Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring

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

[1]College of Engineering and Mathematical Sciences, University of Vermont, 213 Votey Hall, 33 Colchester Ave., Burlington, VT 05405

[2]Sanborn, Head & Associates, Inc. 95 High Street, Portland, ME 04101

[*]Corresponding author – email: arpearce@uvm.edu, Phone: 802-656-XXXX, Fax: 802-656-XXXX

Abstract (max 250 wds.)

We present an application of a clustering method for delineating distinct functional zones for subsurface environmental investigations. The method is data-driven and based on an existing Artificial Neural Network (specifically, a Kohonen Self-Organizing-Map (SOM)) that performs cluster analysis. A non-parametric MANOVA optimizes the number of clusters used for interpretation.

This methodology is applied to a set of microbial data collected from 25 groundwater wells to test the hypothesis that microbial community structure can distinguish between different regions of contamination within a plume. The algorithm successfully distinguished a gradient from clean to contaminated sampling locations using the microbial community structure. At a small number of clusters (2) the method can distinguish between clean and contaminated water. Increasing the number of clusters creates groups along a gradient of contamination. Optimizing the number of clusters using a non-parametric MANOVA suggests a number of clusters for interpretation. The landfill leachate application suggests that microbial communities of may be used to

delineate spatial zones of contamination and the technique could be further developed to support long-term monitoring of contaminated sites.

Keywords: Landfills, Neural Networks, Geomicrobiology

Introduction

Microorganisms have adapted metabolic survival strategies to thrive in extreme environments. For example, *Deinococcus radiodurans* is able to withstand high levels of radiation through specialized methods of DNA repair (Minton 1994). And, a number of acidophilic microorganisms are capable of living at very low pH by actively transporting hydrogen ions out of the cytoplasm against a concentration gradient (Johnson 1998; Canfield et al. 2005). Microbial community composition reflects, and adjusts to changes in, environmental hydrochemistry by virtue of a diverse array of highly adaptable metabolic strategies. Microorganisms are capable of breaking down a variety of common petroleum contaminants (Van Hamme et al. 2003). Dissimilatory reduction of toxic heavy metals by microorganisms can both mobilize arsenic (Stolz and Oremland 1999) and immobilize uranium (Lovley et al. 1991). The reductive dehalogenating metabolism of *Dehalococcoides ethenogenes* has adapted to require anthropogenic chlorinated solvents such as PCE and its breakdown products for its survival (Maymo-Gatell 1997; Magnuson 1998).

Community structure of microorganisms in contaminated areas depends on the array of organisms capable of living in the particular hydrochemical environments. In

uncontaminated water and soils, diverse communities of microorganisms thrive (ref). In an extreme environment such as the acid mine drainage of Iron Mountain, CA, microbial community structure is fairly simple; only a handful of distinct types of organisms have been documented, all of which were related to known acidophiles specifically adapted to life at very low pH (Bond et al. 2000). Common soil and groundwater contamination may not present such an extreme case; but microbial community structure in an uncontaminated environment can be altered by introducing contaminants (Grant et al. 2006). This suggests a dynamic relationship between the 'extreme-ness' of the environment and the complexity of the community structure. Because of the rapid and multifarious adaptations of microbes for coping with (and sometimes necessitating) contamination and harsh conditions of all sorts, their community structure could be valuable for describing the nature and extent of soil and groundwater contamination given the appropriate computational tools.

Challenges of Long Term Monitoring

In the case of leaking landfills or other sites with multiple types of contamination, there are a myriad of contaminants, breakdown products, and complex biogeochemical interactions, which together are difficult if not impossible to identify. If site management is primarily concerned with tracking and delineating the extent of impacted groundwater, each of these must be monitored. Long-term monitoring of contaminated groundwater poses many financial and technical difficulties. Once any active remediation is completed many sites are monitored for decades, and it is vital that screening tools are affordable and accurate.

There are numerous variables and data types (e.g. categorical, continuous) associated with groundwater investigations and these variables are generally autocorrelated in space and time, violating assumptions of parametric statistical tests. Often environmental managers wish to group sample locations with similar hydrochemical features. Also, the optimal number, composition and interpretation of these groups are unknown at the outset. Clustering methods are particularly attractive for these types of exploratory data analyses because they do not require many (if any) assumptions about the data, either the target number of groupings or the structure of the data at the outset. They are useful tools for exploring interrelationships among the data to make initial evaluation of the overall organization (Jain et al. 1996). Non-linear methods have been shown to account for more data variability than linear methods when applied to geochemical and microbiological datasets (Schryver et al. 2006). In particular, artificial neural network (ANN) clustering methods have been shown be more robust than traditional methods for clustering hydrochemical contamination data (Gagne and Blaise 1997, Solidoro et al. 2007).

The Self-Organizing Map as an Environmental Monitoring Tool

Artificial neural networks (ANNs) are data-driven, non-linear data-mining tools roughly based on hypothesized mechanisms of human learning. As a tool for mining hydrochemical and microbial data, ANNs can exploit complex functional relationships within a dataset without explicitly defining them, as is the goal with physics-based modeling approaches. The Self Organizing Map (SOM), or Kohonen Map, a type of ANN, was developed in the 1980's by Teuvo Kohonen, a Finnish researcher of self-

organization, associative memories, neural networks, and pattern recognition (Kohonen 1983, Kohonen 1990). The neurobiological basis for the SOM originates from models of sensory (e.g. auditory, visual, tactile) information processing believed to create topological mappings on the cerebral cortex (Haykin 1999). The SOM creates a topological map of input patterns by competitive (unsupervised) learning (Haykin 1999) and can be used as a non-linear version of a principal components analysis (Ritter 1995) and a non-parametric clustering method (ref). The algorithm has been used in a broad spectrum of applications from document searching (Kohonen et al. 2000; Lagus et al. 2003) to secondary protein structure rendering (Andrade et al. 1993), to pattern classification (Marabini and Carazo 1994) to clustering patterns of gene expression (Tamayo et al. 1999). The SOM has also been used successfully in a variety of water quality applications, including distinguishing between waters of varying trophic status based on hydrochemical measurements (Aguilera 2001), and to distinguish between clean and polluted portions of a river based on bacterial and macroinverterbrate communities (Kim et al. 2007).

As a clustering algorithm, the self-organizing map (SOM) is non-linear and non-parametric. It outperforms many traditional clustering methods (e.g. hierarchical and K-means) on datasets with high dispersion, outliers, irrelevant variables and non-uniform cluster densities (Mangiameli et al. 1996). Extracting different numbers of clusters from an SOM can capture multiple meanings of groupings within a dataset. For example, in a study of a polluted river Kim et al. (2008) found a small number of clusters separated different levels of contamination, while a larger number of clusters separated the dataset

122 based on seasonal biogeochemical differences. Park et al. (2004) found different
123 meanings at successive hierarchical levels clustering the topology of an SOM mapping
124 benthic macroinverterbrate populations in streams with a range of land use impacts.
125
126 Clustering methods as a statistical tool do not assign significance to the clusters they
127 generate or optimize the number of clusters generated. In general, we seek a method to
128 maximize variability between clusters and minimize variability within clusters. Several
129 authors have used a parametric MANOVA (Reyjol et al. 2005; Park et al 2006) to
130 determine significance of the clusters generated from an SOM. Other methods of
131 optimizing the number of clusters in a data set are also based on maximizing the ratio of
132 the between and within cluster sums of squares (Milligan and Cooper 19XX; Calinski
133 and Harabasz 19XX) including the gap statistic (Tibisrani et al. 2001) and the DB index
134 (Davies and Bouldin 1997).
135
136 The goal of this research is to develop a method for exploratory data analysis that clusters
137 data based on microbial community composition. There are three specific objectives.
138 First, we use a non-parametric MANOVA in tandem with an SOM to optimize the
139 number of clusters created by the algorithm. This will work toward creating a repeatable,
140 objective way to evaluate between potential sets of groupings. Second, we apply the
141 technique to a complex spatial dataset of microbial communities and geochemistry from
142 the Schuyler Falls landfill as a proof of concept. Third, we speculate about the meaning
143 of the clusters and discuss scientifically plausibility. Since this is an exploratory method
144 for mining data, not a hypothesis test, there is no correct answer, though results will

ideally guide researchers toward more efficient and effective sampling designs for future monitoring or experiments.

Methods

Field Data Collection & Microbiological Analytical Methods

The field data collection site is the Schuyler Falls landfill, an unlined municipal landfill in Schuyler Falls, NY, USA. Detailed site information can be found in Mouser 2006 (Dissertation) and in documents from the State of NY (refs). The closed and capped landfill is situated near the Saranac River (Figure 1a). Leachate has leaked from the landfill and penetrated nearby groundwater. Water quality is monitored quarterly via an array of monitoring wells in the vicinity of the landfill. This regular monitoring analyzes for approximately 10 different heavy metals and over 30 individual organic compounds (ref). Sampling locations and the approximate extent of contamination visible in Figure 1b, show significant migration in the direction of groundwater flow, toward the Saranac River. The relative overall extent of contamination at this multi-contaminant site is represented as specific conductivity, which correlates well with several of the major contaminants found at the site (Mouser 2006).

Samples for the microbiological analysis were collected by bailing monitoring wells in March 2005 (Mouser et al. in progress). Microbiological analytical methods are detailed in Mouser et al. (in progress). Briefly, DNA was isolated from water samples and 3 different primers were used to amplify regions of DNA specific to Archaea, Bacteria and Geobacteracea via the polymerase chain reaction (PCR) (Mullis XXXX). Through a T-

RFLP analysis and data post-processing, the communities were described by Mouser et al. (in progress) as relative abundance of operational taxonomic units for each of the three chosen taxonomic targets. Principal components of the array of relative abundances were used as input to the computational methods presented where 75% of the variance of each set was accounted for by the first 2, 3, and 3 principal components of the Bacteria, Archaea, and Geobacteracea data, respectively. These 8 metrics for each monitoring well combined to form input patterns for the computational method. The complete dataset is available for 25 monitoring wells at the site.

Computational Methods

The basis of our method is the SOM, a 1-layer (of weights) network based on a 2-dimensional rectangular grid of output nodes (Figure 2). SOM Network architecture is described briefly here and in detail elsewhere (Kohonen XXXX ; Haykin XXXX). All computational methods were implemented by the author using MATLAB (version 7.4). SOM input nodes are fully connected to the output nodes by bundles of synaptic weights. The number of input nodes is defined as the number of parameters in the input patterns. Values of each input parameter are standardized between 0 and 1 as:

$$x_{norm} = \left(x - \min(x)\right) / \left(\max(x) - \min(x)\right)$$

allowing for the variation in each parameter to be considered equally by the ANN algorithm.

Unsupervised training begins by individually calculating the distance between the parameters of an input pattern to the bundle of synaptic weights connected to each node

191 on the 2-dimensional map. For this application we use Euclidian distance, though other
192 distance measures could be used. The node on the 2-dimensional map closest (minimum
193 Euclidian distance) to the input pattern is chosen as the best match. A neighborhood of
194 nodes in the 2-dimensional map is selected and all weights of the best matching node and
195 nodes in the surrounding neighborhood are updated according to the rule:

$$w(i,j)^k_{new} = w(i,j)^k_{old} + \alpha\left(pattern(k) - w^k_{old}\right)$$

196

197 where α is a learning parameter with range (0, 1)

198 w_{ijk} is the weight at node (i,j) for variable k

199 and pattern(k) is the k_{th} variable of the current input pattern

200

201 One training iteration is complete after each input pattern has been presented to the
202 network once and the appropriate weights are updated. Every iteration input patterns are
203 presented in a new random order. As an unsupervised method, SOM network training is
204 executed for a predetermined number of iterations. In this application the network trains
205 in two phases, an ordering phase and a fine-tuning phase. The neighborhood radius
206 (around the best matching node) and learning parameter decrease exponentially during
207 the ordering phase and linearly during the fine-tuning phase with more iterations in the
208 fine tuning phase than the ordering phase (Haykin 1999).

209

210 A method for choosing the optimal size of an SOM map considers quantitative error (qe),
211 calculated as the mean distance between training patterns and their final best matching
212 units (Kohonen 1991), and topographic error (te), the percentage of training patterns for

213 which the best matching unit and the second best matching unit are not adjacent (Kiviluto
214 1996). Minimizing these two metrics can optimize the map size which best captures the
215 topology of the dataset in 2-dimensions. However, as map size increases these two
216 metrics both approach zero and as map size becomes much larger than the number of
217 input training patterns they lose any value. As map size increases quantization error can
218 decrease toward zero because map nodes can learn individual input patterns. The
219 topographic error will decrease toward zero as map size increases because unused nodes
220 will surround the pattern they trained from and eventually the second best match will
221 always be adjacent. For the small number of input patterns available in this application
222 (25), using the qe and te metrics isn't a satisfactory way to optimize map size. From a
223 survey of the ecological studies cited in this manuscript, the ratio of sample size, n to map
224 size ranges from approximately 1 to 10. We will use choose measures of cluster
225 significance determine the optimal map size rather than these map specific metrics.
226
227 Visualizing the final weights or output of the SOM algorithm can take several forms. In
228 this application our number of input patterns is relatively small, n = 25. While a small
229 number of map nodes (i.e. 2 - 10) can directly cluster the data, a larger number of nodes
230 can illustrate the overall organization of the input patterns. Examining a unified distance
231 matrix, or U-matrix, of SOM output is a convenient way to visualize the 2-dimensional
232 organization of the data. A U-matrix is the average Euclidian distance between each
233 node of the map and its immediate neighbors (reference) creating topographic divides
234 between regions of dissimilar data.
235

To compare relevance between the different numbers of clusters generated, we calculated the F-statistic from a non-parametric MANOVA suitable for unbalanced designs (Anderson 2001; McArdle and Anderson 2001; Jones 2003). In the generic form this method allows for the use of any type of distance metric to define the difference between samples, and is appropriate to use with non-parametric microbial community datasets. For the landfill application, final best matching units for each input pattern as determined by the SOM algorithm represented group membership. A Euclidian distance matrix created from the input patterns, and group membership of each pattern are input to the MANOVA.

Results and Discussion

The SOM U-matrix using the Schuyler Falls Landfill microbial data shows a separation of clean and contaminated monitoring locations in the 2-dimensional data space (Figure 3). The white 'ridges' separate regions of similarity (dark valleys) on the map. The most striking map feature is the cluster of clean sampling locations (C) in the upper right hand corner in the contiguous dark valley, implying that these points were fairly homogeneous. The remainder of the map is somewhat more difficult to interpret since there are sharp gradients between many neighboring patterns. Samples collected from three polluted wells (P) are mapped to a region on the left hand side and samples with small but detectable contamination (F) spread out over the remaining area. The isolation of single points within the contours of the U-matrix indicates that community structure is more similar in clean wells than in contaminated wells. It also suggests that the number of nodes used to generate this map is over-fitting the data. This was anticipated given the

sample of 25 wells and 80 nodes. Regardless, it is still useful for observing the overall 2-dimensional mapping of this data set.

This research focuses on using the SOM as a clustering method. Therefore we determine cluster membership with the SOM for each input pattern, as well the F-statistic for 9 different sized SOMs, from 2 to 10 nodes (Table 1). The F-statistic is the ratio of the between and within group mean square error, increasing as groupings become more meaningful. Maximizing this F-statistic should indicate the optimal number of clusters for the given dataset. The clusters generated by the algorithm for 8 and 10 nodes are identical. We do not want to consider an increasing the number of clusters due to concerns of over-fitting the data. With such a small number of data points to be clustered we need to be aware that as the number of points within each cluster shrinks our within cluster variability will continue to shrink and our between cluster variability will continue to grow, falsely implying greater significance. We suggest that 4 clusters may be significant. The calculated value of the F-statistic plateaus between 4 and 7 clusters (17.1 and 18.8) then increases substantially again at 8 clusters to 24.

It is ultimately the spatial location of the clusters in Table 1 that reveal how well this method can delineate regions of contamination. Numbers (for 2, 3 and 4 clusters) superimposed on a site map showing the generalized extent of the plume (Figures 4a, 4b, and 4c) illustrate the spatial continuity of the clusters. Since the clusters were generated using only microbial abundance data, it is remarkable how well the clusters reflect the gradient of contamination at the site.

Since the relative abundance of different microorganisms is so closely tied to favorable metabolic pathways in a particular hydrochemical environment, we can speculate that the cluster divisions are related to the most available carbon sources and electron acceptors. Performing the same analysis separately for the three sub-groups of microorganisms reveal that Bacteria and its subset Geobacter are more adept at categorizing the relative level of contamination at this site than the community of Archaea. Thought they are detected throughout the site, many Archaea are adapted to particularly extreme environments and as a result, we might expect them to be most prolific in a methanogenic region of a contaminant plume (reference). Including only the Bacteria and Geobacter data as input to the SOM algorithm does change the spatial distribution of the clusters (Figure 4d). The SOM performs better when irrelevant variables are removed from the input dataset (Mangiameli et al. 1996), so our analysis may benefit from removing Archaea from the input. Future analysis of similarly contaminated sites might benefit from focusing on other subsets of organisms within specific regions of the bacterial domain. For example, Geobacter are known to reduce heavy metals and may be particularly effective indicators for specific types of contamination (Lovley 19XX). Even though we will likely never find a one to one mapping between contaminants and microorganisms, communities of microorganisms can be good indicators of gradients in electron acceptor and nutrient availability.

At a variety of levels of clustering, MW219 (Table 1) consistently clusters separately from the uncontaminated sampling locations, implying that the microbial communities

305 share more in common with the contaminated wells than the clean wells. Although
306 significant contamination was not detected by hydrochemical monitoring at MW 219,
307 there may be fingering extending southeast toward MW 218 and 219. Profiles thorough a
308 contaminant plume can show differences in microbial metabolic activity related to
309 contaminant concentration, suggesting degradation may occur more rapidly in the fringes
310 of a plume rather than in the most concentrated area (Pickup et al. 2001; Windrel et al.
311 2008). Perhaps MW219 is in an area where the community of microorganisms has
312 changed due to the introduction of low levels of contamination and is an advanced
313 indictor of migrating pollution.
314

315 **Conclusions and Implications**

316 Geochemical gradients and transformations in the environment are intricately coupled to
317 microbial communities and their metabolic processes, yet it is virtually impossible to
318 explicitly describe one as a function of the other based on mechanistic or predictive
319 models. We demonstrate that non-linear methods such as the self-organizing map
320 artificial neural network are effective at distinguishing between different communities of
321 microorganisms and suggesting the spatial extent of functional zones of a plume. This
322 research is a first step toward using microorganisms to delineate spatial patterns of
323 subsurface contamination. Additional work in the area could be useful for site
324 characterization and long-term monitoring.
325

326 **Acknowledgements**

The authors thank Casella Waste Services and the US EPA, for site access, sampling assistance and site information, N. Gotelli and L. Stevens for statistical help, as well as the students in GEOL 371 at UVM for their critical reviews. This research is supported in part by a fellowship to Pearce under NSF grant NSF EPS-XXXXXXX (Vermont EPSCoR).

References – In Progress…

Figure Captions

Figure 1: Schuyler Falls Landfill, 1a) Site location map 1b) Landfill site plan with extent of contamination (electrical conductivity). Figure from Mouser 2009 (in progress)

Figure 2: Architecture of the self-organizing map. Input patterns are presented to the network one at a time for training, and $w(i,j)^1$ through $w(i,j)^K$ compose the bundle of synaptic weights associated with map node (i,j) connecting it to the K features of each input pattern.

Figure 3: Unified distance Matrix of the SOM output created from an 80-node (8 by 10) map using Archaea, Bacteria and Geobacter community data. Locations of the colored letters show the best matching nodes for the 25 monitoring wells. The colors and letters correspond with three approximate classes of contamination at that monitoring location determined by Mouser et al (in progress): clean, fringe and polluted.

Figure 4: Spatial locations of SOM generated clusters using Archaea, Bacteria, and Geobacter as input superimposed on the site map using a) 2 clusters, b) 3 clusters and c) 4 clusters. d) Spatial locations of SOM generated clusters using only Bacteria and Geobacter community data.

Figures

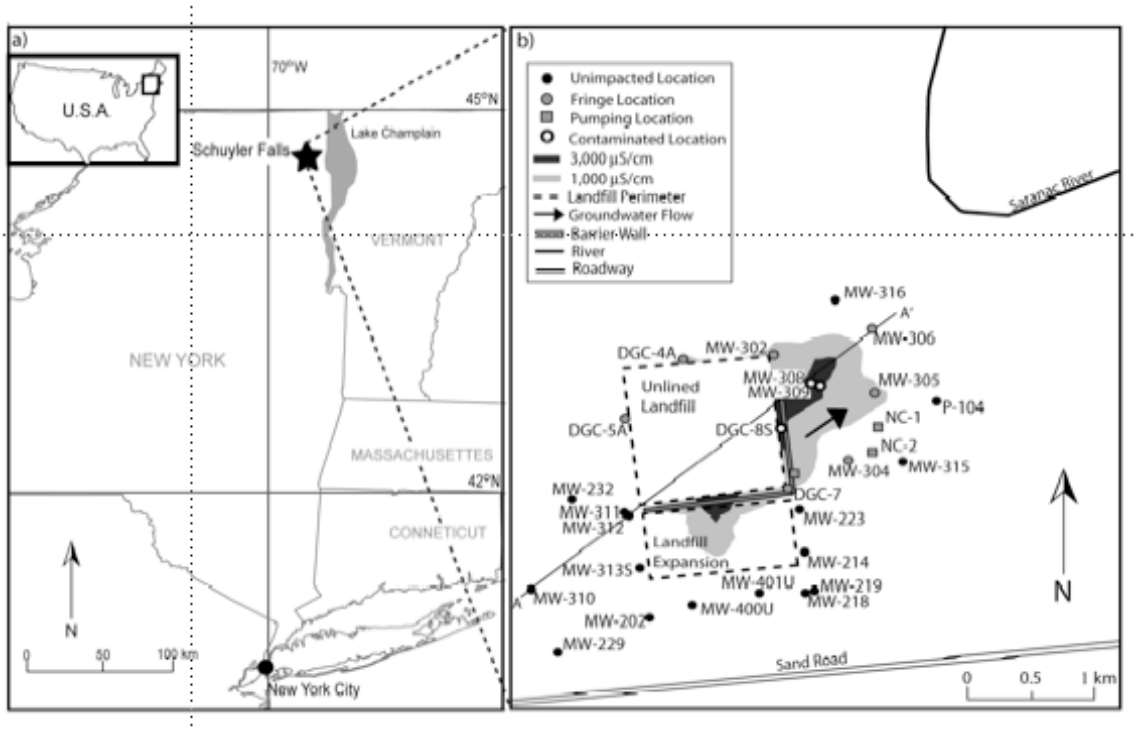
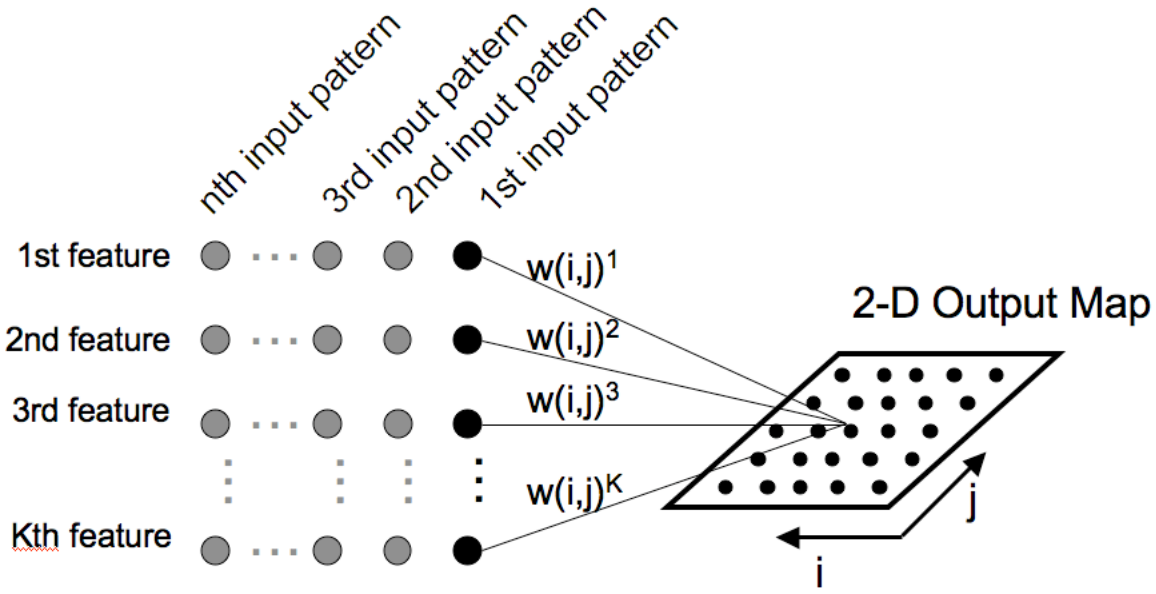
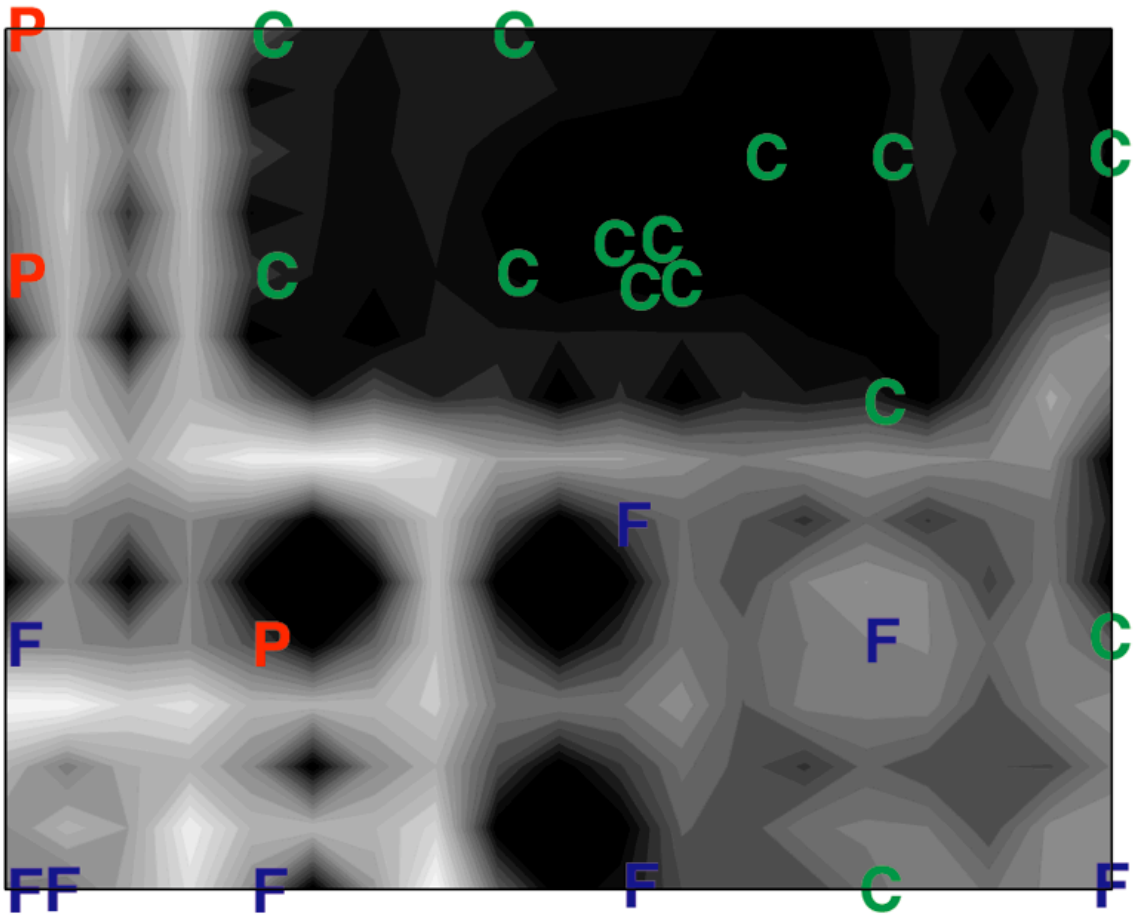


Figure 1



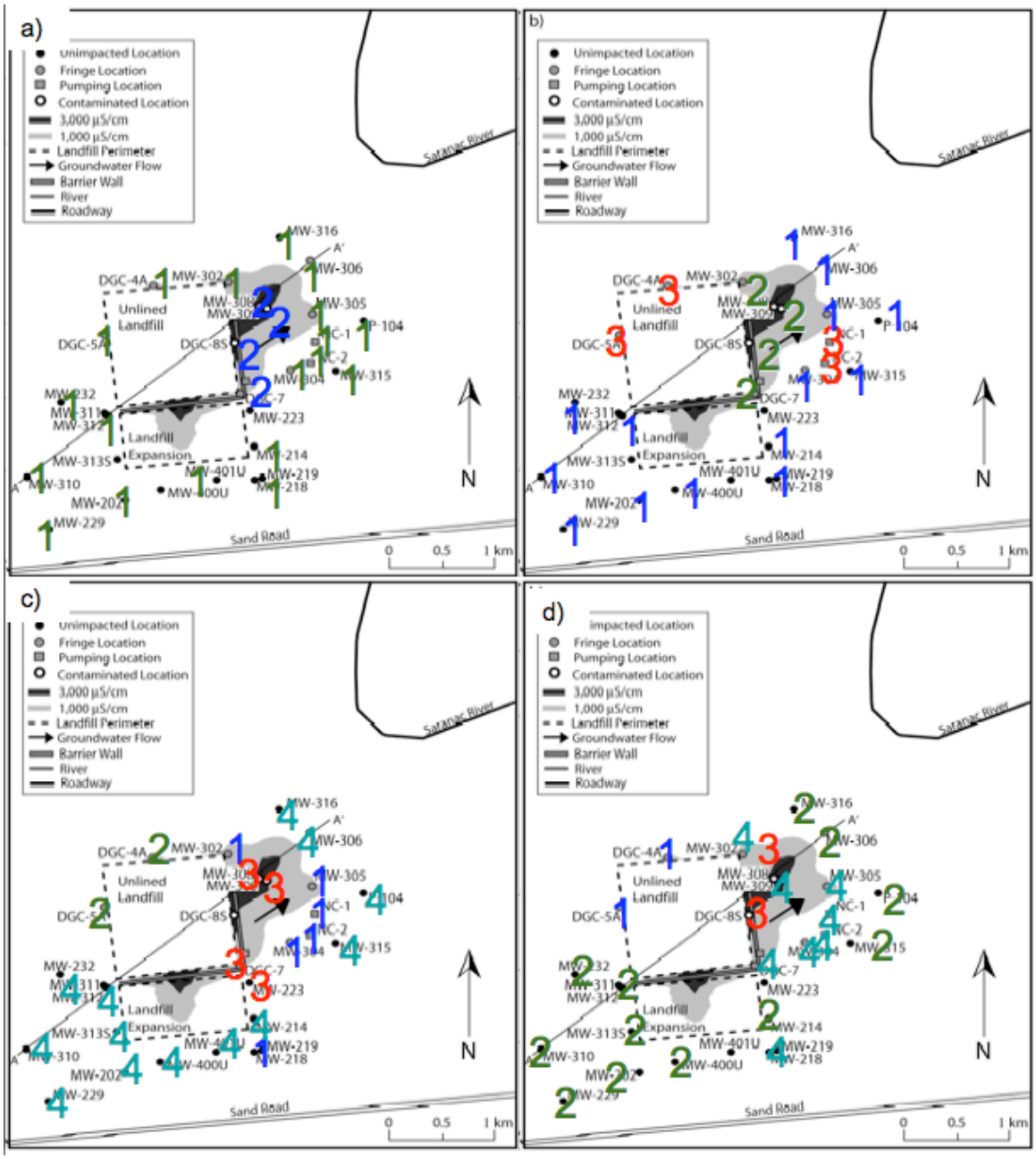
359

360 Figure 2



361

362 Figure 3



363

364 Figure 4 – I know this is clunky-don't yet have the data to reproduce background.

365

366 **Table Captions** (Table will also be presented in-line in the manuscript text)

367 Table 1 – Results of the SOM clustering using microbial data describing the Archaea,

368 Bacteria and Geobacteracea communities for 2 through 10 clusters.

Tables

Table1

Map Nodes	2	3	4	5	6	7	8	9	10	
Monitoring Well	Group Membership									Clean, Fringe or Polluted?
MW-202	1	1	4	3	3	2	1	3	1	Clean
MW-214	1	1	4	3	3	2	1	3	1	Clean
MW-219	1	1	2	5	2	3	6	6	4	Clean
MW-229	1	1	4	3	3	2	1	3	1	Clean
MW-232	1	1	4	3	3	2	1	3	1	Clean
MW-306	1	1	4	3	3	2	1	3	1	Clean
MW-310	1	1	4	3	3	2	1	3	1	Clean
MW-312	1	1	4	3	3	1	1	1	1	Clean
MW-313	1	1	4	3	3	2	1	3	1	Clean
MW-315	1	1	4	3	3	2	1	2	1	Clean
MW-316	1	1	4	3	3	2	1	2	1	Clean
MW-400U	1	1	4	3	3	2	1	3	1	Clean
MW-401U	1	1	4	3	3	2	1	3	1	Clean
P-104	1	1	4	3	3	2	1	3	1	Clean
DGC-4A	1	3	1	1	5	7	8	9	7	Fringe
DGC-5A	1	3	1	1	5	7	8	9	7	Fringe
NC1	1	3	2	5	4	6	5	8	2	Fringe
NC2	1	3	2	5	2	3	7	5	8	Fringe
MW-302	1	1	2	5	4	6	5	8	2	Fringe
MW-304	1	1	2	5	2	3	6	6	4	Fringe
MW-305	1	1	2	5	2	3	6	6	4	Fringe
DGC-7	2	1	3	4	6	4	4	4	6	Fringe
DGC-8S	2	2	3	2	1	5	3	7	5	Polluted
MW-308	2	2	3	2	1	5	3	7	5	Polluted
MW-309	2	1	3	4	6	4	2	4	3	Polluted
Between Group Covariance	61.2	63.8	54.1	44.1	37.9	32.4	29.6	25.8	29.6	
Within Group Covariance	7.2	4.6	3.1	2.6	2.0	1.9	1.2	1.4	1.2	
F-Statistic	8.4	14.0	17.3	17.1	18.8	17.3	24.0	19.1	24.0	