Chemometrics and Intelligent Laboratory Systems 149 (2015) 205–214

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Chemometrics applied to environmental systems

Philip K. Hopke

Institute for a Sustainable Environment and Department of Chemical and Biomolecular Engineering, Clarkson University, Potsdam, NY 13699 USA

ARTICLE INFO

Article history:
Received 1 June 2015
Received in revised form 30 July 2015
Accepted 31 July 2015
Available online 7 August 2015

Keywords:
Pattern recognition
Mixture resolution
Probability
Environmental data

ABSTRACT

To understand the processes that give rise to the observed concentrations of chemical species in environmental samples, it is often useful to apply chemometric methods to help convert chemical data into environmental information. In this review, various chemometrical methods including pattern recognition, mixture resolution methods, and probability estimation will be described, and their applications to a variety of environmental systems are presented. However, there are important differences between applying chemometrical methods in the laboratory as compared to environmental data. The environmental system can be monitored, but in general, it is not possible to plan experiments and use known distributional properties of the system under study. Given the complexity of these systems, large data sets are typically needed to provide the basis for obtaining a good understanding of its functioning. Expanding computer capabilities coupled with improved analytical methods have allowed the production and analysis of larger and larger data sets with more and more computationally intensive sophisticated algorithms that offer the promise of even greater information retrieval in the future.

1. Introduction

Environmental systems function with complex interactions between biotic and abiotic components. In order to understand the functioning of environmental systems, it is often necessary to chemically characterize these systems. Samples may be collected and analyzed for their composition, or chemical species can be measured directly in the environment using in situ monitors or sensors. Thus, like for any chemical system, it is necessary to use appropriate methods to relate the analytical signals with the species concentrations (analytical chemistry) as well as then examine the relationships among the measured species in various ways to describe the functioning of the chemical processes that gave rise to the observed values. Thus, chemometric methods can provide valuable tools to extract information from the often large sets of environmental data.

Environmental applications have been published in *Chemometrics and Intelligent Laboratory Systems* beginning with the first issue [1] and continuing with the papers derived from several of the workshops organized by the U.S. Environmental Protection Agency (Vol. 3, No. 1, 1–158, 1988; Vol. 37, No. 1, 1–214, 1997; Vol. 60., No.1, 1–281, 2002). There has also been a parallel development of environmetrics with its own journal starting in 1990 that focuses more on ecological rather than chemical data analyses. In this paper, the application of major chemometric methods to environmental chemical data will be reviewed. These examples will provide illustrations of the methods and problems to which chemometrical methods can be applied.

2. Pattern recognition

There are various ways that pattern recognition methods [2] can be applied to explore environmental multivariate data. For these problems, unsupervised methods are needed since there are rarely exemplars of specific properties that can serve as the basis of a supervised analysis. Methods like cluster analysis have been widely applied to a variety of data.

Hopke [3] reports the application of cluster analysis to the composition of surficial sediments collected from Chautauqua Lake. Chautauqua Lake is a 24-km-long, narrow lake in southwestern New York State. The abundances of fifteen elements was determined by neutron activation analysis for grab samples of the bottom sediments [4] which had been taken for sediment particle size analysis [5]. Ruppert *et al.* [6] reported the particle size distribution as characterized by percent sand percent silt, and percent clay, as well as the %organic matter and water depth above the sample. In addition, parameters describing the particle size distribution were determined including measures of the average grain size, mean grain size; median grain size, and parameters describing the shape of the distribution; sorting (standard deviation), skewness, kurtosis, and normalized kurtosis. Fig. 1 presents the cluster output using squared Euclidean distance as the dissimilarity measure and the mean within-cluster distance over all pairs of points. It can be seen that the sites fell into four main groups.

In order to interpret these clusters, it is useful to indicate which sites belong to each cluster. Fig. 2 presents this view by giving each of the four major clusters a distinct symbol. It can be seen that the top cluster (A) in Fig. 1 contains only sites in the center of the lake. The other three clusters contain all of the near-shore sites and the three near-shore

---

E-mail address: phopke@clarkson.edu.

http://dx.doi.org/10.1016/j.chemolab.2015.07.015
0169-7439/

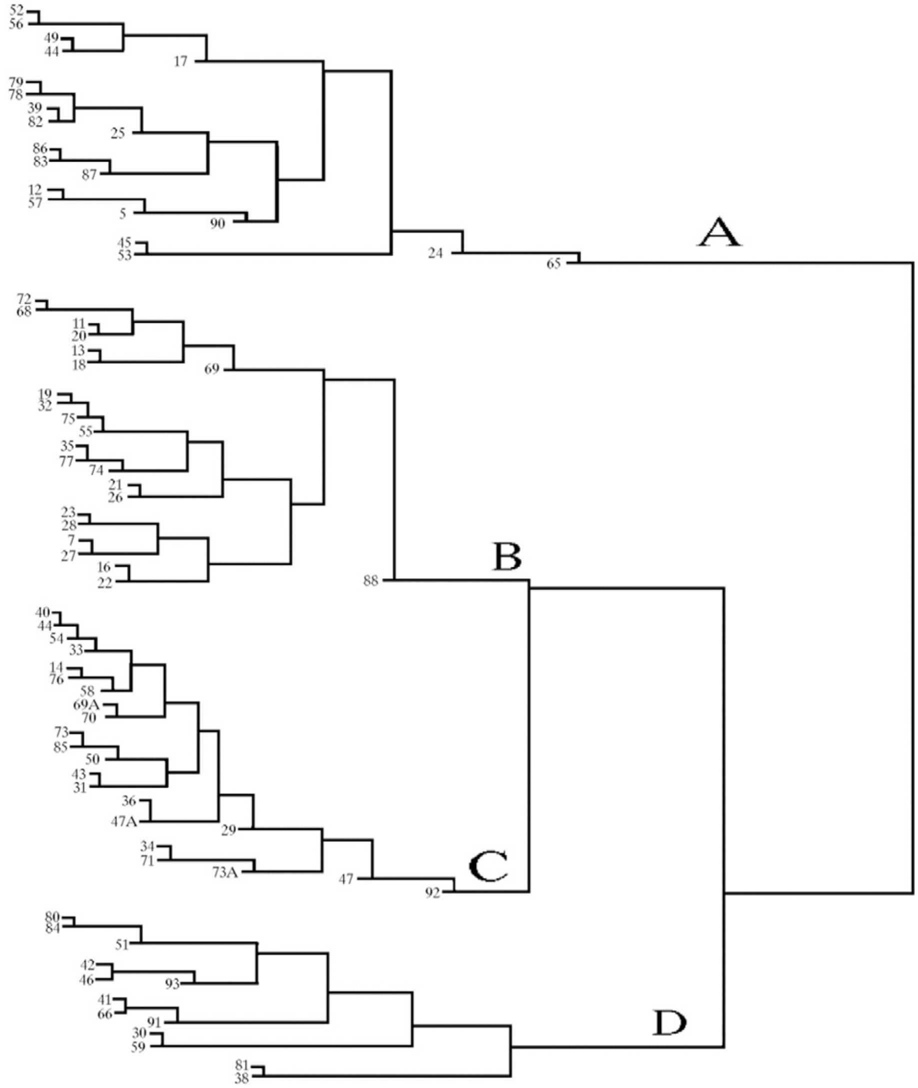

Fig. 1. Dendrogram of the hierarchical cluster analysis of sediment composition data from Chautauqua Lake, NY.

clusters come together before merging with the basin sites. There are differences between the three near-shore clusters. The bottom cluster (D) is very sandy and moderately well sorted. Most of the sites for this group fall in the northern end of the lake. There seems to be little active sedimentation in these sites and only very limited current or wave action. These are sites which have an above-average value for coarse-grained material. The second cluster (B) represents areas of active sedimentation and high wave action. All of the active delta areas fall in this cluster as well as those sites through the narrows at the center of the lake. Although the currents have not been measured, this region would be expected to have the strongest flow as the lake reaches its narrowest point. These high-energy environments are where there is active transport of sediments. In the deltas, the silty material is being deposited at a sufficiently high rate that the material cannot be sorted as it is deposited. The largest value of organic carbon is found in this cluster. This result agrees with the earlier observation that there is rampant weed growth in these deltic areas. The third cluster (C) represents an intermediate energy environment. In these areas, there is sufficient energy to sort the sediments and remove the fine-grained material. Thus, the material in these areas is moderately well-sorted. Thus, the three near-shore clusters do represent reasonable different environments and source material. The cluster pattern of these sites has helped to confirm the interpretation of the causal factors in this study.

Another application of pattern recognition has been employed to provide a basis for developing quantitative results from qualitative or semi-quantitative data [7]. There are several techniques available that can provide only qualitative or semi-quantitative characterization of samples. Of particular interest is the chemical and/or physical characterization of individual airborne particles. Such data provide information on the nature of the mixtures in the particle compositions. If the aerosol is an external mixture, then individual particles have distinct chemical and physical characteristics determined by their sources and the processes by which they were formed. An internally mixed aerosol is one in which all of the particles have uniform compositions that are the same as the bulk particle composition. In reality, the ambient aerosol is rarely either a true internal or external mixture. Typically, distinct composition particles are modified as they are transported from the source to the sampling location. Particularly, cloud or fog processing will result in much more uniform particle compositions. Thus, individual particle characteristics can provide information on both the particle sources and atmospheric transformation processes.

Techniques for characterizing individual particles include computer-controlled scanning electron microscopy (CCSEM) and aerosol time-of-flight mass spectrometry (ATOFMS). CCSEM is described in detail by Casuccio and Hopke [8]. Particles collected on a filter can be characterized for their physical size, shape, and elemental composition. However,