 2003. **89**: p. 221-232.
20. Sharma, K. and L.B. Reutergardh, *Exposure of preschoolers to lead in the Makati area of Metro Manila, the Philippines*. *Environmental Research*, 2000. **83**: p. 322-332.

21. Vyskocil, A., Z. Fiala, V. Chenier, L. Krajak, E. Ettlerova, J. Bukac, C. Vian and S. Emminger, *Assessment of multipathway exposure of small children to PAH*. Environmental Toxicology and Pharmacology, 2000. **8**: p. 111-118.
22. Chen, T.B., J.W.C. Wong, H.Y. Zhou and M.H. Wong, *Assessment of trace metal distribution and contamination in surface soils of Hong Kong*. Environmental Pollution, 1996. **96**(1): p. 61-68.
23. Chen, T.B., Y.M. Zheng, M. Lei, Z.C. Huang, H.T. Wu, H. Chen, K.K. Fan, K. Yu, X. Wu and Q.Z. Tian, *Assessment of heavy metal pollution in surface soils of urban parks in Beijing, China*. Chemosphere, 2005. **60**: p. 542-551.
24. Chuang, J.C., P.J. Callahan, C.W. Lyn and N.K. Wilson, *Polycyclic aromatic hydrocarbon exposures of children in low-income families*. Journal of Exposure Analysis and Environmental Epidemiology, 1999. **2**: p. 85-98.
25. De Miguel, E., I. Iribarren, E. Chacon, A. Ordonez and S. Charlesworth, *Risk-based evaluation of the exposure of children to trace elements in playgrounds in Madrid (Spain)*. Chemosphere, 2006. **66** (3): p 505-513.
26. Ljung, K., O. Selinus, and E. Otabbong, *Metals in soils of children's urban environments in the small northern European city of Uppsala*. Science of The Total Environment, 2006. **366**(2-3): p. 749-759.
27. McKone, T.E. and J.I. Daniels, *Estimating human exposure through multiple pathways from air, water, and soil*. Regulatory Toxicology and Pharmacology, 1991. **13**(1): p. 36-61.
28. Mandal, B.K. and K.T. Suzuki, *Arsenic round the world: a review*. Talanta, 2002. **58**(1): p. 201-235.
29. Hopke, P.K., R.E. Lamb, and D.F.S. Natusch, *Multielemental characterization of urban roadway dust*. Environmental Science and Technology, 1980. **14**(2): p. 164-172.
30. Vogt, N.B., F. Brakstad, K. Thrane, S. Nordenson, J. Krane, E. Aamot, K. Kolset, K. Esbensen and E. Steinnes, *Polycyclic aromatic hydrocarbons in soil and air: statistical analysis and classification by the SIMCA method*. Environmental Science and Technology, 1987. **21**(1): p. 35-44.
31. Mielke, H.W., K.J. Berry, P.W. Mielke, E.T. Powell and C.R. Gonzales, *Multiple metal accumulation as a factor in learning achievement within various New Orleans elementary school communities*. Environmental Research, 2004. **97**(1): p. 67-75.
32. Mielke, H.W., C.R. Gonzales, M.K. Smith and P.W. Mielke, *The urban environment and children's health: soils as an integrator of lead, zinc, and cadmium in New Orleans, Louisiana, U.S.A.* Environmental Research, 1999. **81**(2): p. 117-129.

33. Zhang, C., W. Longhua, Y. Luo, H. Zhang and P. Christie, *Identifying sources of soil inorganic pollutants on a regional scale using a multivariate statistical approach: Role of pollutant migration and soil physicochemical properties*. Environmental Pollution, 2008. **151**(3): p. 470-476.
34. Heinrich, A., *The Application of Multivariate Statistical Methods for the Evaluation of Soil Profiles*. Journal of Soils and Sediments, 2007. **7**(1): p. 45-52.
35. Chen, T., X. Liu, X. Li, K. Zhao, J. Zhang, J. Xu, J. Shi and R.A. Dahlgren, *Heavy metal sources identification and sampling uncertainty analysis in a field-scale vegetable soil of Hangzhou, China*. Environmental Pollution, 2009. **157**: p. 1003-1010.
36. Zhang, W., S. Zhang, C. Wan, D. Yue, Y. Ye and X. Wang, *Source diagnostics of polycyclic aromatic hydrocarbons in urban road runoff, dust, rain and canopy throughfall*. Environmental Pollution, 2008. **153**: p. 594-601.
37. Zhang, C., *Using multivariate analyses and GIS to identify pollutants and their spatial patterns in urban soils in Galway, Ireland*. Environmental Pollution, 2006. **142**: p. 501-511.
38. Mostert, M.M.R., *Levels of pollutants on the surfaces of children's playgrounds situated in public parks.*, in *School of Physical and Chemical Sciences*. 2008, M.App.Sc. Thesis, Queensland University of Technology: Brisbane. p. 173.
39. Wellmer, F.W., *Statistical Evaluations in Exploration for Mineral Deposits*. 1998, Berlin: Springer. 379.
40. Zheng, Y.-M., T.-B. Chen, and J.-Z. He, *Multivariate geostatistical analysis of heavy metals in topsoils from Beijing, China*. Journal of Soils and Sediments, 2008. **8**(1): p. 51-58.
41. Goovaerts, P., *Geostatistics in soil science: state-of-the-art and perspectives*. Geoderma, 1999. **89**(1-2): p. 1-45.
42. Goovaerts, P., *Estimation or simulation of soil properties? An optimization problem with conflicting criteria*. Geoderma, 2000. **97**(3-4): p. 165-186.
43. Goovaerts, P., *Geostatistical modelling of uncertainty in soil science*. Geoderma, 2001. **103**(1-2): p. 3-26.
44. Van Meirvenne, M. and P. Goovaerts, *Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold*. Geoderma, 2001. **102**(1-2): p. 75-100.
45. Facchinelli, A., E. Sacchi, and L. Mallen, *Multivariate statistical and GIS-based approach to identify heavy metal sources in soils*. Environmental Pollution, 2001. **114**(3): p. 313-324.
46. Kanevski, M., R.Parkin, A. Pozdnukhov, V. Timonin, M.Maignan, V. Demyanov and S.Canu, *Environmental data mining and modeling based on machine learning algorithms and geostatistics*. Environmental Modelling & Software, 2004. **19**(9): p. 845-855.

47. Jarosinska, D., M. Biesiada, and M. Miszynska-Graca, *Environmental burden of disease due to lead in urban children from Silesia, Poland*. Science of the Total Environment, the, 2006. **367**: p. 71-79.
48. Hertz-Picciotto, I., R.J. Baker, P. Yap, M. Dostal, J.P. Joad, M. Lipselt, T.Greenfield, C.E.W. Herr, I. Benes, R.H. Shumway, K.I. Pinkerton and R.J. Sram, *Early Childhood Lower Respiratory Illness and Air Pollution*. Environmental Health Perspectives, 2007. **115**(10): p. 1510-1518.
49. Bostrom, C., P. Gerde, A. Hanberg, B. Jernstrom, C. Johansson, T. Kyrklund, A. Rannug, M. Tornqvist, K. Victorin and R. Westerholm, *Cancer Risk Assessment, Indicators and Guidelines for Polycyclic Aromatic Hydrocarbons in Ambient Air*. Environmental Health Perspectives, 2002. **110**(Supplement 3): p. 451-488.
50. Herngren, L., A. Goonetilleke, and G.A. Ayoko, *Analysis of heavy metals in road-deposited sediments*. Analytica Chimica Acta, 2006. **571**(2): p. 270-278.
51. Brzoska, M.M. and J. Moniuszko-Jakoniuk, *Interactions between Cadmium and Zinc in the organism*. Food and Chemical Toxicology, 2001. **39**: p. 967-980.
52. Alcock, R., V. Bashkin, M. Bisson, R.W Brecher, L. van Bree, R. Chrast, C.Colosio, H. Desqueyroux, E. Evstafjeva, M. Feeley, G. Fily, R. Guardans, H. Hansson, T. Hausmann, I. Holobek, M. Jabukowski, N. Jahansson, B. Krauthacker and M. Krzyzanovski, *Health Risks of persistent organic pollutants from long-range transboundary air pollution*. 2003. Masaryk University, Czech Republic: World Health Organisation, Europe.
53. Berko, H.N., *State of Knowledge on PAH's in Australia*. 1999, Department of Environmental Protection: Perth, Western Australia
54. National Environment Protection Agency, *PAH Health Review*, W. Department of Environment, Heritage and the Arts, Editor. 2003, Australian Government. p. 23.
55. Wang, Z., J. Chen, P. Yang, X. Qiao and F. Tian, *Polycyclic aromatic hydrocarbons in Dalian soils: distribution and toxicity assessment*. Journal of Environmental Monitoring, 2006. **9**: p. 199-204.
56. Petry, T., P. Schmid, and C. Schlatter, *The use of toxic equivalency factors in assessing occupational and environmental health risk associated with exposure to airborne mixtures of polycyclic aromatic hydrocarbons (PAHs)*. Chemosphere, 1996. **32**(4): p. 639-648.
57. Desaulles, A., S. Ammann, F. Blum, R.C. Brandli, T.D. Bucheli and A. Keller, *PAH and PCB in soils of Switzerland-status and critical review*. Journal of Environmental Monitoring, 2008. **10**: p. 1265-1277.

58. Cebulska-Wasilewska, A., I. Pawlyk, A. Panek, A. Wiechec, T. Popov, T. Georgieva and P.B. Farmer, *Exposure to environmental polycyclic aromatic hydrocarbons: Influences on cellular susceptibility to DNA damage.* . Mutation Research, 2007. **620**: p. 145-154.
59. Cebulska-Wasilewska, A., A. Wiechec, A. Panek, B. Binkova, R.J. Sram and P.B. Farmer, *Influence of environmental exposure to PAH's on the susceptibility of lymphocytes to DNA-damage induction and on their repair capacity.* Mutation Research, 2005. **588**: p. 73-81.
60. Schneider, K., M. Roller, F. Kalberah, and U. Schumacher-Wolz, *Cancer Risk Assessment for Oral Exposure to PAH Mixtures.* Journal of Applied Toxicology, 2002. **22**: p. 73-83.
61. USEPA, *National Air Toxins Program: The Integrated Urban Strategy*, USEPA.
62. Khalili, N.R., P.A. Scheff, and T.M. Holsen, *PAH source fingerprints for coke ovens, diesel and gasoline engines, highway tunnels and wood combustion emissions.* Atmospheric Environment, 1995. **29**(4): p. 533-542.
63. Singh, K.P., A. Malik, R. Kumar, P. Saxena and S. Sinha, *Receptor Modeling for source apportionment of polycyclic aromatic hydrocarbons in urban atmosphere.* Environmental Monitoring and Assessment., 2007. **136**: p 183-196.
64. Harris, D.C., *Quantitative Chemical Analysis*. Sixth Edition ed. 2003, New York: W.H. Freeman and Company.
65. Zhang, H.B., Y.M. Luo, M.H. Wong, Q.G. Zhao and G.L. Zhang, *Distributions and concentrations of PAHs in Hong Kong soils.* Environmental Pollution, 2006. **141**: p. 107-114.
66. Kokot, S., M. Grigg, H. Panayiotou and T.D. Phuong, *Data interpretation by some common chemometrics methods.* Electroanalysis, 1999. **10**(16): p. 1081-1088.
67. Ayoko, G.A., K. Singh, S. Balerea and S. Kokot, *Exploratory multivariate modeling and prediction of the physico-chemical properties of surface water and groundwater.* Journal of Hydrology, 2007. **336**: p. 115-124.
68. Hopke, P.K., *The Evolution of Chemometrics*. Analytica Chimica Acta, 2003. **500**: p. 365-377.
69. Kokot, S. and G.A. Ayoko, *Multicriteria Decision Making*, in *Chemometrics and Statistics*. 2005, Elsevier: Brisbane.
70. ESRI, *ArcGIS Geostatistical Analyst: Statistical Tools for Data Exploration, Modeling, and Advanced Surface Generation*. 2001, ESRI: Redlands, California. p. 19.
71. Yu, X.Z., Y. Gao, S.C. Wu, H.B. Zhang, K.C. Cheung and M.H. Wong, *Distribution of polycyclic aromatic hydrocarbons in soils at Guiyu area of China, affected by recycling of electronic waste using primitive technologies.* . Chemosphere, 2006. **65**: p. 1500-1509.

72. Chiou, P., W. Tang, C. Lin, H. Chu and T.C. Ho, *Atmospheric Aerosol over Southeastern Region of Texas: Chemical Composition and Possible Sources*. Environmental Modeling and Assessment, 2007: p. 333-350.
73. Larsen, R.K.I. and J.E. Baker, *Source apportionment of polycyclic hydrocarbons in the urban atmosphere: a comparison of three methods*. Environmental Science and Technology, 2003. **37**(9): p. 1873-1881.
74. Simcik, M.F., S.J. Eisenreich, and P.J. Lioy, *Source Apportionment and source/sink relationships of PAH's in the coastal atmosphere of Chicago and Lake Michigan*. Atmospheric Environment, 1999. **33**: p. 5071-5079.
75. Manoli, E., D. Voutsas, and C. Samara, *Chemical characterization and source identification/apportionment of fine and coarse air particles in Thessalonik, Greece*. Atmospheric Environment, 2002. **36**: p. 949-961.
76. Thurston, G.D. and J.D. Spengler, *Multivariate Assessment of Meteorological Influences on Inhalable Particle Source Impacts*. Journal of Climate and Applied Meteorology, 1985. **24**: p. 1245-1256.
77. Brandli, R.C., T.D. Bucheli, S. Ammann, A. Desaulles, A. Keller, F. Blum and W.A. Stahel, *Critical evaluation of PAH source apportionment tools using data from the Swiss monitoring network*. Journal of Environmental Monitoring, 2008. **10**: p. 1278-.
78. Harrison, R.M., D.J.T. Smith, and L. Luhana, *Source Apportionment of Atmospheric Polycyclic Aromatic Hydrocarbons collected from an urban location in Birmingham, U.K*. Environmental Science and Technology, 1996. **30**: p. 825-832.
79. Lim, M.C.H., *Chemical and Physical Characterization of Aerosols from the Exhaust Fumes of Motor Vehicles.*, PhD thesis. 2007, Queensland University of Technology: Brisbane. p. 358.
80. Lim, M.C.H., G.A. Ayoko, and L. Morawska, *Characterization of elemental and polycyclic aromatic hydrocarbon compositions of urban air in Brisbane*. Atmospheric Environment, 2005. **39**(3): p. 463-476.
81. Anderson, M.J., E.P. Daly, S.L. Miller and J.B. Milford, *Source apportionment of exposures to volatile organic compounds: II. Application of receptor models to TEAM study data*. Atmospheric Environment, 2002. **36**: p. 3643-3658.
82. Kim, E., S.G. Brown, H.R. Hafner and P.K. Hopke, *Characterization of non-methane volatile organic compounds sources in Houston during 2001 using positive matrix factorization*. Atmospheric Environment, 2001. **39**: p. 5834-5946.
83. Paatero, P. and U. Tapper, *Analysis of Different Modes of Factor-Analysis as Least-Squares Fit Problems*. . Chemometrics and Intelligent Laboratory Systems, 1993. **18**: p. 183-194.

84. Kim, E., P.K. Hopke, and E.S. Edgerton, *Improving source identification of Atlanta aerosol using temperature resolved carbon fractions in positive matrix factorisation*. Atmospheric Environment, 2004. **38**: p. 3349-3362.
85. Ramadan, Z., B. Eickhout, X-H. Song, L.M.C. Buydens and P.K. Hopke, *Comparison of Positive Matrix Factorization and Multilinear Engine for the apportionment of particulate pollutants*. Chemometrics and Intelligent Laboratory Systems, 2003, **66**: p. 15-28
86. Wang, D., F. Tian, M. Yang, C.Liu and Y-F. Li, *Application of positive matrix factorization to identify potential sources of PAHs in soils of Dalian, China*. Environmental Pollution, 2009. **157**: p. 1559-1564.
87. Vaccaro, V., S. Sobiecka, S. Contini, G. Locoro, G. Free and B.M. Gawlik, *The application of positive matrix factorisation in the analysis, characterisation and detection of contaminated soils*. Chemosphere, 2007. **69**: p. 1055-1063.
88. Lu, J., P. Jiang, L. Wu and A.C. Chang, *Assessing soil quality data by positive matrix factorization*. Geoderma, 2008. **145**: p. 259-266.
89. Mattiuzzo, E., L. Favero, F. Zennaro and D. Franco, *Heavy Metal Behaviour in an Experimental Free Water Surface Wetland in the Venice Lagoon Watershed*. Water, Air, & Soil Pollution, 2007. **183**: p 143-151.
90. Moret, S., G. Purcaro, and L.S. Conte, *Polycyclic aromatic hydrocarbon (PAH) content of soil and olives collected in areas contaminated with creosote released from old railway ties*. Science of the Total Environment, 2007. **386**: p. 1-8.
91. Oliva, S.R. and A.J.F. Espinosa, *Monitoring of heavy metals in topsoils, atmospheric particles and plant leaves to identify possible contamination sources*. Microchemical Journal, 2007. **86**: p. 131-139.
92. Department of the Environment, *Draft Guidelines for the Assessment & Management of Contaminated Land in Queensland.*, Department of the Environment, Editor. 1998, Queensland Government. p. 81.
93. European Union Working Group, *PAH Position Paper*, in *Ambient Air Pollution by Polycyclic Aromatic Hydrocarbons (PAH). Position Paper*. 2001, Office for Official Publications of the European Commission: Luxembourg.
94. Health Council of the Netherlands, *BaP and PAH from coal-derived sources*. 2006, 2006/01OSH: The Hague. p. 204.
95. Imray, P. and A. Langley, *Health-based Soil Investigation Levels*, Ministry for the Environment, Editor. 2001, enHealth Council.

96. Ostrom, N., M. Wilson, and J. Frampton, *Draft Lead Report*. 2004, Department of Toxic Substances Control: Sacramento. p. 119.
97. World Health Organisation, *Criteria for the Derivation of Toxic Equivalency Factors for Dioxin-like PCBs*, in *Target 19: Environmental Health Management*. 1993, World Health Organisation: Copenhagen. p. 18.
98. World Health Organisation, *IPCS International Programme on Chemical Safety. Health and Safety Guide No. 102*. 1996, International Labour Organisation, WHO: Geneva.
99. Ayoko, G.A., M. C. H. Lim, L. Morawska, Z.D. Ristowski, E.R. Jayaratne and S. Kokot, *A comparative study of the elemental composition of the exhaust emissions of cars powered by liquefied petroleum gas and unleaded petrol*. *Atmospheric Environment*, 2006. **40**(17): p. 3111-3122.
100. Lim, M.C.H., G.A. Ayoko, L. Morawska, Z.D. Ristovski and E.R. Jayaratne, *Effect of fuel composition and engine operating conditions on polycyclic aromatic hydrocarbon emissions from a fleet of heavy-duty diesel buses*. *Atmospheric Environment*, 2005. **39**(40): p. 7836-7848.
101. Lim, M.C.H., G.A. Ayoko, L. Morawska, Z.D. Ristovski and E.R. Jayaratne, *Influence of fuel composition on polycyclic aromatic hydrocarbon emissions from a fleet of in-service passenger cars*. *Atmospheric Environment*, 2007. **41**(1): p. 150-160.
102. Keller, H.R. and D.L. Massart, *Multicriteria decision making: a case study*. *Chemometrics and Intelligent Laboratory Systems*, 1991. **11**: p. 175-189.
103. Brans, J.P., P. Vincke, and B. Mareschal, *How to select and how to rank projects: The PROMETHEE method*. *European Journal of Operational Research*, 1986. **24**: p. 228-238.
104. Thielemans, A. and D.L. Massart, *The Use of Principal Component Analysis as a Display Method in the Interpretation of Analytical, Chemical, Biochemical, Environmental, and Epidemiological Data*. *Chimia*, 1985. **39**(7/8): p. 236-242.
105. Bruno, P., M. Caselli, G. de Gennaro and A. Triani, *Source apportionment of gaseous atmospheric pollutants by means of an absolute principal component scores (APCS) receptor model*. *Fresenius Journal of Analytical Chemistry*, 2001. **371**(8): p. 1119-1123.
106. Guo, H., Wang, T., L. and P.K.K. Louie, *Source Apportionment of ambient non-methane hydrocarbons in Hong Kong: Application of principal component analysis/absolute principal component scores (PCA/APCS) receptor model*. *Environmental Pollution*, 2004. **129**: p 489-498.