*Environ. Sci. Technol.* **2005**, *39,* 7551−7559

# A Multivariate Statistical Approach to Spatial Representation of Groundwater Contamination using Hydrochemistry and Microbial Community Profiles

PAULA J. MOUSER,*,† DONNA M. RIZZO,† WILFRED F. M. RÖLING,‡ AND BORIS M. VAN BREUKELEN§

*Department of Civil and Environmental Engineering, University of Vermont, Burlington, Vermont 05405, Department of Molecular Cell Physiology, and Department of Hydrology and Geo-Environmental Sciences, Faculty of Earth and Life Sciences, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands*

Managers of landfill sites are faced with enormous challenges when attempting to detect and delineate leachate plumes with a limited number of monitoring wells, assess spatial and temporal trends for hundreds of contaminants, and design long-term monitoring (LTM) strategies. Subsurface microbial ecology is a unique source of data that has been historically underutilized in LTM groundwater designs. This paper provides a methodology for utilizing qualitative and quantitative information (specifically, multiple water quality measurements and genome-based data) from a landfill leachate contaminated aquifer in Banisveld, The Netherlands, to improve the estimation of parameters of concern. We used a principal component analysis (PCA) to reduce nonindependent hydrochemistry data, *Bacteria* and *Archaea* community profiles from 16S rDNA denaturing gradient gel electrophoresis (DGGE), into six statistically independent variables, representing the majority of the original dataset variances. The PCA scores grouped samples based on the degree or class of contamination and were similar over considerable horizontal distances. Incorporation of the principal component scores with traditional subsurface information using cokriging improved the understanding of the contaminated area by reducing error variances and increasing detection efficiency. Combining these multiple types of data (e.g., genome-based information, hydrochemistry, borings) may be extremely useful at landfill or other LTM sites for designing cost-effective strategies to detect and monitor contaminants.

## Introduction

The cost of waste disposal does not end with the construction and filling of landfills but continues with the operation of long-term monitoring (LTM) and possible remediation systems that last 30 years or more after site closure. Landfills are the primary mechanism of municipal solid waste disposal in most developed nations, with the U.S. and Europe placing over 100 million tons into landfills each year (*1*, *2*). Since 1990, over 6000 disposal sites have been closed in the U.S. and Europe alike (*1*, *2*). Many of these historic landfills are unlined and poorly sited and, as a result, constitute the second largest pollutant source to groundwater in the U.S. (*3*). In an ideal LTM application, landfill owners would measure relevant in situ parameters at optimal locations and depths; early detection of contamination within site boundaries would result in significantly decreased risks, cleanup costs, and remediation time (*4*). Unfortunately, quantitative in situ detectors have not been developed that measure many chemical species, and the number of landfill detection wells are typically limited, reducing detection efficiency.

One improvement to LTM strategies is to utilize multiple types of data (e.g., water chemistry or soil structure) to increase knowledge and reduce uncertainty associated with estimating relevant parameters of concern (e.g., plume concentrations, hydraulic conductivities). Subsurface microbial ecology is a unique source of data that has been underutilized in groundwater LTM schemes. Organism distribution, type, and abundance can provide insight into plume geochemical evolution, drivers of oxidation and reduction reactions, and the potential for attenuation (see refs *5*−*8*). Monitoring microbial communities at an ecological scale provides valuable information that may not be detected using traditional hydrochemical analysis, resulting in lower human health risks and/or decreased LTM costs (*9*). For example, Maymo-Gatell et al. (*10*) and Seshadri et al. (*11*) have isolated and sequenced the bacterium *Dehalococcoides ethenogenes,* responsible for dechlorinating tetrachloroethene (PCE) to the nontoxic ethene. Monitoring the distribution and abundance of this bacterium when PCE concentrations are reduced to slightly below analytical detection limits may increase monitoring efficiency.

Of the two broad categories of microbial community profiling, nonmolecular and molecular, the genome- or molecular-based methods, such as 16S rDNA/rRNA polymerase chain reaction (PCR) (see refs *12*−*14* for further explanation), are being incorporated into the analysis and solution of environmental engineering problems. These tools provide exciting opportunities for obtaining microbiological information at multiple scales that range from identifying individual taxa responsible for specific chemical reactions (*11*, *15*) to the development of broad community profiles involved in major geochemical processes occurring within or outside a contaminated groundwater plume (*16*). For the specific purpose of improving our understanding of the physical processes and the corresponding complexities associated with subsurface contamination, however, genome-based information must (i) represent the organisms that are responding to changes in their living environment, (ii) exhibit spatial structure/correlation across regions or scales of interest, and (iii) quantitatively and/or qualitatively describe the community of interest in ways that advance geostatistical estimation of parameters of interest. Associations between microorganisms and their surrounding subsurface environment have been shown previously (*16*−*18*) as have the spatial correlation of microorganisms at several experimental scales (*19*−*23*). Yet, the incorporation of this valuable microbial-based information into existing geostatistical estimation techniques for the purpose of improving our understanding of subsurface properties has not been well-researched.

* Corresponding author phone (802)656-1937; fax(802) 656-8448; e-mail: Paula.Mouser@uvm.edu.
† University of Vermont.
‡ Department of Molecular Cell Physiology, Vrije Universiteit.
§ Department of Hydrology and Geo-Environmental Sciences, Vrije Universiteit.

10.1021/es0502627 CCC: $30.25 
Published on Web 08/23/2005

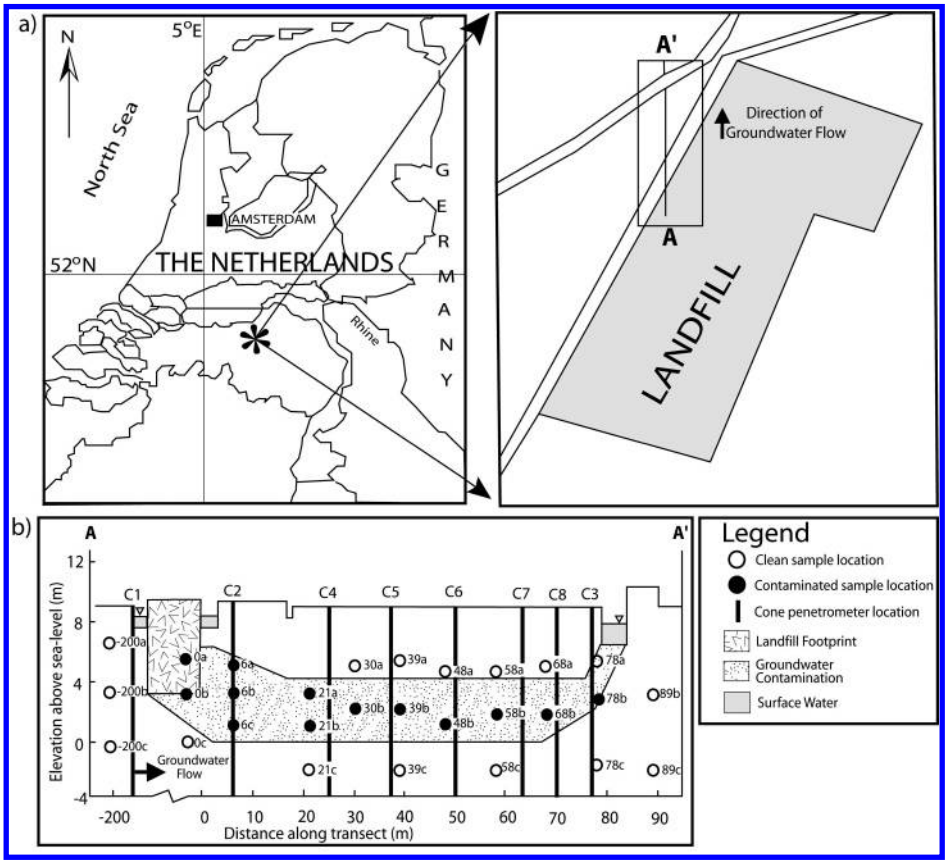


**FIGURE 1. Location of the Banisveld landfill. (a) Location of the transect in the direction of groundwater flow and (b) cross-section of transect with sample classification (clean or contaminated). Circles represent monitoring locations and vertical bar represent boring locations.**

To utilize microbial profiles in geostatistical methods for estimating subsurface concentrations, these data must first be converted from a categorical response such as presence/absence, intensity, density, phylogenic tree, or similarity matrix to an appropriate measurement or classification. We present a method for reducing dependent hydrochemical and microbial data (16S rDNA denaturing gradient gel electrophoresis (DGGE) profiles of *Bacteria* and *Archaea*) collected from a landfill leachate-contaminated aquifer into independent variables using principal component analysis (PCA). The principal components exhibit spatial structure, are correlated to other subsurface parameters along a cross-section of interest, and are incorporated as a traditional type of LTM data for geostatistical estimation and site management. PCA has been used alone for the analysis of microbial data (*18*, *24*−*26*). It has also been used in tandem with geostatistical methods for classification and prediction of environmental data (*21*, *22*, *27*−*29*). The objective of this paper is to go one step further and show the potential for using microbial information: (i) as a sentry or early warning signal when combined with hydrochemical information and (ii) as secondary data for the purpose of improving subsurface parameter estimates and reducing uncertainty associated with subsurface site characterization.

## Materials and Methods

**Site Background and Available Data.** The Banisveld landfill, a 6 ha unlined waste disposal site near Boxtel, The Netherlands (Figure 1a) operated in a 6 m deep sand pit underlain by an 11 m groundwater aquifer between 1965 and 1977. Although the majority of waste was characterized as municipal or household garbage, aromatic hydrocarbons such as benzene, ethylbenzene, xylenes, and naphthalene were detected at several monitoring locations (*30*). In 1998, contaminated groundwater was delineated, and monitoring well samples were found to have a slightly acidic pH, high concentrations of alkalinity and dissolved organic carbon (DOC), a variety of ions, and petroleum byproducts at concentrations above drinking water limits (*30*).

Van Breukelen (*30*) characterized the horizontal extent of the Banisveld landfill leachate plume using an extensive electromagnetic survey, EM-34, with 10 m horizontal and vertical intercoil loop spacing. The vertical extent of the leachate plume was delineated across transect A−A′ using electrical formation conductivity measurements from eight detailed cone penetrometer borings (C1−C8) in the direction of groundwater flow (Figure 1b) (*30*). Groundwater is estimated to be flowing at approximately 4 m/year in a northerly direction (*30*). The horizontal extent of the plume was identified at distances up to 80 m from the landfill footprint, while the vertical extent was estimated between 4 and 9 m below the ground surface (Figure 1b). The electrical conductivity of the subsurface (soils and groundwater) has been used elsewhere to map in situ leachate contamination (*32*, *33*) because landfill leachate typically exhibits high conductivity (low electrical resistivity) (*34*). Observation wells were installed at 11 locations (−200, 0, 6, etc.), and sampling points were placed at multiple depths (a, b, c) within the observation wells for a total of 29 monitoring locations across transect A−A′ (Figure 1b). Monitoring locations were characterized as clean or contaminated by Van Breukelen (*30*).

The groundwater monitoring data for this case study were collected in September 1998 and used previously to relate water quality and the microbial ecology of the leachate plume (*17*). Hydrochemical data consisted of 24 variables (alkalinity, pH, electrical conductivity (EC), Cl, Na, K, Ca, Mg, $NH_4$, $NO_2$, $NO_3$, Mn(II), Fe(II), $SO_4$, $H_2S$, Si, Al, $H_2$, benzene, toluene,

ethylbenzene, xylene, naphthalene, and dissolved organic carbon (DOC)) and 16S rDNA microbial communities (PCR-DGGE profiles) of *Bacteria* and *Archaea* taken from 29 locations within, upgradient, and downgradient of the landfill (Figure 1b). Specific details of sampling procedures and analysis may be found elsewhere (*17*, *30*, *31*, *35*), but briefly, methods used to produce PCR-based fingerprints are described as follows. Groundwater samples were taken from their natural environment using a peristaltic pump; samples were pelletized by vacuum filtering and centrifuging; DNA was isolated using a purification kit; the 16S rDNA gene was amplified to a measurable concentration using PCR; and the amplified DNA was visualized using DGGE. In DGGE, the amplified DNA was subjected to a denaturing gradient that spread DNA across a gel based on its nucleotide sequence. The *Bacteria* and *Archaea* DGGE images were processed using Gelcompar software (v. 4.0, Applied Maths) in a band-independent manner by dissecting the profile into 400 equally spaced intervals. The pixel intensities for the intervals were used to calculate the Pearson product moment correlation coefficient (*17*). This method was not dependent upon manual band assignments and is less sensitive to variations in the amount of PCR product (*36*). By targeting specific groups of organisms, in this case *Bacteria* and *Archaea*, the PCR amplification process is limited to organisms that fall only within those domains. A discussion of the benefits and limitations of PCR-based rDNA/rRNA methods is beyond the scope of this paper but can be found in detail elsewhere (*12*−*14*).

**Multivariate Data Analysis.** Principal component analysis (PCA) is a multivariate statistical technique used to (i) transform a set of interrelated variables into statistically independent variables (termed eigenvectors or principal components) and (ii) gain insight into the relationships between variables. When variables are correlated, PCA is useful in reducing the data to a smaller number of orthogonal linear combinations, and a large proportion of the dataset variance may be accounted for in a few principal components (*37*). However, if variables are uncorrelated, approximately the same number of principal components as variables will be needed to describe the dataset variation. We used PCA to detect the structure and produce statistically independent qualitative variables for three types of data from the Banisveld landfill: hydrochemistry and 16S rDNA DGGE profiles of *Bacteria* and *Archaea*. The PCA of these three types of data allow us to combine a large amount of information into a reduced number of principal components without losing much of the information and provide insight into what may be responsible for the similarities between the groundwater samples.

For this analysis, a separate PCA was performed on each type of data (hydrochemistry, *Bacteria*, and *Archaea*). When hydrochemical variables did not meet normality requirements, we applied a logarithmic transformation ($\log(1 + x)$) to improve the distribution before statistical analysis. Correlation coefficients from processed DGGE gel images of *Bacteria* and *Archaea* were used in the microbial principal component analyses. These multidimensional variables, or principal components, may then be used in spatial mapping to estimate hydrochemical or microbial community principal component scores at unsampled locations for characterizing the zone of contamination and to combine multivariate descriptions of hydrochemistry, *Bacteria*, and *Archaea* with other types of subsurface data to model joint spatial distribution and confidence (defined as error variances) at unknown locations.

**Spatial Data Analysis.** Geostatistical methods for describing and interpolating spatially correlated data take advantage of the common observation that, on average, values closer together in space will be more similar than those further from each other. The steps in applying these methods include developing analytical models that describe the spatial variation between pairs of spatially or temporally related samples and then using these models to estimate sample parameters and their error variances at unknown locations. Although originally used in the geological sciences (*38*), geostatistics has also been frequently applied in agricultural and ecological sciences (i) to evaluate spatial dependence of subsurface properties and/or ecological communities or (ii) for interpolation of these parameters (*25*, *26*, *29*, *39*−*45*).

To gain an understanding of the spatial structure, experimental data were binned into lag distances with approximately the same number of data; and semivariogram values were calculated for each bin (denoted by individual points shown in Figure 2) using MATLAB version 6.1 (MathWorks, Inc, Natick, MA). We selected a given form (e.g., spherical, Gaussian) for the analytical model and fit the model parameters (i.e., range, sill, nugget) to the experimental variogram using nonlinear model fitting functions in JMP statistical software (SAS Institute Inc., Cary, NC) (see refs *41* and *46* for details on calculating semivariance, common variogram models, and model fitting). At large separation distances between pairs of data points ($h$ of Figure 2a), the analytical models oscillate around a plateau known as the sill. The separation distance beyond which no spatial correlation exists and estimates of covariance values remain essentially constant is called the range. If data are discontinuous near the origin due measurement error or spatial correlation occurring at distances smaller than the sample interval, semivariograms sometimes exhibit a nugget effect and jump from the origin to some $y$-intercept or initial error variance (*45*).

Experimental directional semivariograms with fitted models were developed for monitoring locations [samples located at −200 m upgradient of the landfill were not used in the kriging and cokriging estimates because of their distance from the transect of interest, and PCR results for *Archaea* were negative in 7 monitoring locations] leaving a total of 26 hydrochemistry samples, 26 *Bacteria* profiles, and 19 *Archaea* profiles for estimating the spatial structure associated with the first two principal components (Figure 2a−c). Semivariograms were also fit to electrical conductivity measurements in the vertical direction and over the transect A−A′ to estimate the spatial structure in the vertical and horizontal directions, respectively.

Cross-semivariograms provide information about the spatial distance over which pairs of variables are related. Oftentimes, cross-semivariograms are developed between two parameters where one (termed hard or primary data) is the more quantitative, difficult, and/or costly in terms of time or expense to obtain, while the other (termed soft or secondary data) is often more qualitative, less expensive, or easier to measure on-site. At the Banisveld site, electrical formation conductivity measurements collected from eight cone penetrometer borings, spaced approximately every 10 m along the horizontal transect A−A′ and approximately every 0.03 m in the vertical direction, were treated as primary data. Cross-semivariograms were fit between electrical formation conductivity measurements (primary data) and the first PC (PC1) for hydrochemistry, *Bacteria*, and *Archaea* (secondary data) using MATLAB version 6.1 (MathWorks, Inc, Natick, MA).

The kriging methods use a linear regression technique to estimate a parameter and the uncertainty or error variance associated with the parameter at unsampled locations (*46*, *47*), thereby producing an estimate that minimizes the error between surrounding know values and unsampled location in an unbiased way. In ordinary kriging, the estimate at some

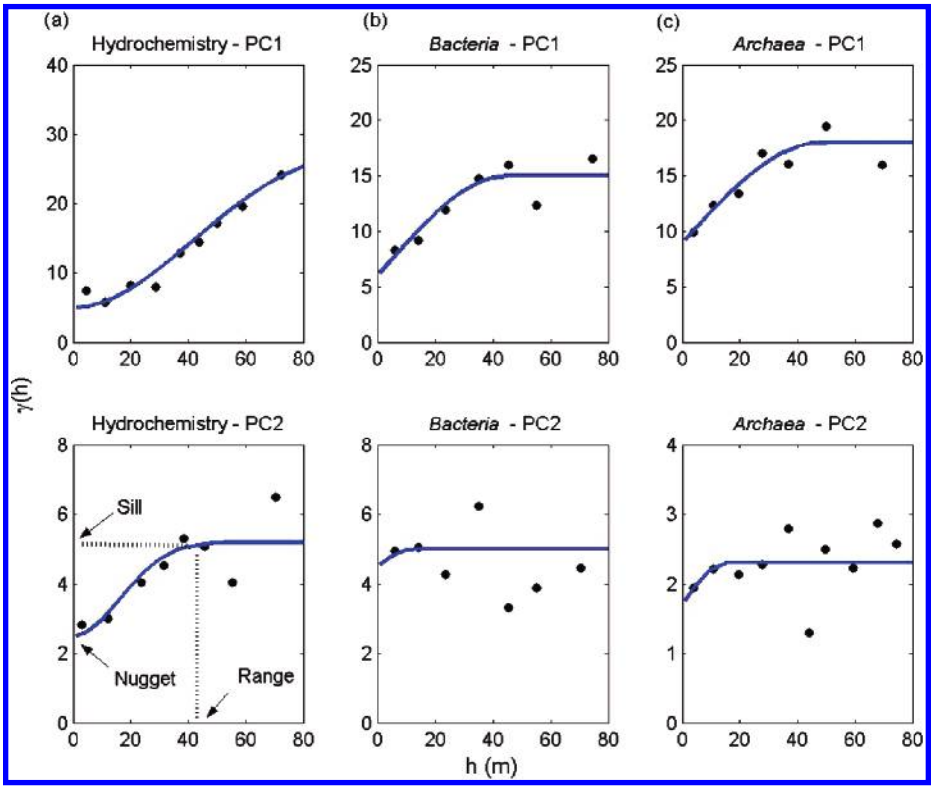


**FIGURE 2. Experimental semivariograms and fitted models for (a) hydrochemistry, (b) *Bacteria*, and (c) *Archaea*. The first principal components (PC1s) are on the top row, and the second principal components (PC2s) are on the bottom row.**

arbitrary point based on $n$ measured (surrounding) values is given as

$$\hat{v} = \sum_{i=1}^{n} w_i v_i \tag{1}$$

where $v_i$ are the surrounding measured (observed) values. The weights $w_i$ are selected such that the estimate is unbiased ($E[\hat{v} - v] = 0$) and has a minimum variance ($E[\hat{v} - v]^2$ is minimum). Spatial interpolation is optimized by encapsulating the spatial correlation of samples in the variogram functions. The variogram functions provide a means for computing a vector of kriging weights ($w_i$ of eq 1)

$$\mathbf{w} = C^{-1}\mathbf{d} \tag{2}$$

where $C$ is a covariance matrix developed using the spatial structure (semivariogram/cross-semivariogram) and the distance between measurements, and **d** is a vector involving the variogram and the distance between sample locations and the (unknown) estimation point. To ensure an unbiased estimator, we imposed the condition that the weights $w_i$ sum to 1 (see ref *48* for details). The variance of the estimation error for unsampled locations may be calculated as

$$\sigma_r^2 = \sigma^2 - \left(\sum_{i=1}^{n} w_i d_i + \lambda\right) \tag{3}$$

variance at the unsampled location, $\sigma^2$ is the sample variance, and $\lambda$ is the Lagrange parameter.

The linear system of ordinary kriging equations and associated error variance (eqs 2 and 3) was developed, solved, and plotted in MATLAB version 6.1 (MathWorks, Inc, Natick, MA) for each point over the vertical transect A−A′ at which an estimate was required. (For more detail on their derivation and assumptions, see ref *38*). Ordinary cokriging was performed with WinGslib version 1.03 (Statios LLC, Stanford, CA) to produce estimates and error variances between primary and secondary variables. Note that the quality of the kriging weights (and therefore, of the interpolation and estimate of the variance of error) depends on the model selected to represent the variogram.

## Results and Discussion

**Principal Component Analysis.** The percent of variance explained in the first two principal components (PC1 and PC2) was quite high for each of the three data types (hydrochemistry: PC1 = 57% and PC2 = 14%; *Bacteria*: PC1 = 48% and PC2 = 20%; *Archaea*: PC1 =68% and PC2 = 18%). This implies that the majority of the information (68−86%) may now be represented with two new variables, PC1 and PC2, that are uncorrelated, linear combinations of the original variables. Figure 3 illustrates how the reduced principal components may be used to separate samples into two classes, clean or contaminated. For hydrochemistry and *Bacteria*, the higher the PC1 score, the more contaminated a sample location. The contaminated *Archaea* samples grouped positively, while the clean samples grouped negatively along the PC1 component axis. *Archaea* had complete separation between the type of sample (clean vs contaminated) within PC1, while hydrochemistry had separation with both PC1 and PC2; see Figure 3, panels c and a, respectively. The clean and contaminated *Bacteria* samples were not fully separated within the first two principal components; however, there is a clustering of clean and contaminated samples along these components (Figure 3b). The grouping and/or separation of clean and contaminated samples is extremely useful from a management or regulatory standpoint. Rather than focusing on a particular constituent as an indicator of pollution, which is often misleading due to the sheer number of potential contaminants at landfill sites, sample parameters may be combined to form new variables that are definitive indicators of pollution for any location along the transect. This type of comparison requires that samples be taken from

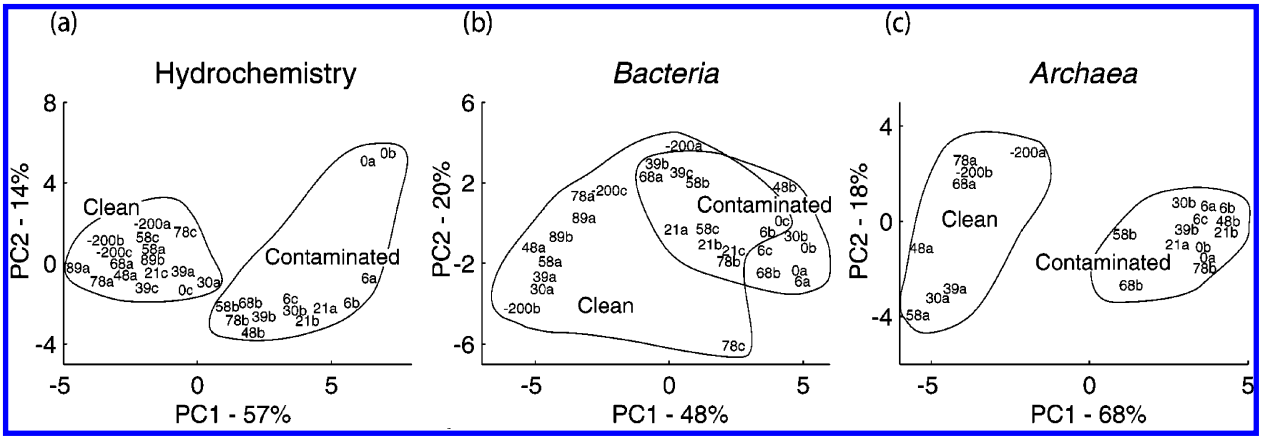


**FIGURE 3. The variance contributed by the first and second principal components (PC1 and PC2) are plotted along the horizontal and vertical axes, respectively.**

both clean (preferable up- and downgradient) and contaminated monitoring locations.

The hydrochemistry principal components give insight into what constituents are driving the separation of clean and contaminated samples, similar to a gradient or isocontours. Hydrochemistry PC1 scores correlate positively to toluene, DOC, alkalinity, EC, $NH_4$, Ca, K, Cl, Na, and Mg, while PC2 scores correlate positively to ethylene, xylene, naphthalene, and Si and negatively to Fe(II) and Mn(II). The second PC primarily separates contaminated samples taken within the landfill (0a and 0b) from contaminated samples in other portions of the groundwater plume.

The principal component scores for the microbial community profiles offer landfill managers additional insight to organisms that have adapted to surrounding environmental stressors. For example, the first principal component for *Bacteria* correlates positively to microbial communities common between contaminated samples and correlates negatively to communities common between clean samples, with some overlap between the two groups. For PCA scores where clean samples cannot be distinguished from contaminated samples (Figure 3b), it is interesting to note that 4 of 11 monitoring locations are below the characterized plume, in particular, along the fringe or boundary of the highly contaminated areas (see 0c, 21c, 39c, and 58c of Figure 1b). The second principal component for *Bacteria* represents uniqueness within the community structure or differences among organisms within each class.

The first principal component for *Archaea* separates samples into clean and contaminated groups, with positive correlation for contaminated samples and negative correlation for clean samples based on their communities (Figure 3c). The clarity of this separation is not surprising as *Archaea* are a taxonomic group that typically flourish in more extreme environments; therefore, we might expect only samples within the landfill or those that have been in contact with leachate to contain a diverse *Archaea* community. The second principal component of *Archaea* indicates slight structural differences among classes. As was the case with *Bacteria*, the grouping of *Archaea* samples located within the landfill (0a and 0b) with other contaminated samples suggests a common core community for all contaminated locations, regardless of their position within the plume.

PCA is often used to determine which variables are important for explaining dataset variance and which variables could potentially be dropped from the sampling scheme, resulting in a decrease in LTM costs without a significant increase in unexplained variance. While this example does not highlight this attribute of a PCA, as more than 67% (or 16 of 24) of the hydrochemical parameters were highly correlated to the first two principal components, it may be extremely useful at landfill sites to address which of the typically 100+ sampled constituents are not contributing to knowledge of contamination. In addition, there is likely a minimum set of microbial communities represented here as DGGE bands that are contributing to the PCA separations that may be addressed with additional statistics and/or image processing algorithims. For example, a discriminant analysis conducted with hydrochemistry and most probable numbers based on Eco-BioLog substrate utilization (*35*) (Biolog Inc, Hayward, CA) taken at a later date from the sample locations revealed that three hydrochemical compounds (Fe, Mg, and $CH_4$) and three carbon substrates indicative of microbial utilization (Glycogen, Glycyl-L-glutamic acid, and D-malic acid) separated clean and contaminated sample locations 100% of the time (data not shown).

**Geostatistical Analysis.** Semivariograms developed for the first principal component for hydrochemistry, *Bacteria*, and *Archaea* had relatively good model fits ($R^2$ between 0.804 and 0.947) and were significant at the $\alpha = 0.01$ level (Table 1). All three show some discontinuity at the origin and horizontal correlation distances between 40 and 50 m (see Figure 2 and Table 1). The second principal components had noticeable discontinuities at the origin resulting in large nugget effects, and small horizontal spatial correlation (Figure 2 and Table 1). The PC2s were not used in the geostatistical analyses because they explained only a small portion of the variance as compared with PC1 (14−20%, see Figure 3) and had relatively poor model fits ($R^2$ between 0.026 and 0.709, prob > $F$ between 0.0087 and 0.7313; see Table 1). Variograms developed in the vertical direction for individual electrical formation conductivity borings (C2−C8) showed highly significant fits ($R^2 > 0.97$, $p < 0.0001$), while the variogram developed across the transect A−A′ for all electrical formation conductivity borings showed a moderate fit ($R^2 = 0.575$) due to the natural soil heterogeneities (see Table 1). Cross-semivariograms developed between electrical formation conductivity (primary data) and the first principal components of hydrochemistry, *Bacteria*, and *Archaea* (secondary data) showed spatial correlations between 20 and 30 m and moderate model fits ($R^2$ between 0.536 and 0.641) (see Table 1). Combining electrical formation conductivity (data range of 0−100 mS/m) and principal component scores (data range of −6 to 6) required data transformation and standardization.

It is rare that municipal landfills have a sufficient number of detection or LTM wells for estimating primary data for geostatistical applications. Landfills usually have less than 10 LTM wells located across large areas (> 10 ha). A suggested rule for variogram development is at least 30 sample pairs per experimental semivariogram point or greater than 25 monitoring locations (*38*). The number of data for the Banisveld site [19−26 monitoring locations and eight CPT

**TABLE 1. Semivariograms and Cross-Semivariograms for Hydrochemistry, *Bacteria* and *Archaea* PCs, and Electrical Formation Conductivity[a]**

| data type | model type | nugget | sill | range (m) | nugget effect (nugget/sill) | $R^2$ | prob > $F$ |
|---|---|---|---|---|---|---|---|
| | | **Semivariograms** | | | | | |
| hydrochemistry PC1 | Gaussian | 5 | 22 | 50 | 0.23 | 0.947 | <0.0001 |
| *Bacteria* PC1 | spherical | 6 | 15 | 40 | 0.40 | 0.804 | 0.0062 |
| *Archaea* PC1 | spherical | 9 | 18 | 40 | 0.50 | 0.844 | 0.0035 |
| hydrochemistry PC2 | Gaussian | 2.5 | 5 | 30 | 0.50 | 0.709 | 0.0087 |
| *Bacteria* PC2 | spherical | 4.5 | 5 | 10 | 0.90 | 0.026 | 0.7313 |
| *Archaea* PC2 | spherical | 1.7 | 2.3 | 10 | 0.74 | 0.068 | 0.4657 |
| formation conductivity | spherical | 170 | 950 | 50 | 0.18 | 0.575 | 0.0027 |
| C2 formation conductivity | spherical | 0 | 116 0 | 9 | 0 | 0.998 | <0.0001 |
| C3 formation conductivity | spherical | 0 | 148 | 5 | 0 | 0.971 | <0.0001 |
| C4 formation conductivity | Gaussian | 0 | 725 | 3 | 0 | 0.986 | <0.0001 |
| C5 formation conductivity | spherical | 0 | 424 | 6.3 | 0 | 0.988 | <0.0001 |
| C6 formation conductivity | spherical | 0 | 422 | 6 | 0 | 0.993 | <0.0001 |
| C7 formation conductivity | spherical | 0 | 258 | 4.3 | 0 | 0.995 | <0.0001 |
| C8 formation conductivity | spherical | 0 | 644 | 5.7 | 0 | 0.982 | <0.0001 |
| | | **Cross Semivariograms[b]** | | | | | |
| formation conductivity and HChem PC1 | Gaussian | 2 | 28 | 30 | 0.07 | 0.636 | 0.0100 |
| formation conductivity and *Bacteria* PC1 | Gaussian | 3 | 28 | 20 | 0.11 | 0.536 | 0.0247 |
| formation conductivity and *Archaea* PC1 | Gaussian | 3 | 28 | 20 | 0.11 | 0.641 | 0.0054 |

[a] Columns 1−6 indicate data type, model type, nugget, sill, range, and nugget effect, respectively. The model fit and significance are indicated by the regression coefficient, $R^2$, and the prob > $F$, respectively. [b] Denotes data that were standardized.

borings with sampling at 0.03 m intervals in the vertical direction for transect A−A′] is unique and would not typically be found at historic landfill sites unless research or extensive remediation had taken place. We must point out that there is very likely small-scale microbiological correlation occurring at intervals less than the distances between these horizontal monitoring locations (<10 m) that cannot be detected with this data set. While it may be difficult to obtain reliable estimates from small sample sets, geostatistics can be applied to undersized data sets provided caution is used during development and interpretation. Other sources of data (e.g., grain size) or knowledge of similar sites can be used to supplement sample information and can be especially useful in identifying soil layering or large changes in horizontal and vertical (anisotropic) conditions that are otherwise not visible. Because of the near continuous vertical sampling density at this site, semivariograms developed in the vertical direction for electrical formation conductivity indicate vertical spatial correlations between 5 and 9 m and horizontal spatial correlations of 50 m, suggesting that parameters associated with the electrical formation conductivity may exhibit a 5:1 or greater horizontal−vertical spatial scaling effect.

Electrical formation conductivity measurements from cone penetrometer borings were used as our ground truth and were plotted for transect A−A′ with red colors indicating contaminated areas and dark blue colors representing clean locations (Figure 4a). Given the spatial relationships defined by semivariograms, ordinary kriging was used to produce estimates of the first principal component for hydrochemistry, *Bacteria*, and *Archaea* along the transect A−A′ (Figure 4b−d). The estimates of PC1 for hydrochemistry and *Archaea* show good contrast between clean and contaminated zones (Figure 4b,d), with remarkable similarity to ground truth measurements, particularly in the vertical direction. Estimates of PC1 for *Archaea* depict a similar contaminated zone as estimates of PC1 for hydrochemistry, although only 19 samples (vs 26 for hydrochemistry) were used (Figure 4d). We would expect a clear distinction between clean and contaminated zones due to the excellent separation of *Archaea* along PC1 (Figure 3). This is an indication that the core community of *Archaea*, originating within the landfill or from leachate-contaminated groundwater, will be present for any groundwater that has been in contact with the leachate.

In contrast, the estimates of PC1 for *Bacteria* appear to be more useful for describing the size of the fringe effect, with a wider and longer zone of higher PC1 for samples outside the characterized plume (Figure 4c). The PCA has identified similarities of the DGGE *Bacteria* profiles that may be indicative of organisms adapted to biogeochemical processes occurring at the lower (vertical) plume fringes that are not detected with other methods (e.g., electrical formation conductivity, hydrochemistry PC1). We believe that the wider plume reflects fringe effects, a result that could mean increased efficiency and sensitivity for detecting and delineating edges of plumes. However, an alternative explanation is that these effects are solely a result of increased variability in *Bacteria* profiles in general. Additional research for this application may indicate that as a broad ecological measure, *Bacteria* communities respond quickly to changes in groundwater quality along the base and front edges of the plume.

The electrical formation conductivity data (primary data) and the first principal components of hydrochemistry, *Bacteria*, and *Archaea*, respectively (three separate sources of secondary data), were combined in a cokriging model to compare the parameter estimates and error variances produced with ordinary kriging of electrical formation conductivity alone (Figure 5a−d). The use of highly resolved electrical conductivity measurements (on the order of 3 cm spacing) provide more fine-scaled parameter estimates than the screened monitoring locations along the vertical extent of transect A−A′. This is particularly evident when comparing concentration estimates in the 55−85 m range of Figures 4 and 5. In practice, it is common to characterize the landfill leachate using a select suite of contaminants of concern (i.e., by creating a series of individual concentration plumes). However, principal component scores (Figure 4b−d) are a combination of many variables (e.g., 24 hydrochemical parameters) representing multiple attributes of the leachate contamination. The addition of the secondary data at 21 m downgradient provided lower estimates of electrical formation conductivity in all three maps (Figure 5b−d), causing a shorter (horizontal) and slightly wider (vertical) source plume between 10 and 20 m.

By combining primary and secondary data in a cokriging model, error variances were reduced by as much as 25% using the first PCs, which account for 48−68% of the variance in the data (Figure 6a−c). Because the hydrochemistry data are similar in type to electrical formation conductivity, we

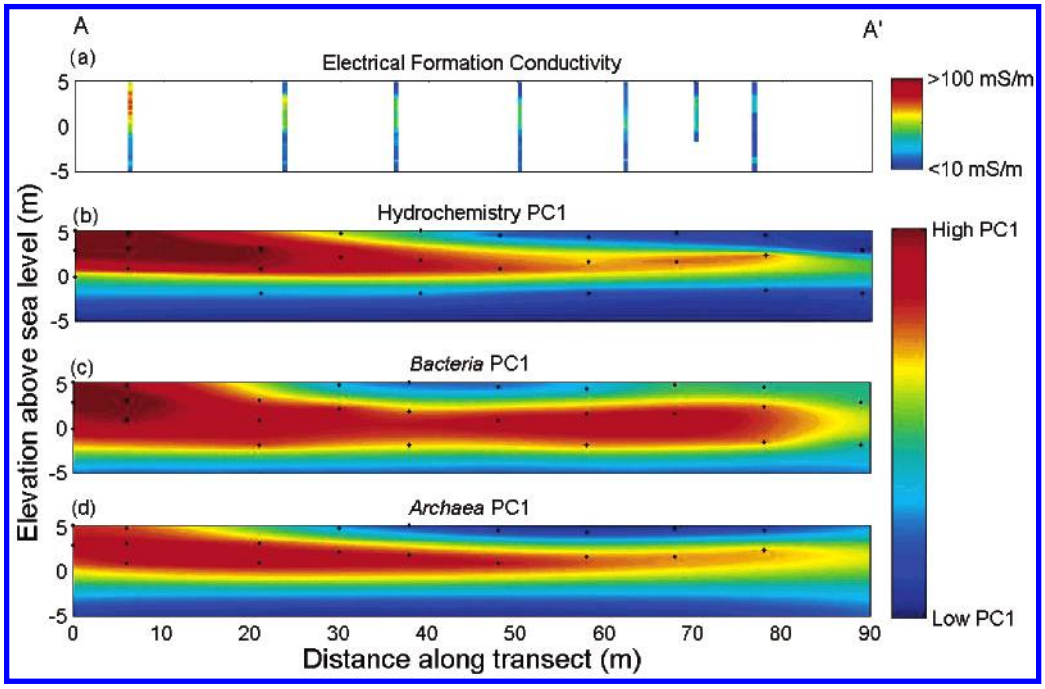


**FIGURE 4. (a) Electrical formation conductivity measurements for soil borings at the Banisveld landfill. Warm colors (e.g., green, red (>50 mS/m)) correspond to leachate contaminated groundwater, while cool colors (e.g., blue (<10 mS/m)) designate clean areas. Ordinary kriging estimates for the first principal components of (b) hydrochemistry, (c) *Bacteria*, and (d) *Archaea*. Black circles represent monitoring locations.**

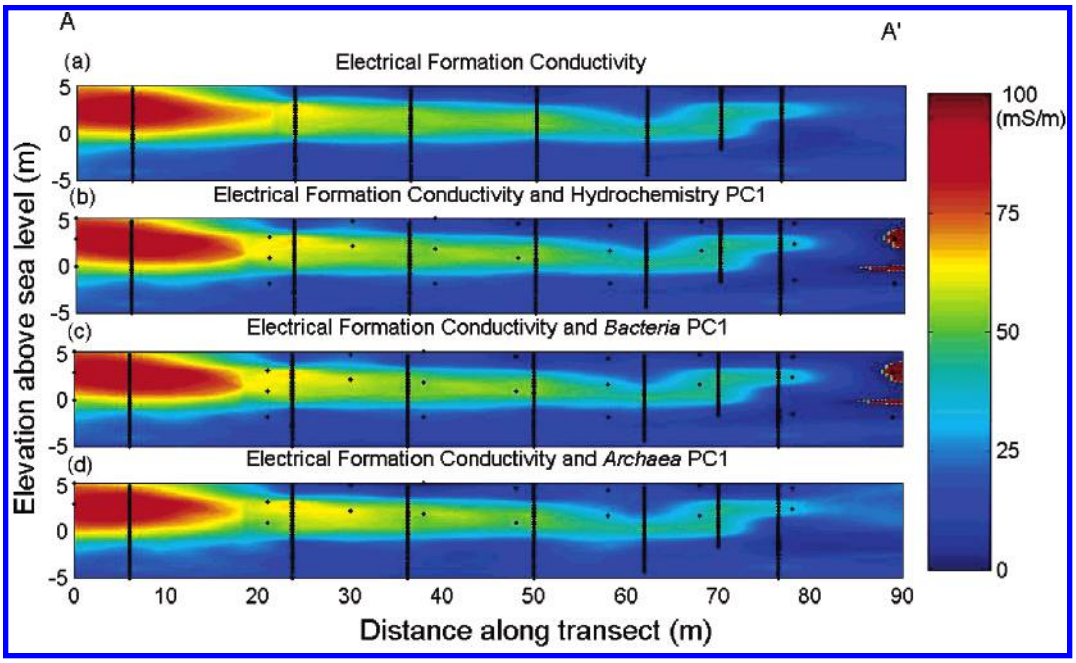


**FIGURE 5. Comparison of (a) ordinary kriging estimates of electrical formation conductivity to cokriging estimates of electrical formation conductivity using the first principal component of (b) hydrochemistry, (c) *Bacteria*, and (d) *Archaea*. Black circles represent monitoring locations, and black lines represent boring locations.**

would expect the addition of the hydrochemistry PC1 in a cokriging model to be the most useful in reducing uncertainty (Figure 6a). It is exciting to note, however, that the microbial data (*Archaea* and *Bacteria* PC1) lower error variances up to 10 and 15%, respectively, at many locations across the transect (Figure 6c,d). For this study, the nature of the DGGE data did not lend itself to a collective PCA analysis. In future studies, however, it may be possible to combine multivariate data types (hydrochemistry, *Bacteria*, and *Archaea* profiles) in a collective PCA analysis with the expectation that organisms might be grouped with their associated hydrochemical parameters (i.e., PC1 might represent microorganisms responsible for the degradation of several monoaromatic compounds and the associated constituents). This could lead to the integration of molecular fingerprinting information with geostatistical analyses for predicting contamination along regions of concern (i.e., boundaries, plume edges), assessing remediation strategies, or developing a more cost-effective method of monitoring.

In this case study, we combined multivariate data of groundwater hydrochemistry, *Bacteria*, and *Archaea* community profiles from the Banisveld landfill (The Netherlands) into principal component variables. The PCA of hydrochemistry, *Bacteria*, and *Archaea* demonstrates that combining

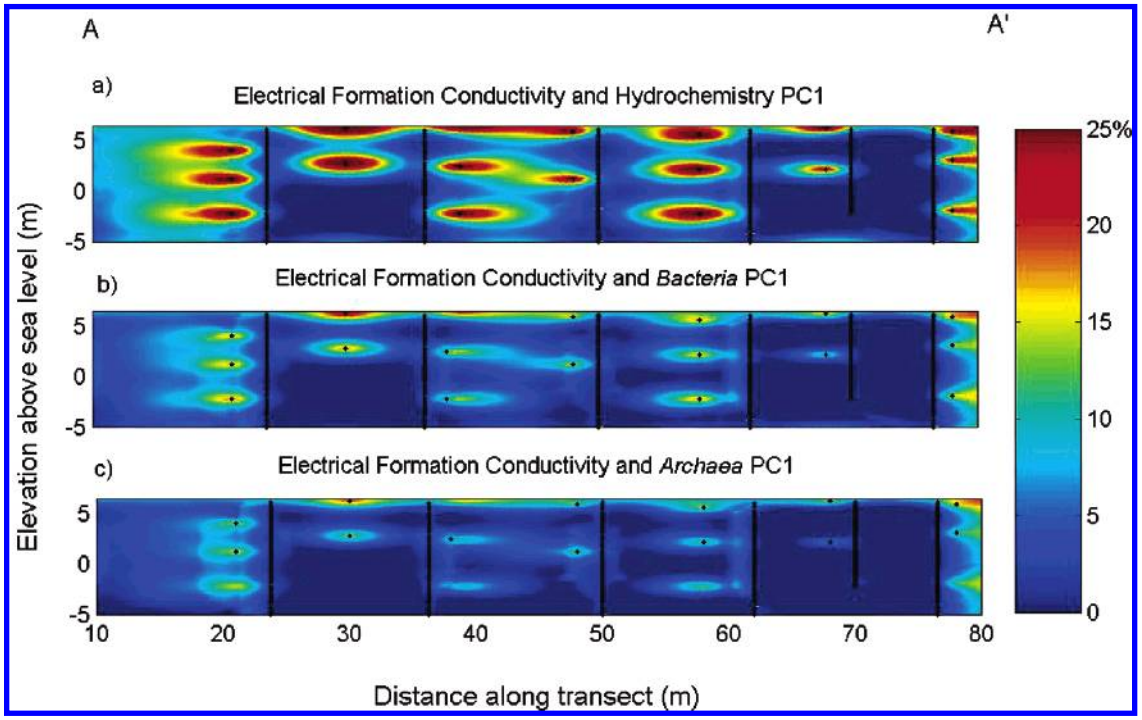


**FIGURE 6. Percent difference between error variance of electrical formation conductivity produced using ordinary kriging and electrical formation conductivity produced using ordinary cokriging with the first principal component of (a) hydrochemistry, (b) *Bacteria*, and (c) *Archaea*. Black circles represent monitoring locations, and black lines represent boring locations.**

multiple chemical measurements and/or microbial profiles into a reduced set of principal component scores can effectively describe a relative degree or class of contamination. The first principal components for both hydrochemistry and *Bacteria* appear to be indicators of a pollution gradient, with contaminated samples having high PC1 scores and clean samples having low PC1 scores, whereas the first principal component for *Archaea* is useful for separating and classifying clean and contaminated sample locations. *Bacteria* PC1 describes microbial structures that may be related to the fringe or plume boundary effects. This PCA analysis offers the managers of LTM sites several benefits since leachate-impacted groundwater can consist of hundreds of organic and inorganic contaminants, individually depicting a separate concentration plume changing with space and time. By describing the water quality in general terms, managers have the ability to (i) combine large datasets into variables that represent the relative amount of contamination across a site, (ii) determine on a site-specific basis which parameters are useful for distinguishing between clean and contaminated areas of groundwater, and (iii) map the shape and direction of a general concentration plume across a transect of interest.

We illustrate that multidimensional principal component variables produced from hydrochemical and/or microbiological data are spatially correlated and can be used with geostatistical techniques to classify or predict contamination at unsampled locations and improve estimates of contamination at landfill sites. Microbiological community profiles provide an extremely valuable source of information at LTM sites. When groups of organisms specific to a contaminated environment are targeted (e.g., *Archaea*), microbial profiles can be used in a manner similar to more traditional data sources for mapping the extent of groundwater contamination. This is not to say that microbiological profiles can at this point in time replace the monitoring of hydrochemical constituents. Instead, we suggest that combining microbial and hydrochemistry information in a long-term monitoring strategy may provide a more sensitive measure of contamination and contribute to reducing uncertainties at undersampled locations. We also hope that as the field of environmental microbiology evolves, the direct monitoring of microbial profiles (without the need for excessive validation of sequences) may result in a useful and cost-effective tool for managers at long-term monitoring sites.

## Acknowledgments

We thank the three anonymous reviewers, Lori Stevens, and Nancy Hayden for very helpful reviews of the manuscript. Work was supported, in part, by a Vermont NSF EPSCoR Graduate Research Assistantship 524474.

## Literature Cited

(1) EEA. *Europe's environment: the third assessment*; Environmental assessment report No. 10. Office for Official Publications of the European Communities, European Environmental Agency: 2003. European Environmental Agency: 2003.

(2) U.S. EPA. *Municipal Solid Waste in the United States: 2003 Facts and Figures*; Office of Solid Waste and Emergency Response (5305W): PA530-F-05-003, Environmental Protection Agency, Washington, DC, April 2005.

(3) U.S. EPA. *National Water Quality Inventory, 1998 Report to Congress*, Office of Water (4503F): EPA841-R-00-001, Environmental Protection Agency, Washington, DC June 2000.

(4) U.S. EPA. *A study of the implementation of the RCRA corrective action program*; Office of Solid Waste, Economics, Methods, and Risk Analysis Division (5307W), Environmental Protection Agency, Washington, DC, 2002, http://www.epa.gov/epaoswer/hazwaste/ca/facility/rcaidfnl.pdf.

(5) Albrechtsen, H.-J.; Heron, F.; Christensen, J. B. Limiting factors for microbial Fe(III)-reduction in a landfill leachate polluted aquifer (Vejen, Denmark). *FEMS Microbial Ecol.* **1995**, *16*, 233−248.

(6) Bekins, B. A.; Godsy, E. M.; Warren, E. Distribution of microbial physiologic types in an aquifer contaminated by crude oil. *Microbial Ecol.* **1999**, *37*, 263−275.

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

(8) Ludvigsen, L.; Albrechtsen, H.-J.; Ringelberg, D. B.; Ekelund, F.; Christensen, T. H. Distribution and composition of microbial populations in a landfill leachate contaminated aquifer (Grindsted, Denmark). *Microbial Ecol.* **1999**, *37*, 197−207.

(9) Oerther, D. B.; Love, N. G. The value of applying molecular biology tools in environmental engineering: Academic and industry perspective in the USA. *Rev. Environ. Sci. Bio/Technol.* **2003**, *2*, 1−8.
(10) Maymo-Gatell, X.; Chien, Y.; Gossett, J. M.; Zinder, S. H. Isolation of a Bacterium that reductively dechlorinates tetrachlorethene to ethene. *Science* **1997**, *276*, 1586−1571.
(11) Seshadri, R.; Adrian, L.; Fouts, D. E.; Eisen, J. A.; Phillippy, A. M.; Methe, B. A.; Ward, N. L.; Nelson, W. C.; Deboy, R. T.; Khouri, H. M.; Kolonay, J. F.; Dodson, R. J.; Daugherty, S. C.; Brinkac, L. M.; Sullivan, S. A.; Madupu, M.; Tran, K.; Robinson, J. M.; Forberger, H. A.; Fraser, C. M.; Zinder, S. H.; Heidelberg, J. F. Genome Sequence of the PCE-dechlorinating Bacterium *Dehalococcoides ethenogenes*. *Science* **2005**, *307*, 105−108.
(12) Madsen, E. L. Nucleic acid characterization of the identity and activity of subsurface microorganisms. *Hydrogeol. J.* **2000**, *8*, 112−125.
(13) Head, I. M.; Saunders, J. R.; Pickup, R. W. Microbial evolution, diversity, and ecology: a decade of ribosomal RNA analysis of uncultivated microorganisms. *Microbial Ecol.* **1998**, *35*, 1−21.
(14) Muyzer, G. Genetic fingerprinting of microbial communities—present status and future perspectives. In *Microbial Biosystems: New Frontiers, Proceedings of the 8th International Symposium on Microbial Ecology*; Bell, C. R., Brylinsky, M., Johnson-Green, P., Eds. Atlantic Canada Society for Microbial Ecology: Halifax, Canada, 1999.
(15) Uz, I.; Rasche, M. E.; Townsend, T.; Ogram, A. V.; Lindner, A. S. Characterization of methanogenic and methanotrophic assemblages in landfill samples. *Proc. R. Soc.London* **2003**, *270*, S202−S205.
(16) Watanabe, K.; Watanabe, K.; Kodama, Y.; Syutsubo, K.; Harayama, S. Molecular characterization of bacterial populations in petroleum-contaminated groundwater discharged from underground crude oil storage cavities. *Appl. Environ. Microbiol.* **2000**, *66*, 4803−4809.
(17) Röling, W. F. M.; Van Breukelen, B. M.; Braster, M.; Lin, B.; Van Verseveld, H. W. Relationship between microbial community structure and hydrochemistry in a landfill leachate-polluted aquifer. *Appl. Environ. Microbiol.* **2001**, *67*, 4619−4629.
(18) Röling, W. F. M.; Van Breukelen, B. M.; Braster, M.; Goeltom, M. T.; Groen, J.; Van Verseveld, H. W. Analysis of microbial communities in a landfill leachate polluted aquifer using a new method for anaerobic physiological profiling and 16S rDNA based fingerprinting. *Microbial Ecol.* **2000**, *40*, 177−188.
(19) Boon, N.; Marle, C.; Top, E. M.; Verstraete, W. Comparison of the spatial homogeneity of physicochemical parameters and bacterial 16S rRNA genes in sediment samples from a dumping site for dredging sludge. *Appl. Microbiol. Biotechnol.* **2000**, *53*, 742−747.
(20) Webster, R.; Boag, B. Geostatistical analysis of cyst nematodes in soil. *J. Soil Sci.* **1992**, *43*, 583−595.
(21) Saetre, P.; Bååth, E. Spatial variation and patterns of soil microbial community structure in a mixed spruce-birch stand. *Soil Biol. Biochem.* **2000**, *32*, 909−917.
(22) Franklin, R. B.; Mills, A. L. Multi-scale variation in spatial heterogeneity for a microbial community structure in an eastern Virginia agricultural field. *FEMS Microbiol. Ecol.* **2003**, *44*, 335−346.
(23) Horner-Devine, M. C.; Lage, M.; Hughes, J. B.; Bohannan, B. J. M. A taxa-area relationship for bacteria. *Nature* **2004**, *432*, 750−754.
(24) Ludvigsen, L.; Albrechtsen, H.-J.; Holst, H.; Christensen, J. B. Correlating phospholipid fatty acids (PLFA) in a landfill leachate polluted aquifer with biogeochemical factors by multivariate statistical methods. *FEMS Microbiol. Rev.* **1997**, *20*, 447−460.
(25) Franklin, R. B.; Blum, L. K.; McComb, A. C.; Mills, A. L. A geostatistical analysis of small-scale spatial variability in bacterial abundance and community structure in salt marsh creek bank sediments. *FEMS Microbiol. Ecol.* **2002**, *42*, 71−80.
(26) Ritz, K.; McNicol, J. W.; Nunan, N.; Grayston, S.; Millard, P.; Atkinson, D.; Gollotte, A.; Habeshaw, D.; Boag, B.; Clegg, C. D.; Griffiths, B. S.; Wheatley, R. E.; Glover, L. A.; McCaig, A. E.; Prosser, J. I. Spatial structure in soil chemical and microbiological properties in an upland grassland. *FEMS Microbiol. Ecol.* **2004**, *49*, 191−205.
(27) Caeiro, S.; Goovaerts, P.; Painho, M.; Costa, M. H. Delineation of estuarine management areas using multivariate geostatistics: the case of Sado Estuary. *Environ. Sci. Technol.* **2003**, *37*, 4052−4059.
(28) Oliver, M. A.; Webster, R. A geostatistical basis for spatial weighting in multivariate classification. *Math. Geol.* **1989**, *21*, 15−35.
(29) Castrignanò, A.; Giugliarini, L.; Risaliti, R.; Martinelli, N. Study of spatial relationships among some soil physicochemical properties of a field in central Italy using multivariate geostatistics. *Geoderma* **2000**, *97*, 39−60.
(30) Van Breukelen, B. M.; Röling, W. F. M.; Groen, J.; Griffioen, J.; van Verseveld, H. W. Biogeochemistry and isotope geochemistry of a landfill leachate plume. *J. Contam. Hydrol.* **2003**, *65*, 245−268.
(31) Van Breukelen, B. M. PhD Thesis, *Natural Attenuation of Landfill Leachate: a Combined Biogeochemical Process Analysis and Microbial Ecology Approach*; Vrije University: Amsterdam, 2003.
(32) Yoon, J. R.; Lee, K.; Kwon, B. D.; Han, W. S. Geoelectrical surveys of the Nanjido waste landfill in Seoul, Korea. *Environ. Geol.* **2003**, *43*, 654−666.
(33) Kayabali, K.; Yuksel, F. A.; Yeken, T. Integrated use of hydrochemistry and resistivity methods in groundwater contamination caused by a recently closed solid waste site. *Environ. Geol.* **1998**, *36*, 227−234.
(34) Lunne, T.; Robertson, P. K.; Powell, J. J. M. *Cone Penetration Testing in Geotechnical Practice*; E and FN SPON: New York, 1997.
(35) Röling, W. F. M.; Van Breukelen, B. M.; Braster, M.; Van Verseveld, H. W. Linking microbial community structure to pollution: Biolog-substrate utilization in and near a landfill leachate plume. *Water Sci. Technol.* **2000**, *41*, 47−53.
(36) Van Verseveld, H. W.; Röling, W. F. M. Cluster analysis and statistical comparison of molecular community profile data. In *Molecular Microbial Ecology Manual*, 2nd ed.; Kluwer Academic Publishers: Amsterdam, 2004; Vol. 1.7.4, pp 1−24.
(37) Afifi, A. A.; Clark, V. *Computer-Aided Multivariate Analysis*, 3rd ed.; CRC Press: Boca Raton, 1997.
(38) Journel, A. G.; Huijbregts, C. *Mining Geostatistics*; Academic Press: New York, 1978.
(39) Brockman, F. J.; Murray, C. J. Subsurface microbiological heterogeneity: current knowledge, descriptive approaches, and applications. *FEMS Microbiol. Rev.* **1997**, *20*, 231−247.
(40) Felske, A.; Akkermans, A. D. L. Spatial homogeneity of abundant bacterial 16S rRNA molecules in grassland soils. *Microbial Ecol.* **1998**, *36*, 31−36.
(41) Goovaerts, P. Geostatistical tools for characterizing the spatial variability of microbiological and physicochemical soil properties. *Biol. Fertil. Soils* **1998**, *27*, 315−334.
(42) Goovaerts, P. Geostatistics in soil science: state-of-the-art and perspectives. *Geoderma* **1999**, *89*, 1−45.
(43) Legendre, P. Spatial Autocorrelation: Trouble or New Paradigm? *Ecology* **1993**, *74*, 1659−1673.
(44) Petrone, R. M.; Price, J. S.; Carey, S. K.; Waddington, J. M. Statistical characterization of the spatial variability of soil moisture in a cutover peatland. *Hydrol. Process.* **2004**, *18*, 41−52.
(45) Robertson, G. P. Geostatistics in Ecology: Interpolating With Known Variance. *Ecology* **1987**, *68*, 744−748.
(46) de Marsily, G. *Quantitative Hydrogeology: Groundwater Hydrology for Engineers*; Academic Press: Orlando, 1986.
(47) Matheron, G. *The Theory of Regionalized Variables and Its Applications*; Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau, 1971.
(48) Isaaks, E. H.; Srivastava, R. M. *An introduction to applied geostatistics*; Oxford University Press: New York, 1989.

*Received for review February 8, 2005. Revised manuscript received June 7, 2005. Accepted July 18, 2005.*

ES0502627