Aerosol and Air Quality Research, 20: 688–689, 2020
Copyright © Taiwan Association for Aerosol Research
ISSN: 1680-8584 print / 2071-1409 online
doi: 10.4209/aaqr.2020.01.0001

Letter to the Editor

Letter to the Editor: Ending the Use of Obsolete Data Analysis Methods

Philip K. Hopke[1,2*], Daniel A. Jaffe[3]

[1] *Department of Public Health Sciences, University of Rochester Medical Center, Rochester, NY 14642, USA*
[2] *Center for Air Resources Engineering and Science, Clarkson University, Potsdam, NY 13699, USA*
[3] *School of Science, Technology, Engineering and Mathematics, University of Washington Bothell, Bothell, WA 98011, USA*

Aerosol and Air Quality Research is taking an editorial stand with regards to outdated data analysis methods, specifically principal component analysis (PCA) and related techniques and enrichment factors (EF). In both cases, they have been replaced with more quantitative data analytical tools that provide much greater information on sources of variation in the data. Enrichment factors were first used in the 1960s when we basically did not have computers. It was a simple way of using double ratios to see if an element were substantially enriched over crustal abundances that had been reported by one of several authors. However, the information is quite crude since it simply says that the element is higher than typical crustal values and does not account for local variations in elemental abundances. When we now have the capabilities to look at correlations, statistical assessment of differences in means or medians, etc., we should provide appropriate quantitative estimates of significance of differences among samples.

In the case of PCA and other eigenvector-based methods, it has been shown by Lawson and Hanson (1974) and Malinowski (2002) that an eigenvector analysis is an unweighted least-square fit to the data. Such fits are going to create problems with heteroskedastic data such as is commonly encountered in atmospheric measurements. Typically, the measurement uncertainties are proportional to the measured values rather than a fixed value for all of the measurements (homoscedastic data). Thus, unweighted least squares fits will not provide the best estimators of the parameters of interest. PCA also typically uses a default of subtracting the mean value from the data points and scales them by the variance such that it apportions the variance rather than the variation of the actual measured concentrations. Although methods like Target-Transformation Factor Analysis (TTFA) avoided subtracting the mean, it still suffers from the problem of improper (absence of) data point weights. In the 1970s when mainframe computing power was less than what we carry in our pockets as a telephone, it was necessary to use simplifying methods like eigenvector decompositions to be able to obtain results in a reasonable time. However, we have long since gained sufficient computing power in personal computers to be able to perform full, explicit least-squares fits with proper data weighting. Factor analysis tools like non-negative constrained alternating least square (Tauler *et al.*, 1993, 1994), positive matrix factorization (Paatero and Tapper, 1993, 1994), and non-negative least squares (e.g., Lee and Seung, 1999; Camp, 2019) have been available for more than 25 years that allow proper weighting of individual data points. Software for these techniques are readily available for download: MCR-ALS (http://www.mcrals.info), PMF (https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses; https://www.psi.ch/lac/sofi-sourcefinder), and NNLS (https://pages.nist.gov/pyMCR/; https://cran.r-project.org/package=NMF) While PCA has value as a preliminary screening tool (e.g., Roscoe *et al.*, 1982) and as an exploratory tool, but it should not be used as a receptor model. For quantitative analyses that form the basis of conclusions in a paper, we strongly recommend more modern and statistically appropriate methods be used.

REFERENCES

Camp, Jr., C.H. (2019). pyMCR: A python library for Multivariate Curve Resolution Analysis with Alternating Regression (MCR-AR). *J. Res. Natl. Inst. Stand. Technol.* 124: 124018.

Lawson, C.L. and Hanson, R.J. (1974). *Solving Least Squares Problems.* Society for Industrial and Applied Mathematics, Prentice-Hall, Englewood Cliff, NY.

Lee, D.D. and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401: 788–791.

Malinowski, E.R. (2002). *Factor Analysis in Chemistry*, 3rd Edition, John Wiley & Sons, Inc., NY.

[*] Corresponding author.
Tel.: 1-585-276-3240
E-mail address: phopke@clarkson.edu

Paatero, P. and Taper, U. (1993). Analysis of different modes of factor analysis as least squares fit problems. *Chemom. Intell. Lab. Syst.* 18: 183–194.

Paatero, P. and Tapper, U. (1994). Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5: 111–126.

Roscoe, B.A., Hopke, P.K., Dattner, S.L. and Jenks, J.M. (1982). The use of principal components factor analysis to interpret particulate compositional data sets. *J. Air Pollut. Control Assoc.* 32: 637–642.

Tauler, R., Kowalski, B. and Fleming, S. (1993). Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. *Anal. Chem.* 65: 2040–2047.

Received for review, January 1, 2020
Revised, February 17, 2020
Accepted, February 29, 2020