Environmental Pollution 184 (2014) 530–539

Contents lists available at ScienceDirect

Environmental Pollution

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

Data-driven modeling of background and mine-related acidity and metals in river basins

Michael J. Friedel [a,b]

[a] Crustal Geophysics and Geochemistry Science Center, U.S. Geological Survey, Denver Federal Center, Box 25046, MS 964, Lakewood, CO 80225, USA

[b] Center for Computational and Mathematical Biology, University of Colorado, Campus Box 170, PO Box 173364, Denver, CO 80217-3364, USA

ARTICLE INFO

Article history:
Received 7 February 2013
Received in revised form
18 September 2013
Accepted 20 September 2013

Keywords:
Cluster analysis
Environmental assessment
Hydrothermal alteration
Mineral-resource assessment
Mining activity
Self-organizing map
Stochastic modeling
Uncertainty
Water quality

ABSTRACT

A novel application of self-organizing map (SOM) and multivariate statistical techniques is used to model the nonlinear interaction among basin mineral-resources, mining activity, and surface-water quality. First, the SOM is trained using sparse measurements from 228 sample sites in the Animas River Basin, Colorado. The model performance is validated by comparing stochastic predictions of basin-alteration assemblages and mining activity at 104 independent sites. The SOM correctly predicts (>98%) the predominant type of basin hydrothermal alteration and presence (or absence) of mining activity. Second, application of the Davies–Bouldin criteria to k-means clustering of SOM neurons identified ten unique environmental groups. Median statistics of these groups define a nonlinear water-quality response along the spatiotemporal hydrothermal alteration-mining gradient. These results reveal that it is possible to differentiate among the continuum between inputs of background and mine-related acidity and metals, and it provides a basis for future research and empirical model development.

Published by Elsevier Ltd.

1. Introduction

In the western U.S., mining activities produce acid-mine drainage with elevated concentrations of potentially toxic trace elements (Church et al., 2007). Although drainage from inactive mines can affect surface-water quality, the background weathering of altered and mineralized bedrock can also be a source of metals and acidity in historical mining districts (Bove, 1996; Miller et al., 1999; Plumlee et al., 1999; Bove and Knepper, 2000). Knowledge of background water quality in mining-affected areas is particularly important for establishing technically feasible remediation goals. The traditional approach for estimating background metal contributions is based on the examination of historical data (Runnells et al., 1998). Given that premining water-quality data are not available from most mined areas, various forms of modeling can be employed to estimate background metal contributions (Mast et al., 2007).

Several issues are likely to confuse modeling the effects of mineral-resources and mining on the environment (Mast et al., 2007). First, the effects of hydrothermal alteration and mining activities on the environment are spatiotemporal and therefore not well understood. For example, possible causes for a modified environmental response can include scale-dependent changes in quantity and(or) quality of both surface and groundwater. Some changes in water quantity may occur in response to increased runoff volume, flood frequency, and peak storm flows; and reduced groundwater recharge and baseflow contributions to streams. Changes in water quality may occur in response to the interaction of water with natural and anthropogenic sources, mixing surface-water and ground-water, chemical reactions, and differences in residence times. Second, the field observations used to characterize the coupled and nonlinear response of environmental forcing at different scales are irregular, noisy and sparse. This makes the construction and calibration of process-based models difficult and trend analysis nonunique and uncertain. For this reason, most studies of mineral-resource effects focus on interpreting the present occurrence and distribution of water quality using univariate (Schmidt et al., 2012) and linear multivariate statistical approaches (Mast et al., 2007).

One alternative nonlinear modeling paradigm is to mine data using an artificial neural network. Artificial neural networks are a generalized data-driven modeling group that includes supervised and unsupervised training methods. The successful application of supervised training methods depends on accurately specifying weights and the output layer of the network prior to its deployment (Hsu et al., 1995). One unsupervised method that requires no a priori

E-mail address: mfriedel@usgs.gov.

0269-7491/$ – see front matter Published by Elsevier Ltd.
http://dx.doi.org/10.1016/j.envpol.2013.09.036

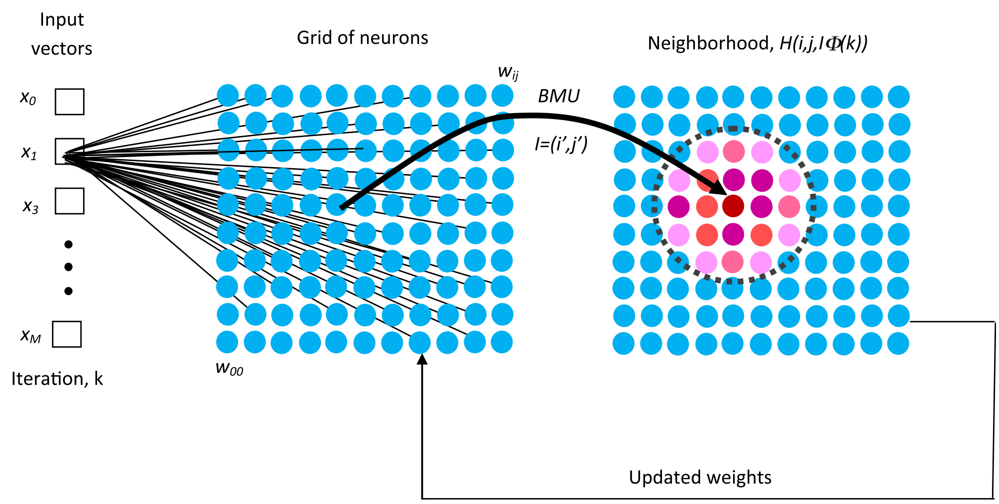

Fig. 1. Self-organizing map learning process. The 2-dimensional grid of neurons is characterized by local coordinates (i,j) and weights w_{ij}. During the kth iteration, an input vector x_m is presented to the grid, and the neuron with the smallest distance to the input vector is considered the best matching unit (BMU) vector. The Gaussian neighborhood function h(i,j,I, r(k)) determines the strength of association among neurons.

knowledge of underlying relations or designation of an output layer is the self-organizing map (SOM) technique (Kohonen, 2001). Some recent SOM-based environmental modeling applications include design-hyetographs (Lin and Wu, 2007), surface-water hydrographs (Kalteh et al., 2008), water toxic risk (Gevrey et al., 2010), stream water-chemistry (Zelazny et al., 2011), hillslope-chemical weathering (Iwashita et al., 2011a), soil properties (Iwashita et al., 2011b), post-fire hazards (Friedel, 2011), soil wettability (Xiong et al., 2011), groundwater exploration (Friedel et al., 2012), climate change trends (Friedel, 2012), stream source-water apportionment (Yang et al., 2013), and rainfall-runoff (Nourani et al., 2013).

In this study, the aim is to understand the relation of water-quality to background and mining-related acidity and metals. We hypothesize that basin measurements sampled across a spatio-temporal gradient can provide sufficient information in the underlying density function for model development, prediction and discrimination. The objectives are to (1) develop and apply a nonlinear classifier for predicting the relative contributions of background hydrothermal alteration and mining activity from water-quality samples, and (2) develop and interpret a conceptual model of the nonlinear water-quality response along a hydrothermal alteration-mining gradient. This approach extends the work of Mast et al. (2007) who sought to characterize sources of metals to surface water in Colorado using a linear multivariate modeling approach, and Friedel (2011, 2012) who used a SOM and multivariate statistical approach to derive conceptual models of hydrologic and geomorphic hazards associated with a post-fire landscape in western U.S., and conceptual models of groundwater hydrogeology in fractured crystalline bedrock in northeastern Brazil.

2. Methodology

Modeling mineral-resource effects on the environment is undertaken using a type of unsupervised neural network (Kohonen, 2001), called a self-organizing map (SOM). Application of the SOM technique to spatiotemporal mineral-resource and stream water-quality data is a three-step procedure. First, these data are used to train the SOM and estimate missing values. Second, the model diversity and performance is evaluated using cross-validation. Third, a conceptual model of mineral-resource effects on the basin water-quality response is derived using nonlinear multivariate techniques.

2.1. Training and estimation

The SOM training process provides a way of representing multidimensional data in a lower dimensional space than the original data set (Fig. 1). Reducing the dimensionality is based on an iterative process (Kohonen, 1984) that is performed each time an input pattern is presented to the map: competition to determine the best matching unit (BMU) vector and cooperative learning (spreading information contained in the current input vector across the map).

In the first training step, a weight vector W_i with the same dimensionality as the input data vectors V_j is assigned to grid neurons in the SOM (Fig. 2). Following an iterative procedure, the SOM is constructed considering the differences between the normalized input vector V_j and the weights W_i of the neurons given by

$$D_{ij} = (W_i - V_j)^T (W_i - V_j), \quad (1)$$

where T is the transpose. Normalization of the input vectors is conducted with respect to their standard deviations.

In the second step, a weight update is determined as a function of the distance to the current BMU, expressed through the Gaussian neighborhood function $\varphi(\Delta, n)$. The rate used to adjust the weight of neurons decreases with distance Δ between each neuron and BMU. Updates of the weights are adjusted according to

$$W_i(n+1) = W_i(n) + \alpha(n)\phi(\Delta, n)[V_i(n) - W_i(n)], \quad (2)$$

where $\alpha(n)$is a scalar value called the learning rate. The association effect takes place at the neighboring nodes but to a lesser degree because of the Gaussian shape. This adaption procedure stretches the weight vectors of the BMU and its topological neighbors towards the input vector. Presenting similar input vectors to the map provides further activations in the same neighborhood and thereby tends to produce clustering of data in the feature space. Association between neurons decreases during the learning process (the width of the neighborhood function $\phi(n)$ is forced to decrease with n preserving large clusters of data while enabling the separation of clusters that are closely spaced). Ultimately, this training process results in a topology where similarities among data patterns are mapped into similar weights of the neighboring neurons, and the asymptotic local density of the weights approaches that of the training set (Ritter and Schulten, 1986).

Following training, the missing (unsampled) or future data values can be estimated based on distances among the available model vectors (Wang, 2003; Kalteh et al., 2008). In the traditional approach, estimates of values are taken directly from the prototype vectors of the best matching units (Fessant and Midenet, 2002; Wang, 2003). Often-times certain training data sets result in biased estimates (Dickson and Giblin, 2007; Malek et al., 2008) requiring a modified scheme that incorporates bootstrapping (Breiman, 1996), ensemble average (Rallo et al., 2002), or nearest neighbor (Malek et al., 2008). This study uses an alternative iterative estimation scheme that minimizes the topological error vector (Fessant and Midenet, 2002). The estimation of values (often referred to as imputation) is done simultaneously for all SOM variables. For more details about SOM training and estimation, the reader is referred to (Vesanto, and Alhoniemi, 2000; Kohonen, 2001).

2.2. Model diversity and performance

According to Rallo et al. (2002), one of the elements necessary for accurate SOM applications is model diversity. Model diversity reflects the incorporation of measurement information characterizing data variables across multiple spatial and temporal gradients. One means of exploring spatiotemporal relations in environmental studies is to implement a gradient study design. In implementing a gradient