6-27-2011

Subsurface characterization of groundwater contaminated by landfill leachate using microbial community profile data and a nonparametric decision-making process

Andrea R. Pearce
University of Vermont

Donna M. Rizzo
University of Vermont

Paula J. Mouser
University of Maine

Recommended Citation

Pearce AR, Rizzo DM, Mouser PJ. Subsurface characterization of groundwater contaminated by landfill leachate using microbial community profile data and a nonparametric decision-making process. Water Resources Research. 2011 Jun;47(6).

Subsurface characterization of groundwater contaminated by landfill leachate using microbial community profile data and a nonparametric decision-making process

Andrea R. Pearce,[1] Donna M. Rizzo,[1] and Paula J. Mouser[2,3]

Received 9 September 2010; revised 26 February 2011; accepted 6 April 2011; published 15 June 2011.

[1] Microbial biodiversity in groundwater and soil presents a unique opportunity for improving characterization and monitoring at sites with multiple contaminants, yet few computational methods use or incorporate these data because of their high dimensionality and variability. We present a systematic, nonparametric decision-making methodology to help characterize a water quality gradient in leachate-contaminated groundwater using only microbiological data for input. The data-driven methodology is based on clustering a set of molecular genetic-based microbial community profiles. Microbes were sampled from groundwater monitoring wells located within and around an aquifer contaminated with landfill leachate. We modified a self-organizing map (SOM) to weight the input variables by their relative importance and provide statistical guidance for classifying sample similarities. The methodology includes the following steps: (1) preprocessing the microbial data into a smaller number of independent variables using principal component analysis, (2) clustering the resulting principal component (PC) scores using a modified SOM capable of weighting the input PC scores by the percent variance explained by each score, and (3) using a nonparametric statistic to guide selection of appropriate groupings for management purposes. In this landfill leachate application, the weighted SOM assembles the microbial community data from monitoring wells into groupings believed to represent a gradient of site contamination that could aid in characterization and long-term monitoring decisions. Groupings based solely on microbial classifications are consistent with classifications of water quality from hydrochemical information. These microbial community profile data and improved decision-making strategy compliment traditional chemical groundwater analyses for delineating spatial zones of groundwater contamination.

Citation: Pearce, A. R., D. M. Rizzo, and P. J. Mouser (2011), Subsurface characterization of groundwater contaminated by landfill leachate using microbial community profile data and a nonparametric decision-making process, *Water Resour. Res.*, *47*, W06511, doi:10.1029/2010WR009992.

1. Introduction

[2] Decision-making tools for contaminated aquifers have historically been based on physiochemical relationships that largely ignore the influence of microbial community dynamics on subsurface biogeochemistry and attenuation [*American Society of Civil Engineers*, 2003]. Microbial biodiversity in groundwater and soil environments presents a unique opportunity for characterizing and monitoring multicontaminant sites (e.g., waste disposal sites) because their abundance and activity integrate the complex amalgamation of contamination, nutrients, site hydrogeology, and subsurface biogeochemical conditions not reflected by individual physiochemical parameters [*Griebler and Lueders*, 2009]. This particular need is epitomized at municipal solid waste landfill (hereafter referred to as "landfill") sites, where disposal impacts are notoriously difficult to detect and monitor in groundwater networks because of the life span of the degrading waste materials [*Barlaz et al.*, 2002], the variability in the chemical composition of landfill leachate [*Kjeldsen et al.*, 2002], and its potential biogeochemical effects in subsurface environments [*Christensen et al.*, 2001]. Landfills pose human and environmental health risks when leachate infiltrates through solid waste and liner materials into the surrounding subsurface environment. Although regulations have been in place for more than 20 years that protect groundwater resources from impacts of solid waste disposal activities (Resource Conservation and Recovery Act of 1976 (Solid Waste Disposal Act) (42 U.S.C. §§ 6901–6992, 1976); Solid Waste Management, as defined in the U.S. Code of Federal Regulations (40 CFR 239–256, 1991)), waste disposal sites still constitute a significant and continued threat to water resources in the United States [*U.S. Environmental Protection Agency* (*EPA*), 2000].

[1] School of Engineering, University of Vermont, Burlington, Vermont, USA.

[2] Department of Civil and Environmental Engineering, University of Maine, Orono, Maine, USA.

[3] Now at Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, Columbus, Ohio, USA.

0043-1397/11/2010WR009992

[3] The inherent variability of groundwater quality at disposal sites necessitates the tracking and statistical analysis of dozens of hydrochemical variables that might be indicative of degraded water quality [*EPA*, 2009]. In addition, groundwater samples are often autocorrelated in space and time and thereby violate underlying assumptions associated with traditional statistical techniques such as independence, normality, and equal variance [*Gibbons*, 1990; *EPA*, 2009]. Because a priori grouping of monitoring wells for statistical trend comparisons between background and potentially impacted locations is often not feasible, better methods are needed for classifying groundwater quality and tracking attenuation trends at landfill sites. The microbial ecology is strongly influenced by subsurface biogeochemical processes, particularly in aquifers contaminated by organic analytes [*Anderson and Lovley*, 1997; *Chapelle*, 2000; *Griebler and Lueders*, 2009]. For example, changes in the abundance and diversity of microbial community members mediating primary terminal electron accepting processes are distinctly evident in groundwater that has been contaminated by nutrient-rich landfill leachate [*Beeman and Suflita*, 1987; *Ludvigsen et al.*, 1999; *Cozzarelli et al.*, 2000; *Christensen et al.*, 2001; *Röling et al.*, 2001]. These complex linkages make microbial biodiversity a useful tool for characterization and monitoring purposes.

[4] Molecular genetic techniques now allow for rapid profiling of the microbial community in pristine and contaminated subsurface environments [*Madsen*, 2000; *Lovley*, 2003; *Weiss and Cozzarelli*, 2008]. Such biotechnology has been applied to characterizing the ecology and biodegradation potential of microbes in groundwater aquifers impacted by organics, metals, and landfill leachate contaminants [*Watanabe et al.*, 2000; *Röling et al.*, 2001; *Holmes et al.*, 2002; *Akob et al.*, 2007; *Brielmann et al.*, 2009]. However, few interpolation or spatial modeling techniques have incorporated community-level genetic data at the field site scale [*Mouser et al.*, 2005; *Besaw and Rizzo*, 2007] because of the high dimensionality, variability, and complex relationships to other physiochemical parameters. Multivariate computational approaches, such as factor analysis, multidimensional analysis, nonmetric multidimensional scaling, and self-organizing maps (SOM) have been used in the past to investigate links between water quality and microbial community dynamics in environmental systems [*Dollhopf et al.*, 2001; *Fields et al.*, 2006; *Feris et al.*, 2009; *Stein et al.*, 2010]. Principal component analysis appears especially useful for assessing multivariate correlations between hydrochemical and microbial information from leachate-contaminated sites [*Ludvigsen et al.*, 1997; *Röling et al.*, 2001; *Mouser et al.*, 2005, 2010], but to date, these computational approaches have largely focused on parametric or linear-based methods.

[5] Clustering methods are particularly attractive for exploring interrelationships among data to make an initial evaluation of the overall organization because they do not require that a target number of groupings or the data structure be specified prior to the analysis [*Jain et al.*, 1999]. However, as a statistical tool, clustering methods do not optimize the number of or assign significance to the clusters generated. Strategies to optimize the number of clusters in a data set generally maximize variability between clusters and minimize variability within clusters [*Milligan and Cooper*, 1985; *Caliński and Harabasz*, 1974] and include the gap statistic [*Tibshirani et al.*, 2001], the Davies and Bouldin index [*Davies and Bouldin*, 1979], or a multivariate analysis of variance (MANOVA) [*Reyjol et al.*, 2005; *Park et al.*, 2006]. A comparison of the groupings created by an individual clustering method (i.e., specifying the number of clusters a priori in successive runs) and between several clustering methods can highlight similarities between multiple potential groupings, if any exist [*Monti et al.*, 2003; *Whitfield et al.*, 2006], and can allow expert knowledge to help guide decisions, a particularly useful tool for the stewardship of contaminated sites.

[6] Nonlinear clustering methods have been shown to account for more data variability than linear methods when applied to hydrochemical and microbial data sets [*Schryver et al.*, 2006]. The SOM, or Kohonen map, is a nonlinear and nonparametric clustering artificial neural network that outperforms many traditional clustering methods (e.g., hierarchical and *k*-means) on data sets with high dispersion, outliers, irrelevant variables, and nonuniform cluster densities [*Kohonen*, 1990; *Mangiameli et al.*, 1996]. Meaningful groupings have been created with the SOM when applied to biological community data [*Giraudel and Lek*, 2001; *Céréghino et al.*, 2005], biogeochemical data [*Solidoro et al.*, 2007], and molecular genetic data [*Dollhopf et al.*, 2001].

[7] Considering the need for improved computational approaches that incorporate variable, highly dimensional microbial data from subsurface environments, we present a nonparametric decision-making strategy to characterize a gradient of water quality in a groundwater aquifer impacted by landfill leachate based solely on the clustering of microbial community data generated from terminal restriction fragment length polymorphism (TRFLP) profiles of the 16S rRNA gene. Our approach compresses the microbial data, modifies the SOM to weight input variables, and then provides statistical guidance for classifying samples into an optimal number of groupings. For comparison, we present results of other clustering techniques (*k*-means and hierarchical methods) and compare microbial classifications to hydrochemical information at the site. This data analysis strategy is designed to guide more systematic, efficient, and effective characterization and monitoring of aquifers contaminated by multiple pollutants using biological-based information.

2. Methodology

2.1. Study Area and Field Data Collection

[8] Field data were collected from 22 groundwater monitoring wells surrounding the Schuyler Falls Sanitary Landfill, a 30 acre (12.14 hectare) unlined landfill in Clinton County, New York (Figure 1), where municipal, commercial, and industrial wastes were deposited between 1977 and 1996 [*Barton and Loguidice*, 1996]. The closed and capped landfill is situated on 15–40 m thick (west to east) Pleistocene age till and outwash soils that overlay dolomite bedrock [*Barton and Loguidice*, 1996]. Advective groundwater transport rates in the sandy soils are estimated at 25 m yr^{-1} northwest toward the Saranac River.

[9] Subsurface contamination from landfill leachate was discovered when anthropogenic organic compounds were detected in down-gradient monitoring wells. Further hydrogeochemical investigations and a detailed subsurface

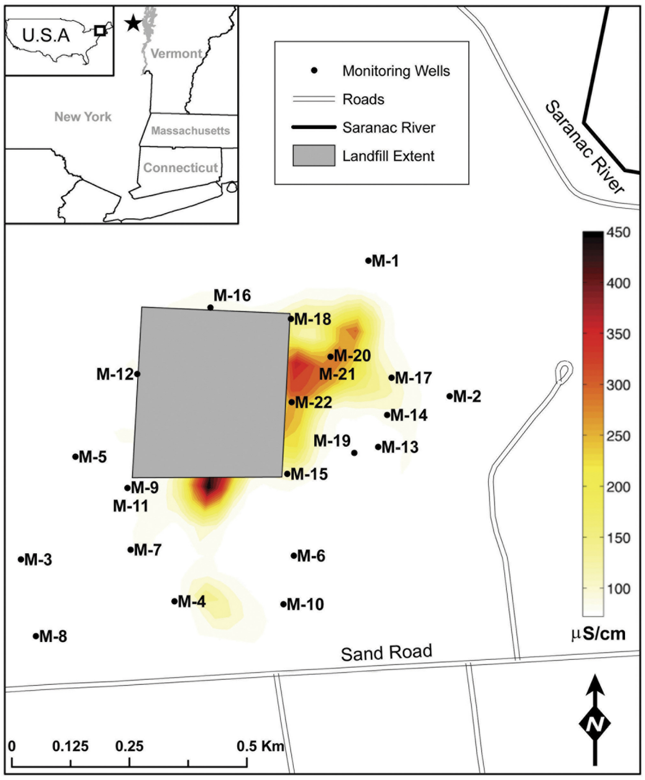


Figure 1. Schuyler Falls Landfill, Clinton County, New York (44.694° N, 73.597° W), site location map and approximate extent of three-dimensional subsurface contamination in plan view. The plume is estimated using conductivity measurements collected from surface electromagnetic surveys (EM-34) and interpolated using the method of ordinary kriging.

electromagnetic resistivity survey helped characterize the spatial extent of subsurface contamination. Indications of leachate contamination include analytical detection of multiple halogenated volatile organic compounds, petroleum by-products, elevated specific conductance, alkalinity, total organic carbon, and other inorganic constituents in downgradient monitoring wells [*Barton and Loguidice*, 1993]. Given multiple types of point measurements, each indicating impact by leachate, many strategies exist for characterizing the three-dimensional extent of contamination at the site.

[10] Subsurface resistivity, consisting of 10 and 20 m horizontal and vertical dipole measurements (four separate surveys with 664 total survey points), was used in this work to estimate the overall magnitude and extent of contamination. Data from the four surveys were interpolated independently using the method of ordinary kriging over a 1064 × 1158 m grid with 30.5 m spacing. The inverse of electromagnetic resistivity values (electrical conductivity) from each of the four interpolated surveys is used to create a conservative, two-dimensional plan view image of overall groundwater contamination (Figure 1). The approximate extent of contamination shows migration in the direction of groundwater flow toward the Saranac River.

[11] The sampling techniques, laboratory methods, and data reduction methodology used for creating microbial community profiles from groundwater samples are described in detail by *Mouser et al.* [2010]. To summarize, bailed groundwater samples were collected from 22 monitoring locations (screened wells) once field parameters stabilized (temperature, turbidity, oxidation reduction potential, pH, and conductance). Samples for hydrochemical analysis were placed on ice and transported overnight to the laboratory and analyzed for specific conductivity (EPA Method 210.1), alkalinity (EPA Method 310.2), ammonia (EPA Method 350.1), iron (EPA Method 200.7), phenols (EPA Method 420.4), biochemical oxygen demand (BOD) (BOD5 test, Standard Method 5210B) [*Clesceri et al.*, 1998], and chemical oxygen demand (COD) (EPA Method 410.4). The 500 mL samples for microbial community analysis were placed immediately on ice and transported to the University of Vermont, where they were pelleted by centrifugation at 20,000 rpm, flash frozen, and stored at −20°C until further extraction.

[12] Nucleic acids were extracted using a MoBio Powersoil DNA Isolation Kit (MoBio Laboratories, Carlsbad, California). Polymerase chain reaction (PCR) amplification of the 16S rRNA gene was conducted using three primer sets targeting Archaea (46F/907R) [*Lane et al.*, 1985; *Ovreås et al.*, 1997], Bacteria (8F/1392R) [*Lane et al.*, 1985], and Geobacteraceae (8F/825F) [*Snoeyenbos-West et al.*, 2000] using reagents and cycling parameters described by *Mouser et al.* [2010]. TRFLP profiles were digested using the *Msp*I restriction enzyme, and digests were quantified in triplicate using capillary electrophoresis (ABI Prism 3100-Avant Genetic Analyzer, Applied Biosystems) at the University of Vermont DNA Analysis Center. TRFLP profiles were analyzed for size calling determinations and minimum fluorescence intensity using GeneMapper software (Applied Biosystems) (see *Mouser et al.* [2010] for a

detailed description of TRFLP digestion methods and fragment binning methods). A total of 40, 115, and 54 terminal restriction fragments representing the relative abundance of Archaea, Bacteria, and Geobacteraceae community members, respectively, were identified across the 22 monitoring locations.

2.2. Nonparametric Decision-Making Process

[13] The following outlines a decision-making process to characterize a water quality gradient across the Schuyler Falls Landfill site using only the microbial community profile data collected from screened groundwater wells. Our methodology uses a nonparametric SOM clustering method in tandem with a nonparametric MANOVA to guide the selection of an appropriate number of groupings. Input data are preprocessed using a nonparametric form of principal components analysis (PCA). The SOM has been modified to allow input variables to be weighted by their relative importance, the percent variance explained by each principal component (PC).

2.2.1. Data Preprocessing

[14] The first step in our methodology is performing a principal components analysis on the combined TRFLP profiles from the monitoring wells to make orthogonal variables and reduce the dimensionality of the input data set. For comparison purposes, PCs were created from both the covariance matrix and the nonparametric Spearman's rank correlation matrix (Figure 2, step 1). The first 21 PCs are unique, explain 100 percent of the variance of the original 209 variable and therefore were all retained in the analysis. Both sets of resulting PC scores are normalized independently between 0 and 1 as

$$\mathbf{x}^{\mathbf{p}}(i) = \{(i) - \min[\mathbf{x}(i)]\}/\{\max[\mathbf{x}(i)] - \min[\mathbf{x}(i)]\},$$
$$p = 1, 2, ..., 22,$$

where $i = 1$–21, the number of PCs, to ensure that differences in magnitude between the variables do not create unwanted bias within the clustering method and to bound the input between 0 and 1 for use in the SOM. These two sets of normalized PC scores comprise the input data for all further analyses.

2.2.2. Self-Organizing Map With Weighting

[15] Next, we cluster the data using an SOM that has been modified to weight input variables by their relative importance. The SOM algorithm is a single-layer (of weights) network developed by *Kohonen* [1990] (Figure 2, step 2). Input patterns $\mathbf{x}^{\mathbf{p}}(i)$ (where $p = 1, 2, \ldots, P$

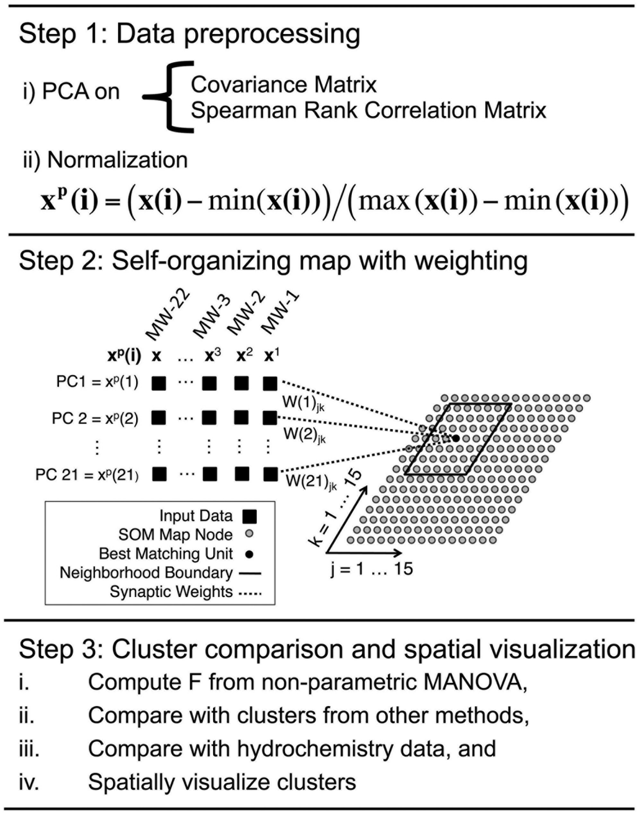

Step 1: Data preprocessing
i) PCA on
Covariance Matrix
Spearman Rank Correlation Matrix
ii) Normalization
$\mathbf{x}^{\mathbf{p}}(\mathbf{i}) = (\mathbf{x}(\mathbf{i}) - \min(\mathbf{x}(\mathbf{i})))/(\max(\mathbf{x}(\mathbf{i})) - \min(\mathbf{x}(\mathbf{i})))$
Step 2: Self-organizing map with weighting
MW-22
MW-3
MW-2
MW-1
$\mathbf{x}^{\mathbf{p}}(\mathbf{i})$ $\mathbf{x}$ … $\mathbf{x}^3$ $\mathbf{x}^2$ $\mathbf{x}^1$
PC1 = $x^p(1)$
PC 2 = $x^p(2)$
PC 21 = $x^p(21)$
$W(1)_{jk}$
$W(2)_{jk}$
$W(21)_{jk}$
k = 1 … 15
j = 1 … 15
Input Data ■
SOM Map Node ○
Best Matching Unit ●
Neighborhood Boundary —
Synaptic Weights ·····
Step 3: Cluster comparison and spatial visualization
i. Compute F from non-parametric MANOVA,
ii. Compare with clusters from other methods,
iii. Compare with hydrochemistry data, and
iv. Spatially visualize clusters

Figure 2. Nonparametric decision-making process. In step 1, data preprocessing, data compression is performed by principal component analysis using a parametric (covariance) or nonparametric (Spearman's rank) correlation matrix. Variables (principal component scores, in this case) were normalized independently. Step 2 is self-organizing map (SOM) with weighting. Step 3, cluster comparison and spatial visualization, is a comparison of the SOM-generated clusters with other clustering methods using an *F* statistic computed by a nonparametric multivariate analysis of variance to examine consensus groupings of the data.

represents a particular monitoring well) are presented to the network and self-organize into clusters during an unsupervised training procedure. In this work, each input pattern (vector $\mathbf{x}^\mathbf{p}$) is fully connected by synaptic weights $\mathbf{w}_{jk}$ to a two-dimensional grid of output nodes, where j and k are indices that map to the output grid. A single bundle $\mathbf{w}(i)_{jk}$ is a vector that has as many components as the number of input data types (i). The initial values of the synaptic weights are constructed by adding small random values (plus or minus up to 5% of the input parameter mean) to each of the original input parameter means.

[16] Unsupervised clustering begins by calculating the distance between an input pattern $\mathbf{x}^\mathbf{p}$ and each of the synaptic bundles $\mathbf{w}_{jk}$ (Euclidian distance is used here, although other distance measures may be used). The weight vector $\mathbf{w}_{jk}$ at node (j, k) with the minimum Euclidian distance to the input pattern $\mathbf{x}^\mathbf{p}$ is selected as the best matching unit (BMU).

[17] The weights associated with the BMU and surrounding neighborhood (here neighborhood is defined by radius b (Figure 2, step 2), normalized by map size) are updated according to the rule

$$\mathbf{w}_{jk}^{\mathrm{new}} = \mathbf{w}_{jk}^{\mathrm{old}} + \alpha(\mathbf{x}^\mathbf{p} - \mathbf{w}_{jk}^{\mathrm{old}}),$$

where α is a learning parameter ranging between 0 and 1 and $\mathbf{x}^\mathbf{p}$ is the current input pattern p.

[18] *Vesanto et al.* [2000] describe an input mask for the SOM as part of the Euclidian distance calculation used to determine the BMU, applying weights to individual input variables depending on their relative importance:

$$\left\{\sum_{i=1}^{I} s(i)\left[x(i) - w(i)_{jk}\right]^2\right\}^{0.5},$$

where $s(i)$ is a scalar value for each input variable. However, a corresponding update rule is not provided in their work. As a result, we modified the original Kohonen update rule as follows:

$$\mathbf{w}_{jk}^{\mathrm{new}} = \mathbf{w}_{jk}^{\mathrm{old}} + \mathbf{s}\alpha(\mathbf{x}^\mathbf{p} - \mathbf{w}_{jk}^{\mathrm{new}}),$$

where the vector $\mathbf{s}$ weights each of the input PC scores by the percent variance explained by each score.

[19] A single training iteration is complete after each of the P input patterns have been presented in random order to the network and the appropriate weights have been updated. The SOM was trained in two phases, an ordering phase and a fine-tuning phase. During the ordering phase, the size of the neighborhood b and the learning parameter α decreased exponentially, from 0.4 toward 0 over 300 training iterations. During the fine-tuning phase, composed of an additional 400 iterations, b and α decreased linearly from 0.3 to 0.05.

[20] The SOM clusters the highly dimensional microbial data onto a two-dimensional output map (in our case, a 20 × 20 grid of output nodes). Visualization of the clusters is aided by what is known as a unified distance matrix (**U** matrix). For each node of the 20 × 20 output map, the values of the **U** matrix are computed as the average Euclidian distance to all adjacent nodes. In this work, we also use the SOM to classify the data into a predefined number of clusters (such as with k-means clustering) by specifying a priori a small number of output nodes, equal to the number of desired clusters.

2.2.3. Cluster Comparison and Spatial Visualization

[21] We next compute an F statistic, or ratio of the between-group variance to the within-group variance, using a nonparametric MANOVA [*Anderson*, 2001; *McArdle and Anderson*, 2001]. The F statistic is used to compare the efficiency of group separation for different clustering methods, including the weighted SOM, unweighted SOM, hierarchical, and k-means (Figure 2, step 3). This particular nonparametric MANOVA allows for the use of any distance metric in defining the distance between samples (Euclidian distance here) and is appropriate for data that do not meet the assumptions necessary for parametric tests. We compare the results of the four clustering methods (i.e., weighted SOM, unweighted SOM, hierarchical, and k-means) to a summary of site hydrochemistry data and classifications presented by *Mouser et al.* [2010]. Finally, we superimpose the clustered wells onto a site map for visualization. Hierarchical and k-means clustering were performed using JMP 9.0.0 (SAS Institute, Cary, North Carolina); the SOM clustering methods and nonparametric MANOVA were implemented by the author using MATLAB R2010a (The Mathworks, Natick, Massachusetts).

3. Results

[22] Figures 3 (left) and 3 (right) show an example of the **U** matrices generated from a weighted and unweighted SOM using the Spearman's rank TRFLP PC scores as input. The **U** matrices highlight the SOM's self-organization of the 22 monitoring wells prior to discrete clustering. Low values are represented as contiguous light shaded areas of the map and indicate clusters or regions of similarity. High values (darker colors) indicate steep boundaries between groupings and are highlighted by hand-drawn lines on the weighted SOM **U** matrix (Figure 3, left). Black dots labeled M1–M22 mark the final BMU for each of the 22 monitoring wells. While there is no physical meaning to the specific j-k placement of the monitoring wells on either of the **U** matrices, the organization (clustering of wells) from the weighted SOM shows more definitive boundaries between groupings than the unweighted SOM.

[23] A comparison of the SOM clustering results with other commonly used clustering methods is presented in Figure 4a. Figure 4a is organized by the type of matrix used in PCA preprocessing (covariance matrix versus Spearman's rank correlation matrix) and by the type of clustering method (i.e., hierarchical, k-means, SOM, and weighted SOM). For each preprocessed clustering scenario, we provide the nonparametric MANOVA F statistic for two, three, or four clusters. Higher values of F indicate better separation between clusters, and all F statistics were statistically significant when compared to a distribution created from random permutations of the data. The wells (M1–M22) are color-coded to identify separation into two, three, or four clusters.

[24] The clustered groupings of Figure 4a are compared with the site hydrochemical information; a subset of this site information is summarized in Figure 4b. Low (L), medium (M), and high (H) divisions were created for key dissolved constituents (specific conductance and alkalinity),

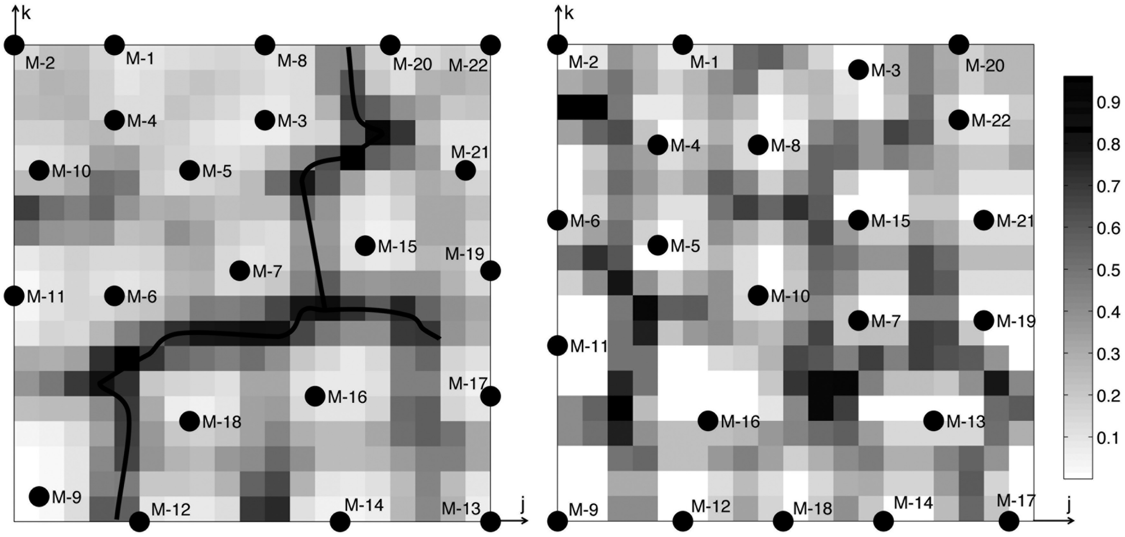


Figure 3. Unified distance matrix (**U** matrix) from (left) weighted and (right) unweighted SOMs using nonparametric (Spearman's rank) principal component scores as input. The **U** matrices display the final organization of the data on the output map. Black lines highlight the dark regions that separate similar groupings on the map. Division between clusters is more evident with the weighted SOM than the unweighted SOM, and no attempt has been made to draw boundaries between clusters.

leachate contaminants (ammonia, COD, and phenols), and biological activity (ORP, Fe, and BOD) on the basis of detection limits and observed breakpoints in data histograms. Individual hydrochemical variables create somewhat different classifications of contamination; however, locations M12–M22 generally classify as M or H, which correspond well to rust or brown clustered groupings.

[25] In addition, *Mouser et al.* [2010] used the site hydrogeology and hydrochemistry to divide water quality into three categories. The bottom row of Figure 4b identifies wells M1–M11 as background (B), M12–M19 as fringe (F), and M20–M22 as contaminated (C). Specific conductivity, alkalinity, total phenols, BOD, and COD values showed the background wells to be statistically different from the fringe and contaminated wells ($p < 0.05$ Tukey-Kramer test for multiple comparisons among means). Iron, ammonia, and ORP values were significantly different between wells with B or F designations and the C wells.

[26] Figure 4a indicates more agreement across impacted monitoring locations when the microbiological input data are preprocessed using a nonparametric Spearman's rank PCA (e.g., highlighted by rust and brown colored monitoring locations M12–M22). These same locations show more variability across and within clustering methods when the input data are preprocessed using the parametric covariance PCA. Figure 4a also shows that the largest F statistics are associated with three clusters, six out of eight times. Note that although the same input data (i.e., molecular genetic based microbial community profiles) are presented to each of the clustering scenarios, the F statistic may only be compared between clustering methods that use the same form of input data (e.g., F results cannot be compared between covariance and Spearman's rank PCA or with clusters created by the weighted SOM). Double lines on Figure 4a separate results where F is comparable.

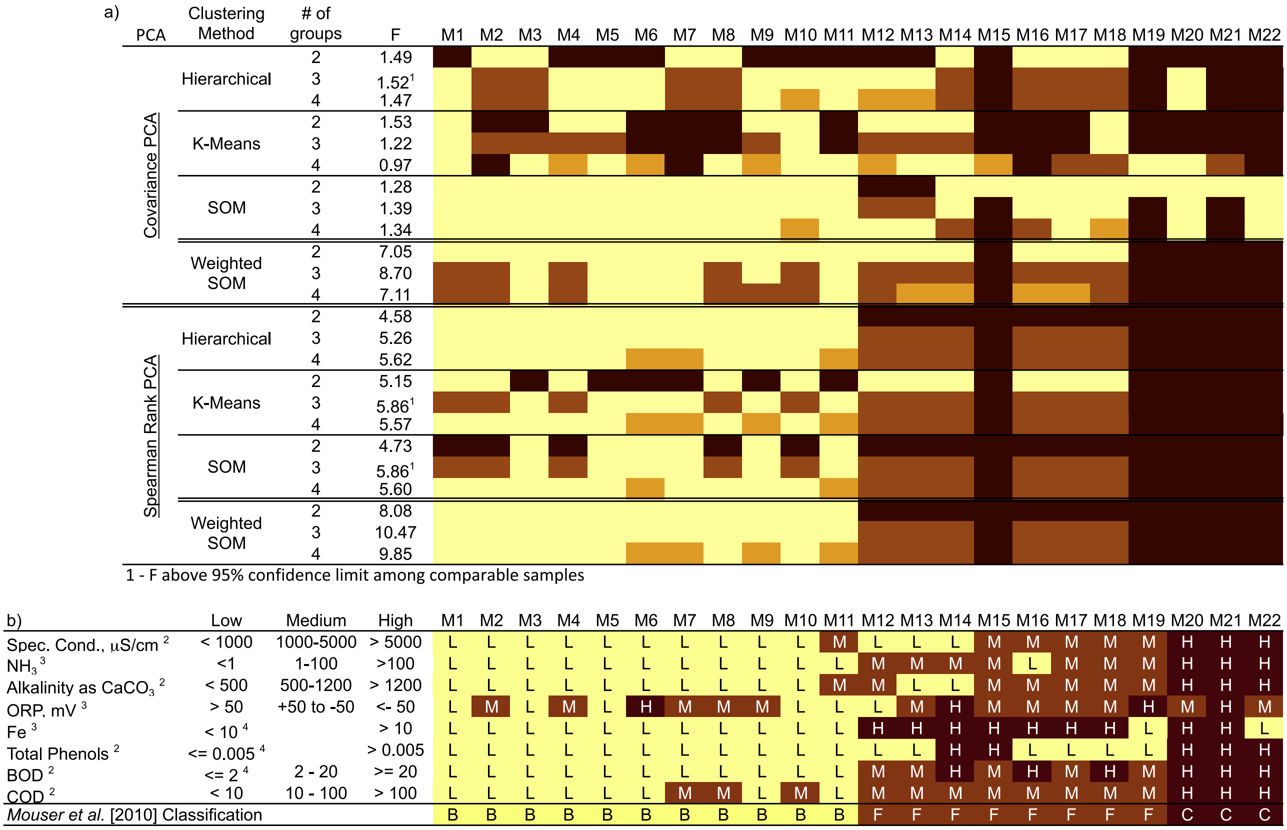
[27] As the number of clusters in Figure 4a increases from two to four using the nonparametric Spearman's rank preprocessing, the monitoring locations are progressively classified across the known contamination gradient (i.e., water quality classified as background and fringe, low and high levels of contamination, etc.). The hierarchical clustering and weighted SOM provide the best overall match to both the hydrochemistry data and the *Mouser et al.* [2010] classifications.

[28] The spatial locations of the clustered wells are shown in plan view along with a conservative 2-D estimate of the contaminant plume (Figures 5 and 6). The well location color aligns with the cluster assignment for key scenarios discussed for Figure 4a. Visual inspection of the changes in monitoring well classification generated with the weighted SOM and inputs preprocessed using the Spearman's rank PCA show progressively more resolution in the plume fringe as the number of clusters increases from two to four (Figure 5). Figure 6a shows the hydrochemistry classifications of *Mouser et al.* [2010] with the results of the three clustering methods (all input data preprocessed using the nonparametric Spearman's rank PCA) that generated the highest F statistic (Figure 6b). Hierarchical clustering results with four clusters are shown in Figure 6b (left), while the k-means and SOM results (three clusters) are shown in Figure 6b (right).

4. Discussion

4.1. Selecting Input Data Structure and the Optimal Number of Output Clusters

[29] This nonparametric decision-making process allows characterization of a water quality gradient in leachate-contaminated groundwater using microbial community profiles. We use a nonparametric MANOVA for guidance in



PCA	Clustering Method	# of groups	F
Covariance PCA	Hierarchical	2	1.49
		3	1.52[1]
		4	1.47
	K-Means	2	1.53
		3	1.22
		4	0.97
	SOM	2	1.28
		3	1.39
		4	1.34
	Weighted SOM	2	7.05
		3	8.70
		4	7.11
Spearman Rank PCA	Hierarchical	2	4.58
		3	5.26
		4	5.62
	K-Means	2	5.15
		3	5.86[1]
		4	5.57
	SOM	2	4.73
		3	5.86[1]
		4	5.60
	Weighted SOM	2	8.08
		3	10.47
		4	9.85

1 - F above 95% confidence limit among comparable samples

b)	Low	Medium	High	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22
Spec. Cond., μS/cm [2]	< 1000	1000-5000	> 5000	L	L	L	L	L	L	L	L	L	L	M	L	L	L	M	M	M	M	M	H	H	H
NH_3 [3]	<1	1-100	>100	L	L	L	L	L	L	L	L	L	L	L	M	M	M	M	L	M	M	M	H	H	H
Alkalinity as $CaCO_3$ [2]	< 500	500-1200	> 1200	L	L	L	L	L	L	L	L	L	L	M	M	L	L	M	M	M	M	M	H	H	H
ORP, mV [3]	> 50	+50 to -50	<- 50	L	M	L	M	L	H	M	M	M	L	L	L	M	H	M	M	M	M	H	M	H	M
Fe [3]	< 10 [4]		> 10	L	L	L	L	L	L	L	L	L	L	L	H	H	H	H	H	H	H	L	H	H	L
Total Phenols [2]	<= 0.005 [4]		> 0.005	L	L	L	L	L	L	L	L	L	L	L	L	L	H	H	L	L	L	L	H	H	H
BOD [2]	<= 2 [4]	2 - 20	>= 20	L	L	L	L	L	L	L	L	L	L	L	M	M	H	M	H	M	H	M	H	H	H
COD [2]	< 10	10 - 100	> 100	L	L	L	L	L	L	M	M	L	M	L	M	M	M	M	M	M	M	M	H	H	H
Mouser et al. [2010] Classification				B	B	B	B	B	B	B	B	B	B	B	F	F	F	F	F	F	F	F	C	C	C

2 - Indicates a significant difference was observed between B and F, C locations ($p < 0.05$ Tukey-Kramer test for multiple comparisons among means)

3 - Indicates a significant difference was observed between B,F and C locations ($p < 0.05$ Tukey-Kramer test for multiple comparisons among means)

4 - Detection limit

Figure 4. (a) Microbial terminal restriction fragment length polymorphism data clustering and multivariate analysis of variance results. The clusters and *F* statistic are shown for each combination of preprocessing and clustering for two, three, and four clusters for monitoring locations M1–M22. Values of *F* can only be compared between clusters created with the same input. Double horizontal lines separate regions between which *F* cannot be compared. Clusters are indicated by color. (b) Hydrochemistry summary including background (B), fringe (F), and contaminated (C) classes created by *Mouser et al.* [2010]. The hydrochemistry classes, low, medium, and high, were defined by looking for naturally occurring breakpoints in a histogram of the data. The upper limit on the low category for phenols, dissolved iron (Fe), and biological oxygen demand (BOD) are defined as the detection limit of the analysis (i.e., all low values were below detection). All values are in mg/L unless otherwise noted. The categories assigned by *Mouser et al.* [2010] were created prior to any clustering analysis and represent classification of the 22 monitoring wells on the basis of an electromagnetic survey and groundwater hydrochemistry information.

selecting the number of clusters to consider, a consensus approach (Figure 4), and expert knowledge of the site to divide the set of monitoring wells into useful management zones (designations of background, fringe, and contaminated locations). Many methods exist to guide the selection of an optimal number of clusters within a data set, but there is no "correct" number, unless it is has been defined a priori. There is likely significant meaning to grouping the data into two, three, or four clusters in this landfill application, and "optimal" groupings will depend on the management objectives or the reason for the analysis. Our *F* statistic suggests, over repeated analyses, that three clusters provide the most significant division of the data set, and these categories are supported by hydrochemical data and existing knowledge of the site [*Mouser et al.*, 2010].

[30] On the basis of the site hydrochemistry, we could speculate that division into two groups would characterize microbes present in background versus leachate-impacted locations; classification into three groups might describe microbes that thrive in background areas, along the plume fringe, or in more heavily impacted source areas. Alternatively, distribution into four or five groups might describe microbes mediating the dominant terminal electron-accepting processes. As a result, the management or research objectives need to be considered when analyzing (or preprocessing) these microbiological data. Interestingly, consideration of more than four clusters using the available site data resulted in at least one group with only one data point, thus artificially raising *F* and overfitting the data (results not shown). The latter suggests there is no mechanistic reason for considering a larger number of categories. It should be noted that most landfill detection or long-term monitoring networks are unlikely to have more than the 22 monitoring wells provided in this application.

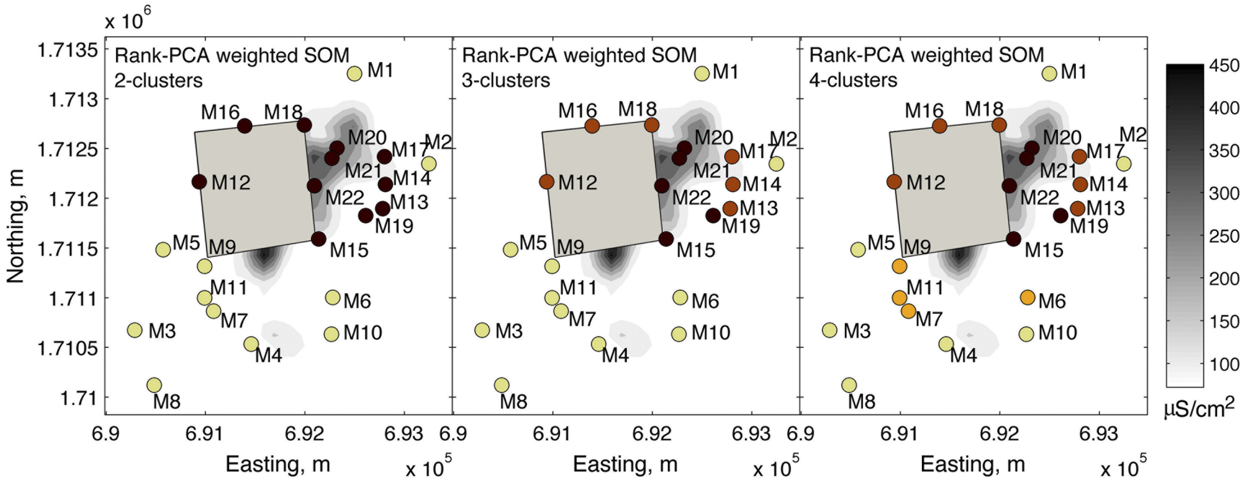


Figure 5. Spatial arrangement of clusters from the weighted self-organizing map using the nonparametric Spearman's rank principal component scores as input for (left) two groups, (middle) three groups, and (right) four groups. Well locations (M1–M22) are color-coded on the basis of the group they fall within, and colors correspond to those output clusters shown in Figure 4.

[31] Part of the attraction of using a nonparametric clustering method such as the SOM is the flexibility of the input data types and relaxation of the assumptions required of most parametric statistical techniques. Preliminary work showed the cluster results are sensitive to the input data structure. As a result, we considered several configurations for preprocessing the microbial community input data, including using (1) raw TRFLP abundance data for all three (Archaea, Bacteria, and Geobacteraceae) microbial communities (209 variables in total), (2) PCA to reduce the 209 community profiles to 21 PCs (21 total variables), and (3) PCA separately on each of the Archaea, Bacteria, and Geobacteraceae community profiles; the latter were concatenated (3 × 21) into 63 total variables. Our initial attempts to use the raw microbial community profile data (i.e., not preprocessed with PCA) did not produce groupings consistent with the site hydrochemistry. Because our objective was to describe a gradient of water quality impacts, preprocessing the combined microbial profiles using PCA (item 2) was the most appropriate method. We also compared the use of PC scores created from parametric (covariance matrix) and nonparametric (correlation matrix) methods. We chose to test PC scores from a covariance matrix because we believe the relative abundance of the organisms to be important. However, output clusters for impacted locations (M12–M22) are more consistent across all four clustering methods when the input data are preprocessed with the nonparametric Spearman's rank correlation matrix, suggesting that choosing a nonparametric preprocessing method when using microbial community profile data is possibly more important than the choice of a specific clustering method.

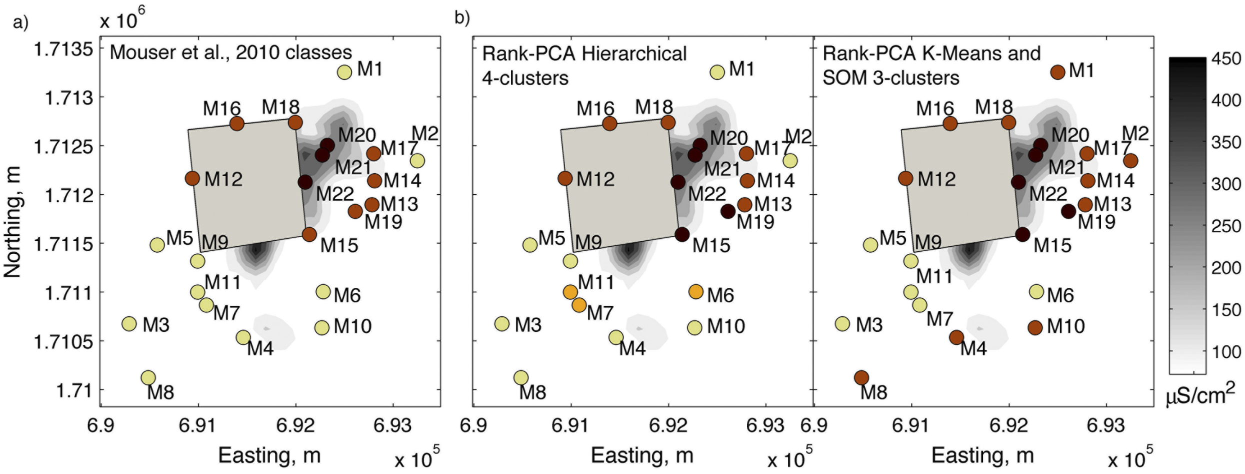


Figure 6. (a) Spatial arrangement of well groupings by *Mouser et al.* [2010] and (b) comparison of output clusters created using nonparametric Spearman's rank principal components. *F* statistics were largest for hierarchical clustering with four groups, while *F* statistics were largest for *k*-means and unweighted self-organizing map clustering with three identical groups. Well locations (M1–M22) are color-coded on the basis of the group they fall within, and colors correspond to those output clusters shown in Figure 4.

4.2. Influence of Microbial Community Data Variability on Classification

[32] The community profiles used in this Schuyler Falls Landfill application are intended to characterize the microbial ecology of dominant and lesser known microorganisms in background and leachate-contaminated groundwater monitoring wells at one snapshot in time and are, by definition, a variable source of data. As such, similarities in TRFLP patterns for clustering purposes may come from the relative abundance of terminal restriction fragments that are shared across monitoring locations and from monitoring locations that share the absence of other fragments observed in the data set. Note that a shared absence is not necessarily indicative of sample similarity but is valuable information nonetheless.

[33] Our modified SOM explores how an input mask that weights input variables (PC scores) by the percent variance explained might influence sample groupings. Clustering methods, including the SOM, are sensitive to the variability of microbial community profiles; therefore, not all members should necessarily be considered equal. Although this nonparametric SOM is more robust to noisy data than most clustering methods, too many irrelevant variables will lower the discriminating power, as is the case with any clustering method [*Mangiameli et al.*, 1996]. Since our objective is to characterize a gradient of water quality impacts, dominant groups of microorganisms responsible for driving larger biogeochemical changes will likely be described by several higher-variance PCs, whereas the less abundant, more biodiverse microorganisms will likely contribute more to the numerous lower-variance PCs. Thus, weighting all of the PC inputs equally in this application could contribute to an incorrect classification or overfitting of the data. Therefore, there is good justification for using the input mask to weight the SOM input by the percent variance explained by each PC score. Although several F values occur above the 95% confidence limit (among comparable results, Figure 4a), there is not enough of a trend to suggest that one clustering method is superior to the others on the basis of these F statistics alone. In this example, the hierarchical clustering and the weighted SOM create nearly identical output clusters when the input data are preprocessed using a nonparametric Spearman's rank PCA. Given that the decision-making methodology outlined here uses entirely nonparametric methods from the preprocessing stage through to the output cluster optimization, we believe that nonparametric weighted SOM will more reliably and systematically characterize microbial community data in other applications.

[34] Although discrepancies exist, the clusters created by this nonparametric methodology generally agree quite well with the site hydrochemistry information. It is difficult to ascertain which monitoring locations have not been impacted by leachate given the waste disposal history and extent of contamination at this site, but there are wells with significantly less contamination. Clustering into two groups (Figures 4 (left) and 5 (left)) defines a division between relatively unimpacted wells and those located within the contaminated or fringe areas of the plume. Figure 5 suggests different conditions prevail up and down gradient of the landfill. Historical remedial efforts at the site include installing extraction wells to the northeast of the landfill to redirect and capture subsurface leachate [*Barton and Loguidice*, 1996]. There is also subsurface contamination extending from the southern boundary of the landfill that has been removed and replaced with clean soil; wells in this region (M6, M7, and M11) appear to cluster differently in some cases.

[35] The spatiotemporal dynamics of microbial communities in aquifers are poorly characterized and influenced at multiple scales [*Griebler and Lueders*, 2009]. While groundwater communities may be spatially correlated at distances greater than 10 m [*Mouser et al.*, 2005], communities extracted from subsurface sediments are thought to exhibit spatial correlation at much smaller distances (less than 1 m) [*Mummey and Stahl*, 2003; *Brad et al.*, 2008]. Thus, considerable variability is likely to exist in community profiles between discrete sampling locations at the landfill site. Attaining vertical resolution of microbial community profiles would be difficult with bailed wells since groundwater samples aggregate vertical regions within the screen and contain biases from the vertical heterogeneity of aquifer transmissivity and variable contaminant concentrations with depth [*Church and Granato*, 1996]. Multilevel samplers are an improvement to characterizing vertical heterogeneity but require more expertise to design and install, are more expensive, and are therefore less common [*Lerner and Teutsch*, 1995]. As such, bailed groundwater samples from screened wells are frequently used to create two-dimensional representations of three-dimensional plumes despite the misrepresentation.

4.3. Implications for Classifying and Long-Term Monitoring at Landfill Sites

[36] Our objective was to extend existing computational methods to guide more systematic, efficient, and effective characterization and monitoring of contaminated aquifers using biologically based information. We show that a decision-making methodology consisting of a nonparametric preprocessing step, weighted SOM clustering, and calculation of a nonparametric F statistic can be used to characterize a water quality gradient in landfill leachate-contaminated groundwater using only microbial community profiles. Microbial community profiles generated using molecular genetic techniques present an opportunity to add value to traditional characterization and monitoring methods because their abundance and activity integrates the complex amalgamation of contamination, nutrients, site hydrogeology, and subsurface biogeochemical conditions not reflected by individual physiochemical parameters.

[37] Biophysiochemical processes are intricately coupled in subsurface systems, yet it is difficult to explicitly describe one as a function of the other on the basis of mechanistic or predictive models. The nonparametric decision-making process outlined here distinguishes between background, fringe, and polluted monitoring wells using only information provided by the communities of microorganisms and provides guidance, suggesting the approximate spatial extent of functional zones of a leachate plume surrounding a landfill. This type of data presents unique challenges that must be respected at each step of the decision-making process, including using nonparametric methods. Modifying the SOM for differential weighting of the input variables allows the clustering method to incorporate the variance explained

by principal components. We believe the differential weighting is necessary to retain the original data structure and that this nonparametric computational methodology is appropriate for microbial community data and compliments standard analytical analyses for the purpose of delineating spatial zones of groundwater contamination.

[38] **Acknowledgments.** This research was supported by VT EPSCoR National Science Foundation grant (NSF EPS 0701410). The authors thank Casella Waste Management, Inc., for assistance with sampling and monitoring well access, Dale Becker with the New York Department of Environmental Conservation for providing access to public data resources and hydrogeology technical insight on the site, and the Clinton County Executive Council for allowing research site access. We also thank Nick Gotelli, Richard Hooper, and Lori Stevens for suggestions with statistical analysis, the VT EPSCoR CSYS group for their feedback on this work, and Lexie Haselton for assistance with Figure 1.

References

Akob, D. M., H. J. Mills, and J. E. Kostka (2007), Metabolically active microbial communities in uranium-contaminated subsurface sediments, *FEMS Microbiol. Ecol.*, *59*, 95–107, doi:10.1111/j.1574-6941.2006.00203.x.

American Society of Civil Engineers (2003), *Long-Term Groundwater Monitoring: The State of the Art*, 103 pp., Reston, Va.

Anderson, M. (2001), A new method for non-parametric multivariate analysis of variance, *Aust. Ecol.*, *26*(1), 32–46, doi:0.1111/j.1442-9993.2001.tb00081.x.

Anderson, R. T., and D. R. Lovley (1997), Ecology and biogeochemistry of in situ groundwater bioremediation, *Adv. Microb. Ecol.*, *15*, 289–350.

Barlaz, M. A., A. P. Rooker, R. Kjeldsen, M. A. Gabr, and R. C. Borden (2002), Critical evaluation of factors required to terminate the postclosure monitoring period at solid waste landfills, *Environ. Sci. Technol.*, *26*, 3457–3464, doi:10.1021/es011245u.

Barton and Loguidice (1993), Clinton County landfill expansion hydrogeologic investigation report, Schuyler Falls Sanit. Landfill, Syracuse, N. Y.

Barton and Loguidice (1996), Clinton County landfill closure—Final closure investigation report, 453 pp., Schuyler Falls Landfill, Syracuse, N. Y.

Beeman, R. E., and J. M. Suflita (1987), Microbial ecology of a shallow unconfined groundwater aquifer polluted by municipal landfill leachate, *Microb. Ecol.*, *14*, 39–54, doi:10.1007/BF02011569.

Besaw, L., and D. Rizzo (2007), Stochastic simulation and spatial estimation with multiple data types using artificial neural networks, *Water Resour. Res.*, *43*, W11409, doi:10.1029/2006WR005509.

Brad, T., B. M. Braster, B. M. Van Breukelen, N. M. van Straalen, and W. F. M. Roling (2008), Eukaryotic diversity in an anaerobic aquifer polluted with landfill leachate, *Appl. Environ. Microbiol.*, *74*(13), 3959–3968, doi:10.1128/AEM.02820-07.

Brielmann, H., C. Griebler, S. I. Schmidt, R. Michel, and T. Lueders (2009), Effects of thermal energy discharge on shallow groundwater ecosystems, *FEMS Microbiol. Ecol.*, *68*, 273–286, doi:10.1111/j.1574-6941.2009.00674.x.

Caliński, T., and J. Harabasz (1974), A dendrite method for cluster analysis, *Commun. Stat. Simul. Comput.*, *3*(1), 1–27, doi:10.1080/03610917408548446.

Céréghino, R., F. Santoul, A. Compin, and S. Mastrorillo (2005), Using self-organizing maps to investigate spatial patterns of non-native species, *Biol. Conserv.*, *125*(4), 459–465, doi:10.1016/j.biocon.2005.04.018.

Chapelle, F. H. (2000), The significance of microbial processes in hydrogeology and geochemistry, *Hydrogeol. J.*, *8*(1), 41–46, doi:10.1007/PL00010973.

Christensen, T. H., P. Kjeldsen, P. L. Bjerg, D. L. Jensen, J. B. Christensen, A. Baun, H.-J. Albrechtsen, and G. Heron (2001), Biogeochemistry of landfill leachate plumes, *Appl. Geochem.*, *16*, 659–718, doi:10.1016/S0883-2927(00)00082-2.

Church, P., and G. Granato (1996), Bias in ground-water data caused by well-bore flow in long-screen wells, *Ground Water*, *34*(2), 262–273, doi:0.1111/j.1745-6584.1996.tb01886.x.

Clesceri, L. S., A. E. Greenberg, and A. D. Eaton (Eds.) (1998), *Standard Methods for the Examination of Water and Wastewater*, 20th ed., 1325 pp., Am. Public Health Assoc., Washington, D. C.

Cozzarelli, I. M., J. M. Suflita, G. A. Ulrich, S. H. Harris, M. A. Scholl, J. L. Schlottmann, and S. Christenson (2000), Geochemical and microbiological methods for evaluating anaerobic processes in an aquifer contaminated by landfill leachate, *Environ. Sci. Technol.*, *34*, 4025–4033, doi:10.1021/es991342b.

Davies, D., and D. W. Bouldin (1979), A cluster separation measure, *IEEE Trans. Pattern Anal. Mach. Intell.*, *1*(2), 224–227, doi:10.1109/TPAMI.1979.4766909.

Dollhopf, S., S. Hashsham, and J. Tiedje (2001), Interpreting 16S rDNA T-RFLP data: Application of self-organizing maps and principal component analysis to describe community dynamics and convergence, *Microb. Ecol.*, *42*, 495–505, doi:10.1007/s00248-001-0027-7.

Feris, K. P., P. W. Ramsey, S. M. Gibbons, C. Frazar, M. C. Rillig, J. N. Moore, J. E. Gannon, and W. E. Holben (2009), Hyporheic microbial community development is a sensitive indicator of metal contamination, *Environ. Sci. Technol.*, *43*, 6158–6163, doi:10.1021/es9005465.

Fields, M. W., C. E. Bagwell, S. L. Carroll, T. Yan, X. Liu, D. B. Watson, P. M. Jardine, C. S. Criddle, T. C. Hazen, and J. Zhou (2006), Phylogenetic and functional biomarkers as indicators of bacterial community responses to mixed-waste contamination, *Environ. Sci. Technol.*, *40*, 2601–2607, doi:10.1021/es051748q.

Gibbons, R. D. (1990), A general statistical procedure for groundwater detection monitoring at waste-disposal facilities, *Ground Water*, *28*(2), 235–243, doi:10.1111/j.1745-6584.1990.tb02251.x.

Giraudel, J., and S. Lek (2001), A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination, *Ecol. Modell.*, *146*(1–3), 329–339, doi:10.1016/S0304-3800(01)00324-6.

Griebler, C., and T. Lueders (2009), Microbial biodiversity in groundwater ecosystems, *Freshwater Biol.*, *54*, 649–677, doi:10.1111/j.1365-2427.2008.02013.x.

Holmes, D. E., K. T. Finneran, R. A. O'Neil, and D. R. Lovley (2002), Enrichment of members of the family *Geobacteraceae* associated with stimulation of dissimilatory metal reduction in uranium-contaminated aquifer sediments, *Appl. Environ. Microbiol.*, *68*(5), 2300–2306, doi:10.1128/AEM.68.5.2300-2306.2002.

Jain, A., M. Murty, and P. Flynn (1999), Data clustering: A review, *ACM Comput. Surv.*, *31*(3), 264–323, doi:10.1145/331499.331504.

Kjeldsen, P., M. A. Barlaz, A. P. Rooker, A. Baun, A. Ledin, and T. H. Christensen (2002), Present and long-term composition of MSW landfill leachate: A review, *Crit. Rev. Environ. Sci. Technol.*, *32*(4), 297–336, doi:10.1080/10643380290813462.

Kohonen, T. (1990), The self-organizing map, *Proc. IEEE*, *78*(9), 1464–1480, doi:10.1109/5.58325.

Lane, D. J., B. Pace, G. J. Olsen, D. Stahl, M. L. Sogin, and N. R. Pace (1985), Rapid determination of 16S ribosomal RNA sequences for phylogenetic analysis, *Proc. Natl. Acad. Sci. U. S. A.*, *82*, 6955–6959, doi:10.1073/pnas.82.20.6955.

Lerner, D., and G. Teutsch (1995), Recommendations for level-determined sampling in wells, *J. Hydrol.*, *171*(3–4), 355–377, doi:10.1016/0022-1694(95)06016-C.

Lovley, D. R. (2003), Cleaning up with genomics: Applying molecular biology to bioremediation, *Nat. Rev. Microbiol.*, *1*(1), 35–44, doi:10.1038/nrmicro731.

Ludvigsen, L., H.-J. Albrechtsen, D. B. Ringelberg, F. Ekelund, and T. H. Christensen (1999), Distribution and composition of microbial populations in a landfill leachate contaminated aquifer (Grindsted, Denmark), *Microb. Ecol.*, *37*, 197–207, doi:10.1007/s002489900143.

Madsen, E. L. (2000), Nucleic-acid characterization of the identity and activity of subsurface microorganisms, *Hydrogeol. J.*, *8*, 112–125, doi:10.1007/s100400050012.

Mangiameli, P., S. Chen, and D. West (1996), A comparison of SOM neural network and hierarchical clustering methods, *Eur. J. Oper. Res.*, *93*(2), 402–417, doi:10.1016/0377-2217(96)00038-0.

McArdle, B., and M. Anderson (2001), Fitting multivariate models to community data: A comment on distance-based redundancy analysis, *Ecology*, *82*(1), 290–297, doi:10.1890/0012-9658(2001).

Milligan, G., and M. Cooper (1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, *50*(2), 159–179, doi:10.1007/BF02294245.

Monti, S., P. Tamayo, J. Mesirov, and T. Golub (2003), Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data, *Mach. Learn.*, *52*, 91–118, doi:10.1023/A:1023949509487.

Mouser, P. J., D. M. Rizzo, W. F. M. Röling, and B. M. van Breukelen (2005), A multivariate statistical approach to spatial representation of groundwater contamination using hydrochemistry and microbial community profiles, *Environ. Sci. Technol.*, *39*, 7551–7559, doi:10.1021/es0502627.

Mouser, P. J., D. M. Rizzo, G. Druschel, S. E. Morales, N. Hayden, P. O'Grady, and L. Stevens (2010), Enhanced detection of groundwater contamination from a leaking waste disposal site by microbial community profiles, *Water Resour. Res.*, *46*, W12506, doi:10.1029/2010WR009459.

Mummey, D. L., and P. D. Stahl (2003), Spatial and temporal variability of bacterial 16S rDNA-based T-RFLP patterns derived from soil of two Wyoming grassland ecosystems, *FEMS Microbiol. Ecol.*, *46*, 113–120, doi:10.1016/S0168-6496(03)00208-3.

Ovreås, L., L. Forney, F. L. Daae, and V. Torsvik (1997), Distribution of bacterioplankton in meromictic Lake Saelenvannet, as determined by denaturing gradient gel electrophoresis of PCR-amplified gene fragments coding for 16S rRNA, *Appl. Environ. Microbiol.*, *63*(9), 3367–3373.

Park, Y., T. Chon, I. Kwak, and S. Lek (2004), Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks, *Sci. Total Environ.*, *327*(1–3), 105–122, doi:10.1016/j.scitotenv.2004.01.014.

Reyjol, Y., P. Fischer, S. Lek, R. Rösch, and R. Eckmann (2005), Studying the spatiotemporal variation of the littoral fish community in a large prealpine lake, using self-organizing mapping, *Can. J. Fish. Aquat. Sci.*, *62*(10), 2294–2302, doi:10.1139/F05-097.

Röling, W. F. M., B. M. Van Breukelen, M. Braster, B. Lin, and H. W. Van Verseveld (2001), Relationship between microbial community structure and hydrochemistry in a landfill leachate-polluted aquifer, *Appl. Environ. Microbiol.*, *67*(10), 4619–4629, doi:10.1128/AEM.67.10.4619-4629.2001.

Schryver, J., C. Brandt, S. Pfiffner, A. Palumbo, A. Peacock, D. White, J. McKinley, and P. Long (2006), Application of nonlinear analysis methods for identifying relationships between microbial community structure and groundwater geochemistry, *Microb. Ecol.*, *51*, 177–188, doi:10.1007/s00248-004-0137-0.

Snoeyenbos-West, O. L., K. P. Nevin, R. T. Anderson, and D. R. Lovley (2000), Enrichment of *Geobacter* species in response to stimulation of Fe(III) reduction in sandy aquifer sediments, *Microb. Ecol.*, *39*, 153–167, doi:10.1007/s002480000018.

Solidoro, C., V. Bandelj, P. Barbieri, G. Cossarini, and S. Umani (2007), Understanding dynamic of biogeochemical properties in the northern Adriatic Sea by using self-organizing maps and *k*-means clustering, *J. Geophys. Res.*, *112*, C07S90, doi:10.1029/2006JC003553.

Stein, H., C. Kellermann, S. I. Schmidt, H. Brielmann, C. Steube, S. E. Berkhoff, A. Fuchs, H. J. Hahn, B. Thulin, and C. Griebler (2010), The potential use of fauna and bacteria as ecological indicators for the assessment of groundwater quality, *J. Environ. Monit.*, *12*(1), 242–254, doi:10.1039/B913484K.

Tibshirani, R., G. Walther, and T. Hastie (2001), Estimating the number of clusters in a data set via the gap statistic, *J. R. Stat. Soc., Ser. B*, *63*(2), 411–423, doi:10.1111/1467-9868.00293.

U.S. Environmental Protection Agency (EPA) (2000), National water quality inventory: 1998 report to Congress, *Rep. EPA 841-R-00-001*, 434 pp., Off. of Water, Washington, D. C.

U.S. Environmental Protection Agency (EPA) (2009), Statistical analysis of groundwater monitoring data at RCRA facilities: Unified guidance, *Rep. EPA 530/R-09-007*, 888 pp., Off. of Resour. Conserv. and Recovery, Washington, D. C.

Vesanto, J., J. Himbert, E. Alhoneimi, and J. Parhankangas (2000), SOM Toolbox for Matlab 5, *Report A57*, Helsinki University of Technology, Helsinki, Finland.

Watanabe, K., K. Watanabe, Y. Kodama, K. Syutsubo, and S. Harayama (2000), Molecular characterization of bacterial populations in petroleum-contaminated groundwater discharged from underground crude oil storage cavities, *Appl. Environ. Microbiol.*, *66*(11), 4803–4809, doi:10.1128/AEM.66.11.4803-4809.2000.

Weiss, J. V., and I. M. Cozzarelli (2008), Biodegradation in contaminated aquifers: Incorporating microbial/molecular methods, *Ground Water*, *46*(2), 305–322, doi:10.1111/j.1745-6584.2007.00409.x.

Whitfield, C., S. Behura, S. Berlocher, A. Clark, J. Johnston, W. Sheppard, D. Smith, A. Suarez, D. Weaver, and N. Tsutsui (2006), Thrice out of Africa: Ancient and recent expansions of the honey bee, *Apis mellifera, Science*, *314*, 642–645, doi:10.1126/science.1132772.

P. J. Mouser, Department of Civil and Environmental Engineering and Geodetic Science, Ohio State University, 470 Hitchcock Hall, 2070 Neil Ave., Columbus, OH 43210, USA.

A. R. Pearce and D. M. Rizzo, School of Engineering, University of Vermont, 301 Votey Hall, 33 Colchester Ave., Burlington, VT 05405, USA. (arpearce@uvm.edu)