Refining Principal Component and Compositional Data Analysis for Understanding Water Quality and Groundwater-Surface Water Interactions

CARLETON BERN[1], MICHAEL HOLMBERG[2], ZACHARY KISFALUSI[3]

[1]U.S. Geological Survey, Box 25046 MS 415, Denver, CO 80225, cbern@usgs.gov

[2]U.S. Geological Survey, 201 East 9th. Street, Pueblo, CO 81003, mholmber@usgs.gov

[3]U.S. Geological Survey, 201 East 9th. Street, Pueblo, CO 81003, zkisfalusi@usgs.gov

Compositional data, such as concentrations of elements in waters, soils, or rocks, convey only relative information because they are measured as proportions of the mass or volume of the whole. Spurious correlations and distortion of actual relationships can arise from improper analysis of such data. Data transforms, such as the centered log ratio (CLR) transform can overcome this obstacle, but each has its advantages and drawbacks. Simultaneously, principal component analysis (PCA) is a tool widely used to discern broad patterns in multivariate data, including compositional data. However, PCA is often used less effectively than it could be, because the contstituents or samples included or excluded from the analysis are not tailored to the question at hand. Here, PCA on CLR-transformed data is used to reveal the dominant patterns in a large compositional dataset from river, tributary, and groundwater samples to discern source areas and processes that decrease water quality along a 152 km reach of river.

The Arkansas River in semiarid eastern Colorado and western Kansas has numerous water-quality challenges. In different reaches and seasons, uranium concentrations exceed drinking water standards, selenium concentrations exceed chronic toxicity standards for aquatic wildlife, and high salinity levels reduce yields of crops irrigated with river water. Cretaceous age marine rocks exposed in the watershed are the largest source of these constituents. Return flows from canal diversions and irrigated agriculture enhance natural mobilization. Correctly identifying source locations for these constituents is necessary to develop and target management practices for water-quality improvements, and the combination of CLR and PCA is a powerful strategy to achieve that goal. The approach presented here can be easily applied to other datasets to locate hot spots and source locations for problem constituents.