Chemometrics in Water Quality Criteria and Appraisal

Chizoba Henrietta Unaeze[1]*, Rasaq Bolakale Salau[2], Johnson Olusanya Jacob[3], Muhammed Muhammed Ndamitso[4], Abdul Kabir Mohammed[5]

[1]*National Biotechnology Development Agency, Abuja, Nigeria*
[2,3,4]*Federal University of Technology, Minna, Nigeria*
[5]*North Carolina Central University, Durham, North Carolina, U. S. A*
*Corresponding Author**

Abstract: **Water quality is extremely important for a number of reasons from the protection of marine organisms and the well being of marine ecosystems to the health of people in the region and the safety of industries such as aquaculture. As a result it is essential that environmental health in water body is monitored. Traditional monitoring methods include assessment of biological indices or direct measurements of water quality, which are based on in situ data collection and hence are often spatially or temporally limited. But the complexity of information requires new analysis techniques that allow us to identify the components and possible causes of spatial and temporal variability. An overview of the application of chemometric data analysis methods to complex chemical mixtures in various environmental media is presented. This paper presents a review of selected research given as examples of the application of principal components analysis and other statistical methods to identify contributions from multiple sources of contamination. This review identifies how these methods can be utilized to address water quality variability in order to foster a wider application of such techniques for water quality assessment and monitoring.**

***Keywords*: Water, Chemometrics, Contaminants, Sediments, Analysis, Heavy metals**

I. INTRODUCTION

Environmental scientists are confronted with the daunting task of assessing ecosystem and human health impacts arising from a multitude of modern day pollutants. From the basic steps of data collection, data analysis/structure interpretation, and predictive model development, scientists and engineers must be able to gain key insights into the sources, migration pathways, and health consequences of contaminants (Lutgarde, *et al*., 2015). Remediation and control strategies require this information to be successful.

Chemometrics is the application of multivariate mathematical and statistical tools to the study of chemical problems, including methods for the analysis and interpretation of analytical instrumentation data. Chemometric analysis can be applied to multichannel spectroscopic data, multi-component chromatographic data or combinations of multivariate and univariate instrumentation that create a broad array of measures characterizing a complex environment. (Wenning, *et al., 2012).*

Pattern recognition methods in chemometrics have been used to reveal and evaluate complex relationships in a wide variety of environmental applications. These methods have contributed to the systematic understanding of sediment trace metal and organic chemical concentrations arising from natural and anthropogenic sources, biological response of selected organisms to natural or toxic factors, identification of pollutant source contributions, and apportioning the spatial or temporal distribution of natural and anthropogenic influences. (Osman, *et al.*, 2018*).*

The purpose of this paper is to review chemometric techniques that can be used to examine spatial variability in water quality data. The differences between techniques, their advantages and limitations, as well as the questions they address and prospects are reviewed in order to identify the most appropriate techniques for differing circumstances.

II. CHEMOMETRIC METHODS OF ANALYSIS

Chemometric data analysis methods include exploratory, mixture component identification, group or class modelling, and calibration techniques. Exploratory techniques include cluster analysis, principal component analysis and factor analysis methods. Mixture component identification techniques include target factor analysis and polytopic vector analysis methods. Class modelling techniques include K-nearest neighbor (KNN), discriminant analysis, Bayes probability and SIMCA principal component modelling methods. Calibration techniques include multivariate regression, step-wise regression, partial least squares (PLS) path modelling, principal component regression (PCR), and non-linear regression methods. Data visualization is a vital component of all these methods for plotting raw or transformed data and displaying results in meaningful ways. (Nekoeinia, *et al.*, 2016).

2.1 Multivariate analytical techniques

Multivariate statistical analysis has been widely applied in environmental studies, which provide an effective way to reveal the relationships between multiple variables and thus they are helpful for the understanding of the influencing factors as well as sources of the chemical components. Numerous multivariate methods are available to analyse

spatial and temporal trends in water quality datasets. (Uría *et al.*, 2009).

2.1.1 Principal Component Analysis and Factor Analysis

These two methods are aimed at finding and interpreting hidden complex and casually determined relationships between features in a data set. Factor analysis is a class of multivariate statistical methods which aim to determine the underlying structure of a multivariate dataset, to summarise and reduce the amount of data. The summary is achieved by obtaining the factors, or underlying dimensions, of the data which describe the data in terms of a smaller number of items than the original variables. Data reduction is achieved by substituting the derived factors for the original variables in the dataset (Vialle, *et al*., 2013).

Principal components analysis is used to reduce the dimensionality of a dataset by explaining the variability of a large set of variables using linear combinations of the original variables, principal components. While many components are usually required to explain the total variability of a system, often the majority of the variability can be accounted for by a smaller number of principal components. (Vialle, *et al*., 2013).

Singh *et al.* (2019) applied both factor analysis and principal components analysis to their water quality sampling data from the Gomti River in India. Principal components analysis identified six principal components that accounted for 71% of the total variance in the water quality data. The first component was found to be related to electrical conductivity, dissolved salts, alkalinity, chloride and sodium and represented 27.9% of the total variance, while a second principal component was related to dissolved oxygen and accounted for 17.3% of the total variance. Factor analysis was applied to reduce the contribution of less significant variables from PCA. Factor analysis revealed six factors, comprised of 14 of the original 24 parameters, which explained 71% of the variance. The six factors from most significant to least significant were found to be related to the mineral composition of the river water, anthropogenic pollution sources, dissolution of soil constituents, mineral related hydrochemistry, fluoride, and suspended sediments.

Figure 1 illustrates the application of principal components analysis on the one year of monthly MODIS chlorophyll-a images of South Australia. PCA was performed in ENVI and it showed a principal component that accounted for 81.6% of the variance, representing the overall high and low chlorophyll-a areas within the region. The second and third principal components identified seasonal influences within the dataset, and represented 5.6% and 4.7% of the variability respectively. The second PC appears to separate areas with relatively high chlorophyll-a in summer from those with relatively high chlorophyll-a in winter. Likewise the third PC contrasts areas with relatively high chlorophyll-a in spring from areas with relatively high chlorophyll-a in autumn.

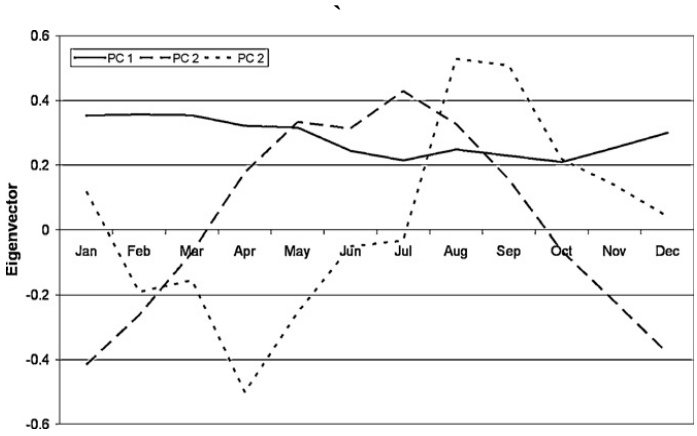

Fig. 1. Band loadings for principal components 1, 2 and 3.

2.1.2 Cluster analysis

The purpose of cluster analysis is to partition a set of objects into two or more groups based upon the similarity of the objects with respect to a chosen set of characteristics, so that similar objects are in the same class. Cluster analysis can be applied in both an exploratory and confirmatory sense: it can be used to either develop a new classification of the objects or to confirm a proposed grouping of the objects. There are two distinct forms of cluster analysis: hierarchical and non-hierarchical.

Hierarchical cluster analysis can be either agglomerative, where each object begins in its own cluster and then in subsequent steps the two clusters that are most similar are combined, or divisive, where all objects start in a single cluster and then in each step the objects that are most dissimilar are split off and made into smaller clusters. The hierarchical relationships between objects are commonly displayed graphically in a dendrogram or tree diagram.

Non-hierarchical cluster analysis, also known as k-means clustering, differs in that it does not involve the construction of a dendrogram and requires the number of classes to be pre-specified. Initially a cluster seed is established, which is an initial guess at the cluster mean, and the objects within a predetermined threshold distance of the seed are included in the resulting cluster. Further cluster seeds are then chosen until all objects are assigned to a cluster.

An example of non-hierarchical cluster analysis is provided by McNeil *et al.,* (2019) who applied k-means clustering to 30 years of surface water chemistry data from freshwater rivers, streams and lakes across Queensland, Australia. Approximately 34,000 measurements of many water chemistry variables were taken between the 1960s and 1990s covering the entire state. A two-stage kmeans cluster analysis of this data first identified 347 groups of measurements, which were then further reduced to nine water quality types. These nine water quality provinces highlighted the natural processes of the regions influencing surface water chemistry.

Unsupervised classification is widely used to derive cover classes or groups from digital remote sensing data. Typically the individual wavebands of multispectral imagery provide the variables for the clustering, but images from several dates may also be used. For example Erkkila and Kalliola (2013) applied an unsupervised classification to six summertime Landsat images of

the Archipelago Sea in Finland. Landsat Thematic Mapper (TM) and Enhanced Thematic Mapper (ETM^+) blue, green and red bands from the six images were used to create an 18-band image. The unsupervised classification of this image grouped together regions with similar spectral/temporal properties in the visible region of the spectrum, identifying groups which form zones with increasing distance from coastal waters to the open sea. These zones correlated well with measurements of chlorophyll-a and Secchi disk depths obtained for the region, which showed greatest chlorophyll-a concentrations and lowest Secchi depths in the inner archipelago closer to land.

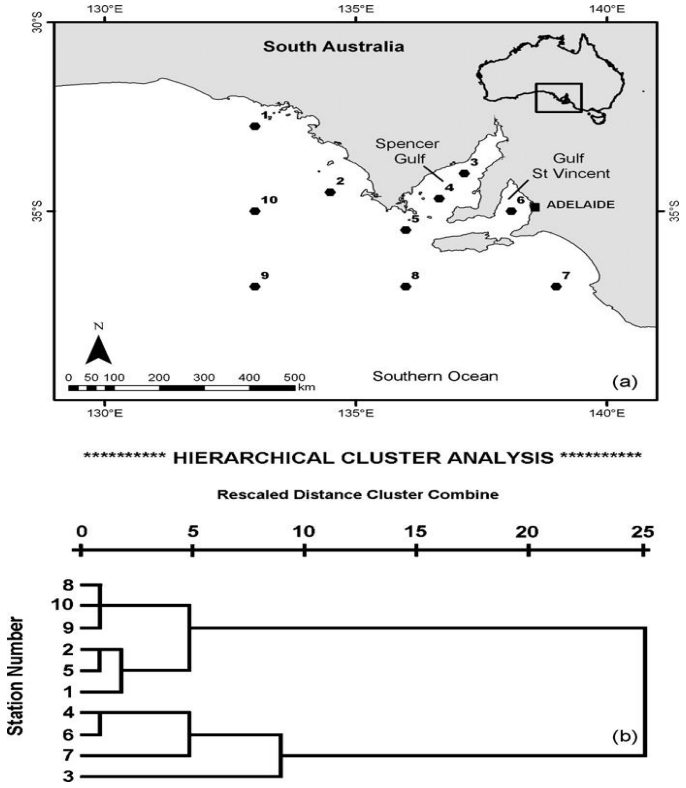

Fig. 2.(a) The location of the 10 stations at which chlorophyll-a temporal profiles were extracted from MODIS monthly chlorophyll-a images for 2006 and (b) an example of dendrogram produced from a hierarchical cluster analysis of chlorophyll-a temporal profiles at these locations.

2.2.3 Discriminant analysis

Discriminant analysis is a multivariate technique used to determine the variables responsible for the separation of objects into groups. Discriminant analysis can address a number of research questions including, but not limited to, determining whether statistically significant differences exist between two or more known groups, determining which independent variables account for the majority of the differences between groups, and establishing procedures for classifying objects into groups. The dependant variable may consist of just two groups, such as good or bad, or multiple groups, such as low, moderate and high pollution for example. The groups or categories must be mutually exclusive as objects may only be placed within one group. Discriminant analysis aims to determine whether these groups or classes can be separated using many independent metric variables, and to identify the variables contributing most to the classification.

Similarly to cluster analysis, discriminant analysis has mostly been applied to analysis of river water quality. Singh *et al.*, (2019) applied both a spatial and a temporal discriminant analysis to their dataset of 24 parameters collected monthly over 5 years in the Gomti River, India. For the temporal discriminant analysis the dataset was divided into three seasons: winter, summer, and monsoon.

Singh *et al*. (2019) also applied a spatial discriminant analysis to determine the variables responsible for variations between the regions of the river determined through cluster analysis. As for the temporal discriminant analysis the standard and forward stepwise methods obtained 92% accuracy using 23 and 17 parameters respectively, while the backwards stepwise method obtained 91% accuracy using just 9 parameters. The backward stepwise method showed that pH, temperature, alkalinity, Ca-hardness, DO, BOD, chloride, sulphate, and TKN are responsible for the majority of variation between the study regions. Therefore discriminant analysis was able to contribute to a significant reduction in the multivariate dataset, while providing insight into the variables contributing to variations in water quality between periods and regions.

III. TRENDS IN CHEMOMETRICS

3.1 Applications in environmental analysis

Examples of the application of chemometrics to environmental problems became increasingly evident in the scientific literature over the last four decades, as scientists explored the interpretive power of the various techniques and as chemometries software systems became more readily available. A wide variety of problems has been tackled in atmospheric, water quality and sediment, biological response, and industrial maintenance and control areas.

3.1.1 Atmospheric pollution and dispersion processes

The need to measure and delineate the concentrations of trace or excess concentrations of organic and inorganic chemicals in different geographical regions is typical in air pollution analysis. Frequently, multiple sources of these elements are present. Pattern recognition methods have been used to evaluate multi-component air pollution measures and to identify source contributions and the spatial extent of their influence over the region of interest. The particular contributions from various sources are important when toxic

materials are involved or where discharges result in concentrations of materials that are above regulatory limits.

3.1.2 Source identification and apportionment

PCA and cluster analysis methods have been used to identify contributing sources of total suspended particulate matter in urban and rural regions of Sicily. Results on elemental concentrations determined by particle induced X-ray emission (PIXE) analysis indicated soil (resuspended road dust) and automotive emissions, as well as specific industrial contributions, as major sources in two Sicilian cities. Cohen *et al*. (2017) used factor analysis to evaluate volatile organic compounds (VOCs), particulate mass and elemental concentrations in the Kanawha Valley, surrounding Charleston, WV, USA. One shortterm, specific factor was attributed to a forest fire in the region. Five other factors were identified as long-term contributions from general VOCs, automotive emissions, acid particulates, combustion, and chlorinated VOCs.

Target transformation factor analysis (TTFA) has been used to apportion measured aerosol concentrations among several sources. TTFA was also applied to atmospheric particulate data from the St. Louis area to derive detailed elemental profiles for anthropogenic sources in the region, including incinerators, a paint pigment plant, iron and steel works, motor vehicles, a copper products plant and a zinc smelter. The contributions of each source were identifiable at specific particulate monitoring stations based on the wind direction.

Vong *et al.* (2014) used PLS for source apportionment in air pollution modelling in comparison to results obtained with the chemical mass balance technique. Kim *et al.* (2018) applied cluster analysis and SIMCA methods to scanning electron microscopy data on atmospheric particulates to evaluate the contribution of automotive emissions to El Paso, TX, USA, air quality. Agglomerative (hierarchical) and partitioning clustering methods were used to identify categories of particulates stemming from similar sources. SIMCA models were developed for different particulate classes and used to assign mass fractions for 86% of automotive emission particles into 13 particle classes.

3.1.3 Emissions and episodic modeling

Exploratory and regression modelling techniques have been successfully used for emissions evaluation. Gonzales *et al.* (2013) utilized PCA, SIMCA and PLS methods to compare gas chromatographic data on PCB residues in municipal incinerator fly ash, for PCB Aroclors 1242, 1248, 1254, and 1260. The percentage composition of Aroclor 1260 found in the fly ash residues was 53-87%, a decrease from nearly 90% in environmental samples reported by other researchers, with a corresponding increase in the percentage composition of the lower chlorinated, more toxic Aroclors in the fly ash.

Exploratory and classification methods have been used to identify and anticipate air pollution episode conditions. Pijpers (2015) used PCA and cluster analysis methods to describe and predict air pollution complaint hours in the industrialized region at the estuary of the River Rhine near Rotterdam.

3.2 Water quality and sediment characterization

Similar to atmospheric pollution, water and sediment quality assessments typically address the problem of characterizing chemical concentrations and sorting out the influence of natural and anthropogenic sources. Additionally, the accumulation of contaminants deposited in sediments introduces a temporal aspect to the analysis that is not usually of concern in water, soil and air investigations.

3.2.1 Chemically-contaminated sediments

In Halifax Harbour, Scott and Haggard (2015) used factor analysis to identify four primary factors contributing to concentrations of trace metals in sediments: primary anthropogenic contamination (sewage effluent), surface drainage with high clay content, secondary contamination from industrial and urban dump sites, and diagenetic alteration yielding remobilization and precipitation of Fe- Mn oxides. Temporal variations in factor contributions were profiled as a function of sediment depth.

Poulton, (2013) used cluster analysis on similarity scores of ratios of concentrations of trace pollutants, including heavy metals and PCBs, in sediment samples from Lake Ontario to identify areas with similar chemical patterns. Sediments from sites near Hamilton Harbour segregated into clusters based on proximity to the main harbor basin, various point sources, and confluence with Lake Ontario and Cootes Paradise. Port Hope samples grouped in clusters representative of the turning basin, the river and west slip, the nearshore zone and an area impacted by a uranium refinery.

3.2.2 Water quality characterization

Francisco *et al*. (2014) applied PCA and FA methods to 33 hydrocarbon and 31 fatty acid variables measured in the Ebro River delta, near Ampost, Spain. PCA results identified one outlier sample affected by a high, specific algal input, and revealed important differences between dissolved and particulate concentrations in the bays and river channels. FA results provided identification of three dissolved phase contribution sources (anthropogenic, algal, and higher plant wax materials), five particulate influences in the bays (algal, terrestrial, microbial, petrogenic and other bacterial) and four particulate influences in the river channels (algal, anthropogenie + terrestrial, mixed terrestrial + microbial, and specific algal).

Rauret *et al.* (2015), used cluster analysis and nonlinear mapping to evaluate ground water quality in the Tenes River aquifer near Barcelona, Spain. Results revealed variables affected by sampling procedure (manual or pumped) and by pollution from surface water. A similar approach was adapted by Librando (2009), using PCA, KNN, linear discriminant analysis (LDA) and SIMCA to evaluate water quality data

from three Sicilian rivers. PCA contributions for the 25 measured variables were evaluated to ascertain the importance of each variable, PCA plots were evaluated for outliers (distinctive samples) and behavioral structure, and classification results for KNN, LDA and SIMCA were evaluated to ascertain similarities and differences between sampling sites. Brown *et al.* (2013) applied several pattern recognition methods, including cluster analysis, PCA, SIMCA and KNN, to evaluate the effect of coal strip mine drainage waters on Trout Creek, in Colorado.

3.3 Petroleum pollution assessment

The identification of sources of hydrocarbons is a particularly challenging application of chemometrics, due to the hundreds of possible sources and fingerprint patterns. For example, Jorge *et al.* (2016) applied KNN and SIMCA to classify samples from 40 crude and residual oils, in unmodified and weathered conditions, using neutron activation analysis of 22 trace elements. Classification accuracies of the samples into their known oil categories up to about 88% were obtained.

Rekadwad *et al.* (2017) classified spilled oil tarballs to crude oil classes from 62 sources of oil shipped to Japan, based on gel permeation chromatography. PCA and cluster analysis methods were used to characterize aromatic and aliphatic hydrocarbons in Madrid urban runoff revealed two main clusters characteristic of residential areas with low concentrations and moderately industrialized, high traffic residential areas with intermediate to high concentrations. PCA results indicated three main factors: a pyrolitic component with strong contributions from chrysene, benzo [a] pyrene, phenanthrene, and methylphenanthrene; a petrogenic component with strong contributions from pristane, phytane, and anthracene; and a natural organic component with strong contributions from oleic, stearic and linoleic acids.

3.4 Rainwater contamination studies

Le *et al.* (2017) used PCA to analyze sulfate and trace metal ion concentrations in rainwater samples near a Puget Sound copper smelter. Factors contributing to ionic concentrations in rainwater included soil/nitrate, the copper smelter, seasalt, and auto/fuel combustion. In a similar fashion, Hooper and Peters (2010) used PCA to identify the primary contributions for solutes in rainwater throughout the USA. Principal contributions were found to be due to three factors: acid, salt and agricultural/soil. The relative contributions of each factor were determined at 194 stations, located throughout the USA. Acidic conditions in the northeastern USA remained steady as a percentage component, in spite of the decline in sulfate deposition, due in part to a decline in agriculture/soil contributions which help neutralize the acidity. In a parallel study, Le *et al.* (2017) evaluated the ionic composition of rainwater and atmospheric samples at a remote station on the Olympic Peninsula, using PCA, PLS and calculated enrichment factors and scavenging ratios. The primary influence on rainwater chemistry was shown to be due to scavenging of submicrometer aerosol particles.

3.5. Biological response modeling

A typical goal for biological response modeling is to understand the influences of natural and anthropogenic factors on biological systems, biochemical mechanisms, organ function, and chemical body burdens in tissues. The data generated in biological investigations can be used to establish the prevalence and levels of exposure at different life stages, to identify trends in exposure, and to assess the effects of source mitigation on subsequent chemical exposures. In a study by Ali *et al.*, (2021) to understand the influences of natural and anthropogenic factors on biological systems, Screen-printed anion-exchange solid-phase extraction was applied for point-of-care determination of angiotensin receptor blockers. The method was applied for the determination of ARA-IIs in human blood plasma samples, and relative recoveries in the range of 89.0–107.8% with relative standard deviation (RSDs (≤8.9% were obtained.

Similarly, Shamim *et al.*, (2019) detected glucose in human blood plasma using fabricated electrode as a point-of-care (POC) biosensor. The detection limit was 1.1 μM, and the sensitivity was 620 $\mu A\, mM^{-1}\, cm^{-2}$ at the linear range of 2–426 μM. Data obtained from these biological studies could be clearly interpreted using chemometric techniques.

3.6 Chemicals in biological tissues

Where a mixture of contaminants such as petroleum hydrocarbons, PCBs, PCDDs, and PCDFs are measured in biological tissues, chemical source fingerprinting offers a viable means of identifying the industrial processes involved in the contamination, defining the geographical boundary of the contamination, and estimating the contributions from sources. For example, Karafistan (2019) used PCA and PLS to evaluate distribution of heavy metals in mussels on the coast of Finland, near the city of Pori. PLS prediction models were developed for percentage of shell deformities, and for distance from two pollution sources, a titanium oxide plant and the river Kokemaenjoki.

Amigo *et al.* (2010) applied PCA and PLS to concentrations of PCDD/PCDF isomers in crab tissues sampled off the southern Norwegian coast. A linear relationship in log-transformed concentrations with distance was found for male crabs in the 18-33 km range from a magnesium plant. Beyond that range, the contribution from the plant was too low and PCDD and PCDF concentrations were influenced by other sources.

Elias *et al.* (2018) used PCA to demonstrate differences in PCB/PCDD/PCDF pollutant concentrations in herring gulls, lake trout and walleye in the Great Lakes. Patterns of PCDD/F residues in the biota of lakes Erie, Huron and Superior were relatively consistent within each lake, with the exception of western Lake Erie, which was influenced by local sources of PCBs. Lake Ontario samples exhibited high concentrations of mirex and photomirex, attributed to manufacturing sources in the Niagara and Oswego Rivers. Pattern recognition methods

have also been useful to assess ecotoxicological factors. Mishra *et al.* (2017) used PCA to correlate sediment chemical concentrations and changes in species populations in Puget Sound. Sensitive benthic species were identified by negative loadings on pollution influenced factors. PCA scores plots of sampling sites versus pollution/response factors indicated the severity of distributed pollution loads.

Vogt (2014), used PCA, SIMCA and PCR to evaluate the relationship between species diversity indices and concentrations in sediments of total organic carbon, total nitrogen and 8 elements from 14 sampling stations in Kristiansandfjord, Norway. Negative effects of pollution on species diversity were revealed, with a clear gradient from polluted stations close to an industrial factory emission to cleaner stations in the main fjord basin and up into the river estuary.

3.7 Forest and plant productivity studies

Moseholm (2019) modeled plant growth of three varieties of *Lolium multijlorum* Lam. in response to climatological and pollutant variables. Controlled concentrations of nitrogen dioxide and sulfur dioxide were applied to test plants, in parallel with plants grown with charcoal filtered air and unfiltered air. PLS and SIMCA methods were used to illustrate the importance of warm weather, sunshine and time of growing season on plant growth. Exposure levels were considerably lower than those previously associated with plant growth inhibition and below short term concentration levels of concern for human health. The study confirmed chronic growth depression at low concentrations.

Zorica *et al.* (2020) used PCA and canonical discriminant analysis to evaluate terpene variation and differences in geographically distributed populations of western red cedar *(Thuja plicata).* The data exhibited low intra- and interpopulational variability, with some small differences between coastal and interior populations; however, latitude and elevation differences were not correlated with terpene content.

Similarly, Burkhardt *et al.* (2016) investigated air pollution damage to Norwegian Spruce, using pyrolysis-field ionization mass spectrometry of tree needles, ICP-AES analysis of needle element concentrations, physiological needle variables, biometric measures, and soil variables. Pattern recognition analysis included PCA score plots and Fischer weights of variables between four categories of tree damage according to the degree of needle loss and canopy discoloration. No single variable could explain the degree of tree damage. Tree discoloration could be partially explained by a combination of soil acidity, aluminum toxicity, water stress and a soil nutrient component. PCA revealed distinctions between trees differing in needle loss and showed that trees with significant needle loss had higher concentrations of ergosterol, p-sitosterol and a-tocopherol.

3.8 Human health effects

Diaz *et al.* (2013) applied multivariate linear regression and step-wise regression to meteorological and air pollution variables to determine the relationship between weather conditions and the rate of hospital admissions for chronic obstructive pulmonary disease and heart failure. The study found a significant influence of suspended particulate and sulfur dioxide pollution on admissions for both diseases, even at concentrations much lower than those given as a warning guide by international legislation.

Vogt (2014) applied PCA, SIMCA and PCR to evaluate the potential mutagenicity of respirable particulate concentrations in a wood-burning community in Norway. Four factors were identified: wood burning, domestic oil heating, automotive emissions, and a copper/sulfur/ atmospheric transformation process or source. Mutagenicity was found to increase with general pollution; however, increases in domestic oil heating or the copper/sulfur factors include toxic compounds which counteract the measurement of mutagenic activity.

3.9 Industrial maintenance and process control

Highly important goals for industry are reduction of waste and elimination of toxic emissions. The primary advantage of the chemometric approach in industrial settings is the relative ease of implementing a highly-focused instrument system for monitoring the quality of a product or raw material, with parallel reductions in waste and unwanted emissions. General purpose analytical instrumentation can be converted into a source of highly specific, quality control information through chemometric calibration. For example, Hollstein *et al.* (2015) utilized neural net classification of near-infrared reflectance spectra to separate recycled plastics into six resin categories with an assignment accuracy of 98%. The minimization of PCDD/PCDF emissions using various incineration plant operating conditions was modelled by Cruciani *et al.* (2011) using linear and non-linear PLS methods. Gonzales *et al.* (2013) utilized SIMCA to optimize the reduction of PCB residues in municipal incinerator fly ash. Hazardous waste assessment and control problems were addressed by Krska *et al.* (2012) using statistical analysis of sequential samples to evaluate cleanup of contaminated sites and by Sarkar *et al.* (2012) who used SIMCA to identify hazardous compounds from their mass spectra.

IV. MAJOR CHALLENGES

One challenge for chemometrics in chemistry is the spectacular growth of databases that contain a large amount of data on molecular structures and their properties, for instance the Cambridge Structural Database (CSD) for organic and organometallic crystal structures, the Brookhaven Protein Data Bank_(PDB) mainly for protein structures, and the Nucleic Acid Database _NDB for nucleic acids. In the beginning, the primary use of these databases was storage and retrieval of data. These are, of course, valid objectives. However, the more data was gathered, the more it was

realised that these databases contained important information that was not explicitly brought into them. It is the information hidden in the relations between all these data, such as similarities and dissimilarities that may reveal important new chemical knowledge. Finding these hidden relations in databases is sometimes called data-mining or knowledge discovery.

4.1 Data-mining: specific issues

4.1.1 Database size

Data-mining is a term that is used to describe the process of extracting information and identifying interesting patterns or features out of large masses of data. Chemometricians are used to dealing with relatively clean, often more or less designed datasets. These datasets are analysed with all kinds of pattern recognition and other exploratory data-analysis tools. The results can be evaluated statistically in order to draw chemically relevant conclusions. The number of objects in databases, interesting for data-mining, is orders of magnitude larger. Consider, e.g., the database resulting from the human genome project already containing gigabytes of data. This leads to problems where neither chemometricians nor statisticians are used to.

One obvious consequence is that with current computer technology it is not possible to keep all data in memory. This means that if one wants to process all data, new algorithms and strategies will have to be developed that can process the data sequentially, or that do not use all data but a limited subset.

A second problem is the high dimensionality of the data. Clearly, not all variables will be relevant in identifying new patterns. A distance measure in the full dimensionality may easily overlook patterns differing in only a small number of variables. In the case of molecular databases, the problem of finding an appropriate distance metric is directly related to the representation of the data. Another less obvious problem associated with the large amounts of data lies in the statistical evaluation of the results. Most statistical tests are based on a fixed level of significance, e.g., as 0.01 or 0.05, the probability for a type I error or for wrongly rejecting the null hypothesis. However, given the large amounts of data, any tiny difference will become statistically significant, even when it has no real meaning. The power of tests, the ability of correctly rejecting the null hypothesis will be much more important in data-mining research.

4.1.2 Contaminated data

Another problem with data sets usually encountered in chemometrics is the way databases are constructed. Data are not gathered according to a design, but rather are stored when measured and interpreted. There are several consequences of this fact. Firstly, old data are probably of a different quality than more recent data. Moreover, not all scientists submitting data use the same techniques and the same equipment, so that even with the newest data some inherent quality differences will be present. Furthermore, the objects in a database are almost certainly a non-random sample of all possible objects, a fact which may lead to unwarranted conclusions. The data almost always will deviate from normal distributions, and are easily contaminated by errors. Missing data represent a further problem. In cases where not all variables are measured or in cases where some variables are impossible to measure, conventional statistical methods to model the dependencies in the data may break down. Taking only those variables and objects without missing values into account may lead to a substantial decrease of the data set size. Apart from the smaller generality of the conclusions, the chance that something interesting is missed increases.

V. PROSPECTS

There are an increasing number of situations where the use of multivariate analysis is necessary in chemistry and may greatly facilitate the interpretation of data. In environmental chemistry the average or mean of two sample sets believed to be representing a polluted and a non-polluted sample set may be the same or, in a statistical sense, non-distinguishable. If the non-polluted sample set represents a true background with no influence from potential pollution sources, then it is likely that the covariance structure in this sample set will not have any resemblance to the covariance structure in the sample sets collected close to, and representing, the pollution sources. By using principal component analysis the variance/covariance structure in the two datasets may be investigated. Loading plots of two and two principal components may then show if the variables have the same variance-kovariance structure. Using PCVD plots the comparison of the principal component loading values may be made directly in one plot, which shows visually if the variancekovariance is different in the two datasets. Compared with the interpretation of loading plots of separate classes, the PCVD plots will allow identification of similarity in covariance structure when the principal components are reversed, i.e. when the first principal component in class A represents the same covariance structure as the second principal component in class B.

Another area of chemistry where PCVD plots should be useful is in process analytical chemistry (PAC). Interpreting non-specific multivariate chemical analytical data, e.g. from NIR spectroscopy of heavy oil residues, to determine which methods, or wavelengths, to include in a rapid on-line analytical procedure is one example. Analysing sample sets of the various types of possible residues to obtain PC models for them allows comparison of the loadings using the PCVD approach. By identifying variables (wavelengths) which do not fall along the lines equidistant from the principal components, it is possible to determine which variables are important for modelling the principal components, or PLS-x components, in the different types of oil and thus which must be used for modelling of the dataset.

Chemometrics can also be applied in analytical spectroscopy. Chemometric techniques are capable of overcoming most of

the challenges associated with analytical spectroscopy as they have the ability to extract important features (e.g. underlying soil properties) from complex

(spectral) data sets and hence can be developed in conjunction with spectrometric methods (Mark and Workman, 2010) to perform rapid and stable SQA. Chemometrics is most popular in experimental design and in the application of multivariate analysis to complex data matrices (such as soil spectra which are combinations of diverse SQIs).

Typical problems related to SQA that can be successfully handled by multivariate chemometric techniques in conjunction with spectroscopy include the following.

1) Determination of the concentration of compounds e.g. nitrate content in soil by EDXRFS and mid-IR spectroscopy
2) Classification of soil in terms of origin and soil type by XRF and mid-IR spectroscopy
3) Prediction of soil chemical composition and contamination by EDXRFS, vis–NIR and LIBS
4) Development of quantitative structure activity relationship (QSAR) in for example soil ecotoxicology by IR and ultra-violet (UV) spectroscopy
5) Remediation process monitoring of contaminated soils by vis–NIR, ICP and mass spectrometry (MS)
6) Discrimination of soil types for intelligence and forensics by IR spectroscopy

VI. RECOMMENDATIONS

All of these techniques are complementary, each providing insight into the problem and the role of factors that contribute to the chemical or biological condition. Rather than rely on a single technique, environmental scientists should apply this broad chemometric methods, combined with standard statistical data analysis, to derive a comprehensive picture of complex environmental problems. When this is applied, informative insights will emerge from this approach.

REFERENCES

[1] Ali, F., Shamim, A., Atena, T., Aristides, D., Camellia, T. (2021). Screen-printed anion- exchange solid-phase extraction: A new strategy for point-of-care determination of angiotensin receptor blockers. Volume 222, 121518, ISSN 0039-9140, https://doi.org/10.1016/j.talanta.2020.121518. (https://www.sciencedirect.com/science/article/pii/S0039914020308092)

[2] Amigo, J., Popielarz, M., Callejón, R., Morales, M., Lourdes, T., Ana, P., Mikael, T., Andersen, T. (2010). Comprehensive analysis of chromatographic data by using PARAFAC2 and principal components analysis. Journal of chromatography. A. 1217. 4422-9.

[3] Andréfouët, S., Payri, C., Hochberg, E., Hu, C., Atkinson, M., & Muller-Karger, F. (2014). Use of in situ and airborne reflectance for scaling-up spectral discrimination of coral reef macroalgae from species to communities. *Marine Ecology Progress Series*, *283*, 161-177.

[4] Brown, L. M., Zhou, H., Zhang, J., Wang, L., Zhang, Z. (2013). "Pattern Recognition" special issue: Sparse representation for recognition. Pattern Recognition, 46(7), 1748-1749.

[5] Burkhardt J., Grantz D.A., Cánovas F., Lüttge U., Matyssek R. (2016) Plants and Atmospheric Aerosols. Progress in Botany Vol. 78. Progress in Botany, 78. 12:124

[6] Cohen S, Karmack T, Mermelsteinm R. (2017). Global measure of perceived stress. J Health Soc Behav; 24:385-96

[7] Cruciani, G., Clementi, S., Curti G. (2011). Some applications of the partial least-squares method. Anal Chim Acta 191:149–160

[8] Diaz, Z., Morgan, G., Corbett, S., Wlodarczyk, J. (2011). Air pollution and hospital admissions. *American Journal of Public Health.* 88. 12 :1761-1766.

[9] Elias, M.S., Ibrahim, S., Samuding, K. (2018).Multivariate analysis for source identification of pollution in sediment of Linggi River, Malaysia. *Environ Monit Assess* 190, 257

[10] Erkkilä, A., Kalliola, R. (2013). Patterns and dynamics of coastal waters in multi-temporal satellite images: support to water quality monitoring in the Archipelago Sea, Finland Estuar. Coast. Shelf Sci., 60 (2), pp. 165-177

[11] Fernandes, L., António. F., Ferreira, A., Cortes, R. Pacheco, F. (2018). A partial least squares –Path modeling analysis for the understanding of biodiversity loss in rural and urban watersheds in Portugal. Science of the Total Environment. 626. 1069-1085.

[12] Francisco, C., Clemente, L., Barrientos, E., López, R., Murillo, J. (2014). Heavy metal pollution of soils affected by the uadiamar toxic flood. The Science of the total environment. 242. 117-29.

[13] Gonzales, M., Qualls. C., Hudgens, E., Neas, L. (2013). Characterization of a spatial gradient of nitrogen dioxide across a United States-Mexico border city during winter. The Science of the Total Environment. 2013; 337:163–173

[14] Hooper, R. P., Peters, N., E. (2013). Use of multivariate analysis for determining sources of Solutes found in wet atmospheric deposition in the United States. Environ Sci Technol 23:1263–1268

[15] Jorge, A., Frederico, L., Marco, A., Renato, L. (2017). Characterization of Gasoline by Raman Spectroscopy with Chemometric Analysis. *Analytical Letters*. *50* (7) , 1126-1138.

[16] Karafistan, A., Gemikonakli, E. (2019). Contaminant Evaluation in Fish from the Mining-Impacted Morphou Bay, Cyprus, Using Statistical and Artificial Neural Network Analysis. *Mine Water Environ* 38, 178–186.

[17] Kim, S., Nadia, B., Alireza, M., SeJoon, P.,David, P. David S. (2018). A cluster analysis approach for differentiating transportation modes using Bluetooth sensor data. Journal of Intelligent Transportation Systems. 22: 353-364

[18] Krska, R., Becalski, A., Braekevelt, E. (2012). Challenges and trends in the determination of selected chemical contaminants and allergens in food. *Anal Bioanal Chem* 402, 139–162

[19] Le, T., Zeunert, S., Lorenz, M., & Meon, G. (2017). Multivariate statistical assessment of a polluted river under nitrification inhibition in the tropics. *Environmental science and pollution research international*, *24*(15), 13845–13862.

[20] Librando, V. (2009). Chemometric evaluation of surface water quality at regional level. *Fresenius J Anal Chem* 339, 613–619

[21] Lutgarde, M.C., Theo, H. R., Mischa, L.M., Ron W. (2015). Molecular data-mining: a challenge for chemometrics. Chemometrics and Intelligent Laboratory Systems 49. 121–133

[22] Mcneil, V., Cox, M. & Preda, M.. (2019). Assessment of chemical water types and their spatial variation using multi-stage cluster analysis, Queensland, Australia. Journal of Hydrology. 310. 181-200.

[23] Mishra, Sidharth & Sarkar, Uttam & Taraphder, Subhash & Datta, Sanjoy & Swain, Devi & Saikhom, Reshma & Panda, Sasmita & Laishram, Menalsh. (2017). Principal Component Analysis. International Journal of Livestock Research. 1. 10.5455/ijlr.20170415115235.

[24] Moseholm, L. (2019). Analysis of air pollution plant exposure data: the soft independent modelling of class analogy (SIMCA) and partial least squares modelling with latent variable (PLS) approaches. Environmental Pollution 53: 313-331.

[25] Nekoeinia, M., Mohajer, R., & Salehi, M.H. (2016). Multivariate statistical approach to identify metal contamination sources in

agricultural soils around Pb–Zn mining area, Isfahan province, Iran. *Environ Earth Sci* 75, 760.
[26] Osman, O., Faust, S. (2018). Chemistry of Water Treatment. Boca Raton: CRC Press, https://doi.org/10.1201/9781315139265
[27] Pijpers, F. W. (2015). Description of Air Pollution by Means of Pattern Recognition Employing the ARTHUR Program. *Environmental Applications of Chemometrics*. 7. 93-105
[28] Poulton D. J. (2013). Statistical zonation of sediment samples using ratio matching and cluster analysis. *Environmental Modeling & Assessment*. 13(2-3):379–404.
[29] Rauret G., López-Sánchez J.F., Sahuquillo A.S., Rubio R., Davidson C., Ure A., Quevauviller P. (2015). Improvement of the BCR three step sequential extraction procedure prior to the certification of new sediment and soil reference materials. J. Environ. Monit. 11:57–61.
[30] Rekadwad, B. & Khobragade, C. (2017). Oil Biodegradation. 10.1007/978-3-319-52666-9_3.
[31] Sarkar, S.K., Bhattacharya, B., Debnath, S., Bandopadhaya, G., Giri, S. (2012). Heavy metals in biota from Sundarban wetland ecosystem, Implications to monitoring and environmental assessment. Aquat Ecosys Health Manage 5:215–222
[32] Scott, j. T., Haggard, B. E. (2015). Implementing Effects-Based Water Quality Criteria for Eutrophication in Beaver Lake, Arkansas: Linking Standard Development and Assessment Methodology, Journal of Environmental Quality, 44, 5, (1503-1512).
[33] Shamim, A., ,Ali, F.,Hassan, S, (2019). Plasma-functionalized Highly Aligned CNT-based Biosensor for Point of Care Determination of Glucose in Human Blood Plasma. Electroanalysis. https://doi.org/10.1002/elan.201800895.
[34] Singh, K., Rathore, A. Bhushan, N. & Hadpe, S. (2019). Chemometrics Applications in Biotech Processes: A Review. Biotechnology progress. 27. 307-15. 10.1002/btpr.561.
[35] Stien, L., Manne, F., Kari, K. (2016). Automated image analysis as a tool to quantify the colour and composition of rainbow trout (Oncorhynchus mykiss W.) cutlets. Aquaculture. 261. 695-705.
[36] Uría, F. A., Cristina, L. M., Enrique, R., Maria, L. F. (2009). Journal of Hazardous Materials. Source identification of heavy metals in pastureland by multivariate analysis in NW Spain. 165: 1008 – 1015
[37] Vialle, C., Jacob, S., Montréjaud, V. (2013). Monitoring of water quality from roof runoff: Interpretation using multivariate analysis. Water Research. 45: 3765-75.
[38] Vong, R. Geladi, P. Wold, S. Esbensen, K. (2014). Source contributions to ambient aerosol calculated by discriminant partial least squares regression (PLS). Journal of Chemometrics. 12. 281-296
[39] Vogt F, Booksh K. (2014). Influence of Wavelength-Shifted Calibration Spectra on Multivariate Calibration Models. *Applied Spectroscopy*. 58(5):624-635.
[40] Zorica, P., Rada, M., Milena, S., Vera, V., Srđan, B. (2020). Chemodiversity in natural plant populations as a base for biodiversity conservation, Biodiversity and Biomedicine, 10.1016/B978-0-12-819541-3.00002-5, (11-41)