Lee Corbett
GEOL 371
4/15/09

Review of: Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring
Authors: A.R. Pearce, P.J. Mouser, and D.M. Rizzo

In this manuscript, the authors present a study that uses a self-organizing map (SOM) to analyze microorganism communities and therefore track contaminants in groundwater from an unlined landfill. The research appears to incorporate an interdisciplinary, synthetic view of groundwater monitoring and it seems like this approach may be applicable to other sites.

Overall, the manuscript is well organized and well written. The authors do a very good job of providing the necessary background information in an approachable manner to that all readers will be able to understand the manuscript regardless of their previous knowledge of SOMs.

One additional note: this manuscript is very far out of my field, and much more quantitative than manuscripts I typically read. Therefore, I am unable to provide an adequate assessment of the methods, data, and interpretations you present here. Hopefully my comments (focused more on the writing and less on the science) will still be useful.

Please refer to the hard-copy edited version of the manuscript for small comments regarding structure and rhetoric. In addition, I have several broader comments:

1.) I think your introduction is very solid and the information you present is helpful. Good work! My only concern with the introduction is that the reader has to wade through almost five pages of information, most of it fairly technical, before they can find out the goal of your study. Unfortunately, this isn't an easy problem to solve because all of this background information is necessary for the reader to understand the detailed description of your goals that you have at the end of your introduction. Do you think it would be possible to work in a brief, simplified statement of the purpose of the study earlier in your introduction?

2.) In the first paragraph of your methods section (where you describe the study site), it might be helpful to give a description of the current water quality (i.e. the reader might like to know whether you are dealing with a very polluted site or a not-so-polluted site).

3.) I think your manuscript successfully argues that your method has potential to be further developed and used at other sites. If this is your goal, however, I would be interested in learning more about the feasibility of doing this type of work (e.g. how long it would take, what type of expertise you need, etc.). I think this information would be vital to people interested in using your method.

Good luck with edits and publication. Well done!

Lee (abcorbet@uvm.edu)

Review by: Carrie Pucko

Authors: Andrea R. Pearce, Paula J. Mouser, Donna M. Rizzo

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

Summary: This paper outlines a new method for using self-organizing maps and MANOVA tests to explore water quality via microbes. It describes the evolution of using cluster analyses in water quality assessment and outlines other methods previously used for this purpose and other types on environmental monitoring. The MANOVA is used to establish the ideal number of clusters while the SOM places each observation into the clusters.

Review:

While I think the fact that this method does very well in addressing the problem at hand, it was unclear to me how this problem is dealt with currently and why using the SOM and MANOVA is preferential. What are the costs involved in doing this? Why would someone measure microbial communities rather than just water samples? There was one sentence that explained that the microbial communities can accurately describe the overall health or pollution level of very complex environments, but I believe that was the only time anything like that was addressed. As far as the details about the evolution of ANN's and clustering methods, I thought this was very clear and useful. I also liked how you addressed real statistical concerns that every scientist goes through when trying to find a statistical test that works. For example, the reasons you gave for using MANOVA's to determine cluster number was very helpful. That is a problem that anyone who has ever used cluster analyses has run into. I also liked the explanation and examples of how clustering procedures can pick up on patterns on multiple levels within data.

Overall I thought it was a well written, clear paper that served its purpose as an outline to a potentially very useful method not only in water treatment and pollution management, but in many other fields as well. I just think you need to stress how (or if) this method, and the data on which this method is based, is preferntial to simply doing water testing. I admittedly do not know very much about this topic, so perhaps some of this is too general for your audience, but it would have been helpful for me to know the advantages of this in order to see the real value in it.

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

By Pearce, Andrea R., Paula J. Mouser, Donna M. Rizzo

4/15/09
Review by Christina Syrrakou

This paper studies the use of an artificial neural network method which performs cluster analysis on microbial community data and aims at proving that community structure is primarily affected and organized according to local contamination patterns within a plume. Additionally to the ANN method and specifically the Kohonen Self Organising Map, in order to optimize the number of clusters in the data set, the writers use a non-parametric MANOVA. The paper concludes that although clustering methods do not provide a straight-forward relationship between microorganisms and contaminants, the method used is a first step in charactering contamination gradients within a plume using microbial data.

The manuscript presented is overall well written and presents in an effective way the writers' arguments. Although the topic presented is very specific I believe that the main points can be easily understood by a wide audience. Something that was really helpful in reading this paper was the division of each section (intro, methods etc.) in smaller subsections with characterizing titles. Also, it is obvious that the writers tried to avoid too much technical detail throughout the paper.

So, I recommend this paper to be published with minor revisions which are:

-In the subsection entitled "Challenges of Long Term monitoring" of the Introduction I was a bit troubled on how the first paragraph (l68-75) connects to the method presented in this paper. The reason is that it is mentioned that for site management it is important to monitor the different kinds of contaminants separately. However, as far as I understood this method cannot distinguish between the different kinds of contaminants but rather gives a spatial perspective of the contamination. So, it might be good to clarify the point mentioned.

-Although as mentioned above, the writers tried to avoid much technical info, I have to admit, and probably the main reason is my absence of background on ANNs, that the Computational Methods in the Methods part was a bit difficult to follow. I think the first part (l 178-208) contains critical terminology and therefore is necessary for the readers. However, l210-225 where the method of choosing the optimal size of a SON is presented could be either avoided or less detailed since as it is mentioned in the last part (l221-225) this method was not used due to the small input patterns available in this application.

- I think it would be good to state in a sentence at the second paragraph of the methods (l163-175) that the analysis of the specific application includes community data of all three types of microorganisms, that is Archaea, Bacteria and Geobacteria. In this way it will be easier to compare to the results from the same analysis omitting the Archaea as it is mentioned in the Results and Discussion (l286-288).

-Finally, in Figure 4 for better visual comparison between the results including all types of microorganisms and the results omitting the Archaea I think that it would be good to keep in one line Figures a,b and c and put in a second line figure d. Also, although it is stated in the figure caption you could add to the first line a caption saying "Clusters using all microorganisms" and to the second "clusters using only Bacteria and Geobacter".

As for the smaller revisions please see the hardcopy of my review.

Generally, I think it was a good paper and although it didn't have a standard "results" format it was easy to read and understand. Good luck!

Joe Bartlett

4/15/09

Review of Pearce et al.

The purpose of this study was to present a new method for delineating distinct functional zones for subsurface environmental investigations. They utilized a clustering method based on an existing Artificial Neural Network. The method was applied to microbial data collected from groundwater wells around a plume of contamination from a leaking landfill in Schuyler Falls, NY. A gradient of contamination was successfully identified using the microbial community structure. This method was shown to be effective at distinguishing presence or absence of contamination, gradient of contamination, and could be further developed to support future monitoring of contaminated sites.

The paper is very well written and contains only a few grammatical and sentence structure issues. It is well organized, concise, and fairly easy to read given the technical nature of the paper. The literature review is somewhat choppy and several of the sentences need to be better integrated. The rest of the introduction is very good, especially the description of SOM's and ANN's, which is the first time I have ever actually understood either of these methods. The introduction could use a few sentences to better frame this study with existing research/knowledge. Also, a few sentences of justification at the end of the introduction would be really helpful. Much of the first methods paragraph might read better if it were incorporated into a study site section in the Introduction. Otherwise the methods are very good. The R&D section reads well and does a great job at conveying the success of the method for this application. A short paragraph on future research considerations could be added to the end of this section. The conclusions are good, the last sentence is a little weak to end on but this could be moved to a future research consideration sub-section.

This article is very appropriate as a research paper for the Journal of the Association of Ground Water Scientists and Engineers. It should be accepted with moderate revisions. Specific recommendations for improvement are as follows:

- L23: Elaborate, what kind of plume?
- L37: Can you use examples that are related to landfills?
- L41: Awkward
- L45: Need to integrate these sentences better, a little choppy
- L53: Passive
- L149: Most of this paragraph might go better as a study site sub-section in the intro
- L163: More detail needed on the monitoring wells, did you install them?
- L295: Why are these irrelevant
- L298: Might want to make a second paragraph, too much going on in this one
- L303: Awkward

April 15, 2008

UVM internal review of:

> Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring
>
> Authors: Pearce, Andrea R., Paula J. Mouser, and Donna M. Rizzo

The overall objective of this paper is to determine if diversity of microorganisms can be used to track migrating contaminant plumes in subsurface soil deposits. The use of artificial neural networks was used to find major differences in microbial communities surrounding a leaking landfill located in close proximity to the Saranac River in New York. Results of this study showed that contaminant plumes of this particular landfill found that differences in bacteria and geobacter were primary indicators of contamination at the study site. Diversity at the site was evaluated using four sets of grouping. Results of this grouping showed remarkable similarities to the current extents of the contaminant plume as defined by chemical evaluations.

The content of this paper is quite good, with far reaching practical implication. Results from this paper clearly demonstrate the ability to track contaminant plumes using biological diversity to track chemical changes in ground water composition. Although the contents of this paper are excellent the flow of this paper makes understanding on the part of the reader difficult, causing the reader to do large amounts of work in order to understand the processes being used. Illustrations created by the author highlight the findings of this paper quite nicely. After understanding of how the grouping is determined the reader can immediately understand the long reaching implications of this paper's findings. The quality of the illustrations are not quite to the point of publishing, however the detail that seems to be desired by the author may not be necessary. Overall due to the content of this paper it should be accepted. However the flow of this paper needs some significant work to aid in the reader's understanding. After these improvements I feel that this paper will be a quite influential manuscript.

- This paper does a wonderful job at showing how the diversity of microorganism in soil deposits can be used to determine the extent of a contaminant plume. However the exact focus of this particular paper is lost in the introduction. Throughout this section the author makes note of improving the clustering of sampling location for long term monitoring. This issue, though important, does not seem to be addressed in the remainder of the paper. This particular section could be left out so that the focus of the paper may be aimed at the ability of these methods to reproduce plume geometry. Or further attention could be paid to this particular section and how using these methods would make a significant contribution to the methods that are already in place for selecting sampling schemes.

After closely reviewing the author's guidelines for this journal everything seems to be in order. I wish you the best of luck with your submission.

Jaron

Jared Nunery
GEOL 371
April 15th, 2009

Review of:

Pearce, A.R., P.J. Mouser, and D.M. Rizzo, **Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring**

For submission to:
Ground Water

This paper presents a new methodology for exploratory data analysis. Through the use of cluster data base analysis of microbial community composition, this methodology is able to address previously difficult questions, through a novel analytical technique.

Overall I thought that this paper was very well written. It is clear that the authors thoroughly understand the various analytical techniques described, and furthermore they do an excellent job of conveying the material to the reader at an understandable level. The de-emphasis of the results and discussion section effectively maintained the spotlight on the new methodology being presented. One area that could potentially be shortened is the introduction, however, the manuscript is not very long so it may not be necessary. I would suggest comparing relative lengths of articles within the journal, and seeing if the length of this manuscript is within accepted page lengths. Below I describe specific comments, separated by section.

Abstract:

In general I thought the abstract was well written and clearly highlighted the purpose and the results of this study. Two suggestions would be combining the two paragraphs as one, as this might make the abstract appear more concise, and cohesive. The second suggestion is to be careful of the use of repeated words. Distinguish is used three times in the second paragraph, and gradient is also repeated. Choosing synonyms will reduce the redundancy between sentences.

Introduction:

This section is very clearly structured. The use of section headings aids in the organization, but I wonder if transitional sentences that segway one section to the next might help the general flow. It is clear that you acknowledge the complexity of the analysis techniques you were describing, and you did an excellent job of explaining each technique. My one concern is the length of the introduction, I am unfamiliar with this journal, but it would be a good idea to check other articles in this journal and see what the common intro length is. One section that I would not cut, as I thought it was great, is your last paragraph where you describe the goals and objectives of the study. One minor comment, on line 144, you imply that hypothesis test have a right and wrong answer, is this accurate, or do these tests simply reject or accept a null hypothesis?

Methods:

As this topic is beyond my knowledge of analytical techniques, I will restrict my comments. Overall, the step by step description of the techniques made sense, though at some points I found my mind drifting (though again this is probably related to my lack of knowledge in this field). One general point that might help the flow is to maintain one constant tense throughout this section. At times you jump between tenses and I was confused (like between the first three paragraphs – see hardcopy). Also, on line 222, what does te and qe represent?

Results and discussion:

I found that in this paper, combining the results and the discussion was very effective. As the meat of this paper is the presentation of a new methodology, having a brief, combined results and discussion maintained the emphasis on the methodology. The conclusion section did an excellent job of summarizing the major conclusions; however the implications were de-emphasized. Adding another sentence or two describing how specifically this methodology will impact this field would help cement the significance of this study. This is the point in the paper to really sell this methodology as a significant contribution to the field.

Great job writing this manuscript, and best of luck with the submission. For more specific comments, see the hardcopy with my edits. If you have any questions about my comments feel free to contact me (jnunery@uvm.edu)

Paper Title: **Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring**
Paper Authors: A.R. Pearce, P.J. Mouser and D.M. Rizzo

Reviewer: **Lance E. Besaw**
Date: April 15, 2009

Summary

The authors present an application of a combined SOM-MANOVA method for classifying subsurface contamination using microbial data. They apply the method at a leaking landfill located in upstate New York. With the method they cluster microbial data and draw inferences as to how it can be used in long-term monitoring of groundwater contaminated sites.

Evaluation

The authors' contribution is noteworthy. A repeatable method for determining the proper number of clusters in and SOM is an extremely important contribution to the field of ANNs. However, the authors do not appear to have finished the contribution or this manuscript.

Regarding the data quality. The data collection methods appear to be very thorough.

Regarding the logic of interpretation, some of the claims made in the abstract do not seem to be fulfilled in the manuscript. As a particular example, the authors state in the abstract that the non-parametric MANOVA optimizes the number of SOM clusters. However, in the results and discussion section, the authors subjectively select 4 clusters to represent their data, when the MANOVA suggests 8 or 10 clusters would be better. This conflict must be addressed.

Overall the figures present good material to the reader. I feel like Figure 3 needs much more explanation and could be made more appeasing to the eye.

Recommendation

Overall, I think the manuscript is well written and it contribution is significant. However, many items need to be addressed before this manuscript can be accepted for publication. I recommend this paper be rejected in its current form. However, when it is completed, its contents would be most useful to many ANN practitioners.

Specific Comments

I feel like the abstract has a lot of good material. However, I would not separate it into 2 paragraphs. The authors should focus on the flow of the abstract to make it a single concise entity.

I feel like the introduction was poorly organized. The authors presented too much background material before they state what the paper is about. I would prefer they prioritize and reorganize the material presented and separate into an introduction section (which includes their goals) and background section. This will increase the readability of the paper and let the readers know what they are getting into (currently the paper's objectives are state on page 6).

The study site section should include more description of the site (e.g. something about geology, dates of landfill operation, etc). It appears many details have been left out.

The computational methods section does not flow as well as it could. The authors need to work on the flow of this section. In addition, they left out numerous key parameters (e.g. square or hexagonal SOM, learning coefficient, neighborhood size and number of training iterations). More detail should be provided about the non-parametric MANOVA and how it was implemented.

The results and discussion section does not seem to be finished but is well written and provides good information to the reader.

Many of the terms used in the paper could use a proper definition, as many of the readers of this journal will not be microbiologists (e.g. community structure)

Meredith Clayton
GEOL 371
15 April, 2008

Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring by Pearce et al., 2009

This paper describes the application of a data driven clustering method useful for subsurface environmental investigations for the delineation of distinct functional zones. The cluster analysis presented in this piece is performed using an existing Kohonen Self-Organizing-Map (SOM) and a non-parametric MANOVA is used to optimize the number of clusters used for interpretation. This methodology was applied to microbial data collected from 25 groundwater wells in order to test whether differences in contamination within a plume can be detected through microbial community structure. These results demonstrate that the use of non-linear methods, such as the SOM, is effective for illustrating differences between different communities of microorganisms. These methods are also useful for providing suggestions regarding the spatial extent of functional zones within a plume.

In general, this paper is well written and well organized. I believe that it is nearly ready for submission following consideration of a few suggested changes. The first change I would suggest would be working on the introduction. You have provided some great information and examples relevant to the work being described but some parts do not flow well from a reader's perspective. They seem a bit choppy as if selected pieces were inserted from another paper or literature review. I would recommend working on linking the paragraphs together with the aid of transitional phrases for more seamless progression through the intro. I would also consider the content of the methods section. I think you have done a great job of providing very detailed explanations of your methods; however, you may want to consider the level of detail provided. Depending on your intentions you may consider whether to include things such as describing changes in map size that correspond to changes in topographic error. This almost seems more like data interpretation than methods description. On the other hand, you may want to simply consider making a paragraph with a heading label that denotes, justification and implications of the methods used. This would allow a more experienced reader to read the methods without getting into the finer details a more novice reader may need. The results and discussion section seems to only need a few grammatical changes made, but I also noted the need to elaborate on a few issues described. I was troubled by your justification for using 80 nodes because you state that you expected that it would be overkill. This is especially curious because you proceed to mention future decisions made to avoid over-fitting data. I can make inferences about this but given the level of detail you have provided previously in this paper I believe you should remain as thorough throughout. Similarly, you should consider providing additional information to explain why Archaea are "irrelevant" (line 293-295). My final suggestion would be to revise the last sentence of the conclusions section. You have done a great job of describing the importance of your data but I feel that the final sentence of the conclusion is too vague. This paper deserves a stronger concluding statement. In summary, you have a few minor

things to consider revisions too but overall you have a solid paper that is almost ready for submission. Nice work.
The biggies:

- Work on linking paragraphs of introduction for fluidity
- Consider the level of detail to include in the methods (what is your goal, what reader)
- Consider creating separate heading under methods for more detailed explanations/interpretations
- Consider a more thorough explanation of why you believe archaea irrelevant and similarly why you chose to use 80 nodes when you mention that you expected this to be way too many (especially because you express caution later on about over-fitting data)
- Work on making a stronger concluding sentence

Mark Isselhardt

Critical Writing (Geology 371)

4/15/09

Pearce, A, et al., 2009. Exploratory data analysis using a self-organizing map and MANOVA for environmental monitoring

The authors describe the novel application of an existing Artificial Neural Network to delineate subsurface water contamination. This project used microbial community data from groundwater monitoring wells placed around a leaking caped landfill to generate a 2-dimentional map of functional unique contamination clusters. Further analysis looked at how many unique clusters were needed to effectively characterize the pollution. The technique appears to be somewhat effective at differentiating between contaminated and uncontaminated areas. Results also suggest that not all microbial communities present the same opportunities for use as contamination proxies.

Although this is was a technical paper and a challenge to review the concept does make sense and the results appear to agree with the initial hypothesis. The work will undoubtedly be continued and expanded in the future. This paper does deserve to be published with a few minor revisions. There might be slight tweaks in the review if the target journal is more data/computational in nature versus one that is more applied in focus. Broadly, this paper could be strengthened by highlighting more clearly why this technique is potentially superior (ease of use, less financial cost, more accurate than other methods, etc.). At the same time the methods section is pretty dense and might more accessible to a wider audience if it included some more basic language or illustrations. The writing was style was technical but clear. The flow of the paper made sense. The following comments are by section:

Abstract: Good concise summary of the research. It might help to include the # of clusters the MANOVA results suggest as the ideal number.

Introduction: This section starts off very strong and includes what appear to be lots of good citations of relevant research. Since this work has such a spatial component, is there space to include some other forms of spatial analysis of ground water contamination(Geographically weighted regression)? The first section of SOM’s as an Environmental monitoring tool seems a bit out of place. Maybe just cite the original work and let curious readers find out the history. The next paragraph does a great job explaining why this technique is well suited to the study. The last two paragraphs are a well written explanation of the work being done (L127-128 might have a place in the abstract). The Goals and objectives are clear.

Methods: The first part of this section relies heavily on the work by Mouser. It does not leave a lot of information for the reader to grab on to about the study area. How was the "extent of contamination" that is mapped on figures 1 & 4 generated? The section that talks about PCA (L170-175) is a little hard to follow. Could this be described in simpler terms? The computational methods section is very dense. Can you use "sample well" in place of input to give the data more context? How is the data formatted? Does it look like a database? The training section could also use a bit less ANN jargon. What about a table of common terms and their definitions? The last section of the Methods (L236-244) is very well written and helps make sense of table 1.

Results and Discussion: How does the data space depicted in Figure 3 relate to the physical space of the landfill. Somehow tying this together will help clarify the results. The section on how the various microbial communities were analyzed could have been set up more in the methods. How was the Archaea determined to be "irrelevant"?

Conclusions and Implications: This is a fairly brief section. What are the implications for long term monitoring if this method is successful?

Figures: Figure 1 is solid, but how is the contamination defined? Figure two is confusing. Could you include a sample page of data and show how the computed analyzes it? Can there be any tie made between Figure 3 and the study site (North Arrow, ground water flow)? The group numbers in figure 4 need to be smaller (or the well sites need to be bigger). Table 1 is good, but can you place boxes around the groups for easier viewing? How about graphing the MANOVA covariance results? How was the Clean, Fringe, Polluted designation applied?

Andrea, great job and good luck editing/submitting.

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

By: Pearce et al.

This paper provided a clear explanation of how ANNs can be used to predict groundwater contamination based on microbial communities. Microorganisms are defined as adapting to the environment where they are, so changes in microbial communities reflect changes in the surrounding hydrochemistry. Parameters to measure contamination in groundwater from landfill leachate are complex and autocorrelated in space and time. Non-parametric statistics need to be used and Kohonen Self Organizing Maps (SOMs) are used for clustering spatial data in this study. DNA was isolated from 25 monitoring wells specific to Archaea, bacteria, and Geobactereae. Eight metrics were determined from the microbiological composition based on principle component analysis. An ANN was used to cluster nodes on a two dimensional map through an iterative process of finding weights for nodes based on distance from other nodes. Measures of cluster significance were used to determine the optimal map size. F-statistics were calculated from a non-parametric MANOVA to compare between the different numbers of clusters generated. The clean sampling locations were found to be homogeneous with a clear clustering location on the SOM output. The optimal number of clusters was based on achieving a large F-statistic, but too many clusters would represent a false increase in significance. Ultimately 4 clusters were considered significant and the clusters fit well with regions of different contamination. It was suspected that removing Archaea from the analysis would improve the validity because Archaea do not have as much metabolic activity in many of the sampling locations.

I think this paper flows very nicely, and after some revision it should be accepted to the Journal of Groundwater. I think that your introduction is a great progression through the different motivations for doing this research. Your first paragraph could maybe be more poignant if you started off with your sentence on line 41 because that is what underlies your analysis. Also, in the introduction I like how you describe the need for a non-parametric analysis, like an SOM (page 4). Are there particular motivations for using microbiological community structures in this kind of analysis beyond what you mention on page 3?

In the methods section the term units is introduced in line 212, I believe for the first time. Following, on page 10, there is a discussion on qe and te. Are these methods for determining the optimal size of a SOM widespread? I think that you could reduce this paragraph because you ultimately do not use this methodology. Then you can put more stress on the methodology that you do use for determining significant clusters.

I have provided some in-line comments on your annotated copy of the manuscript. Following are some additional specific points, which I thought of, for guidance as you edit your paper:

- On line 207 you describe the ordering-stage and the fine-tuning stage; why do you perform these two stages? For speed? Computational necessity with large data sets? And on this same note why is the SOM better/different then other similar statistical methods?
- On line 229 what do you mean by directly? When you have a large number of nodes are you not directly clustering the data?
- On page 11 in the Results and Discussion section, you mention over-fitting. I think that you could provide reason why that could be bad and what it means exactly.
- On page 12 how do you get to the conclusion that 4 clusters are optimal?

Overall I think the paper was very well written, and worthy of being published after some edits. I look forward to discussing it in class. Good luck.

Martin

Paper: 'Exploratory data analysis using a self – organizing map and MANOVA for environmental monitoring' by Pearce Andrea R., Paula J. Mouser, and Donna M. Rizzo

Reviewer: Nikos Fytilis – 04/15/09

In this manuscript, the authors present a methodology that uses microbial data from 25 groundwater wells to test the hypothesis that microbial community structure can distinguish between different regions of contamination. The basis of the method, which is data-driven, is the self – organizing map that performs cluster analysis. Optimizing the number of clusters using a non-parametric MANOVA suggests a number of clusters for interpolation. Applying this technique to a complex spatial dataset of microbial communities and geochemistry showed that the algorithm can successfully distinguish a gradient from clean to contaminated locations using the microbial community structure. So, based on the results, this screening tool, which uses non-linear and non-parametric algorithm, produces accurate spatial patterns of subsurface contamination. This approach is the first step forward using microorganisms to delineate spatial patterns of subsurface contamination.

The abstract is short, concise, well written and quickly describes your motivation. I have one comment that might help to improve your abstract. You don't mention or say anything about the study site in your abstract. You could say the area and provide general information about the contamination in the area. The introduction I thought is well written and provides a good overview of the research presented in this paper. One area I would to focus is the two first paragraphs which I think you should better to inverse them. The next two sections are well thought and ease to follow. I found these two sections very vital to this paper. There are some points that I couldn't understand probably because I am not very familiar with the subject (e.g. lines 118-119,133). At the end, I really liked the paragraph where you clarify the objectives of your research. Overall, I thought that since this paper is strongly correlated with another paper, these two sections bond the two papers nicely and they present a good material.

The methods section, in my opinion, was the part of this paper which I had many problems reading it. Maybe if I had the other paper and was more familiar with the subject, I could have understood more. Nevertheless, as far as I can judge, this subsection provides all the necessary info and the most crucial equations used by the algorithm to create the 2-dimensional map. I also believe that the figures referred to this section helped me get a clear overview of the contaminated area but Figure 2 didn't help me at all. I tried to connect all the variables of the equation mentioned in lines 193-199 but I couldn't. What is the last term inside the parenthesis and k is always superscript? Where can I find the last term inside the parenthesis in Figure 2? My suggestion is to put in line 184 that input parameters are noted in this paper as (x). I know it is a tiny detail but I have seen in other papers other notations and it would be better to clarify that. In addition, while I was reading the last paragraph on page 9, I didn't find the number of iteration you used. You perfectly describe the two training phases but you don't mention the number of iterations needed for this research. One other point I want to make is about the lower case n you use in line 223. What does this stand for? Finally, I believe you did a great job describing the MANOVA and how this tool is useful to optimize the number of clusters created by the SOM.

I don't know why you chose these three specific microorganisms and if there are other microorganisms you could use. From your results they seem to support well your research but you could site a paper showing that these are the most common used microorganisms used in similar studies. I believe that your figures are good (I mentioned my problem with Figure 2) but I have one other question. In Figure 1 what does the dashed line show? I believe that this paper should be published with minor revisions. Please refer to my hard-copy edited version of the manuscript for small comments regarding structure and rhetoric. Good luck (and for your presentation on Friday!!!).