Cluster analysis of geophysical field data: An approach for reasonable partitioning of sites

Daniel Altdorff[A] & **Peter Dietrich[A]**,

[A] UFZ Leipzig, Departement Monitoring- and Exploration Technologies, Permoserstraße 15, 04318 Leipzig, daniel.altdorff@ufz.de

Abstract
For land management options, a sharp partitioning of areas is suitable. However, applied soil science usually provides integrated data and different information from identical sites and does not supply a partitioning in respect to all variables. In this paper, we describe a way for deriving reasonable partitioning of sites from multidimensional data sets by an integrated statistical approach: the clustering of different discrete field data. For this purpose, two separate work packages (WP) are suggested. In the first WP, sets of different data are standardize and discretize in order to get collocated data on predefined locations This step is precondition for further processing of multi-dimensional statistical analysis. The second WP leads from the different sets of collocated data to a map of partitions representing different soil units. Using an example of field data, we could show that cluster analysis works as a useful tool for site partitioning.

Key words
Rapid soil survey, geophysical approaches, soil moitrure, bulk density, electromagnetic induction.

Introduction

Investigation of soil properties will become increasingly important in the future (e.g. Lin 2003, Hartemink and McBratney 2008). In particular, assessment of hydrological soil qualities, such as water content and hydraulic conductivity is essential to define appropriate strategies for sustainable land and water resource management. However, direct data acquisition of these relevant properties by point measurements is complicated and costly. Hence, several possibilities to acquire relevant information of soil properties have been develop in the last decades. An established and accepted way is the derivation of subsurface details from geophysical data, for example apparent electric conductivity, georadar, and gamma ray radiation (e.g. Hayley *et al.* 2007, Weller *et al.* 2007, Beckett 2008, Mendez *et al.* 2009). The main benefit of these indirect methods is the ability to measure soil parameters quickly and in a non-invasive manner. This is a great advantage compared to conventional pedological data acquisition. Regrettably, in respect to the methods these geophysical measuring techniques provide integrated data and different information for the same site and do not reflect the entire area. In the praxis of applied soil science however, often a sharp partitioning of areas is required for land management measures. Hence, one challenge in land use management is the definition of a reasonable partitions from multidimensional data sets. In this paper, we will describe a transferable approach for sharp site partitioning - independently from number and amount of the available data sets.

Proposed Work Flow

Regarding the different data sources and soil parameter databases, a standardized program for site characterization is required. We concern the problem by an integrated statistical approach – the clustering of different discrete geophysical data (Dietrich *et al.* 1995, Dietrich and Tronicke 2009). The aim of this procedure is the partitioning of test sites according to their natural characteristics. With respect to the different established measurement methods and their corresponding data sets, we describe our method first schematically to make the approach transferrable to other data sources. Afterwards, we use geophysical field data as an example for the applicability of the method. The workflow is divided into two separate work packages (mainstays), one for the preparation and discretization of the field data, and another one for statistical approaches and clustering. The first work package contains the following operational sequences: (i) interpolation from point data, (ii) conversion of map data into standardized discrete grid data, (iii) blanking the maps and generation of ASCII files (see Figure 1). The second work package comprises the (i) consideration of data relationships (ii) selection of the appropriate variables, and (iii) the cluster analyses. As example for illustrating this approach, we used electromagnetic induction (EMI) data sets from a test site in Graswang / Bavaria, Germany obtained by a near surface survey. Four different data sets from different depths (integral up to 0.75 m, 1.50 m 3 m, and 6 m below surface) were exploited to conduct site

partitioning. These data sets contain values from apparent electric conductivity measured with the instruments GEONICS EM38DD and EM31 (Geonics Limited, Mississauga, Ontario Canada) in horizontal and vertical configurations and the related spatial information, (GPS coordinates northing and easting). At step i) a variogram from each of these data sets was plotted and a function was fitted with respect to the spatial dependency of the data in the field; a typical view of an unconditioned data set and its corresponding variogram is shown in Figures 1a and 1b using Surfer 8 (Golden Software).

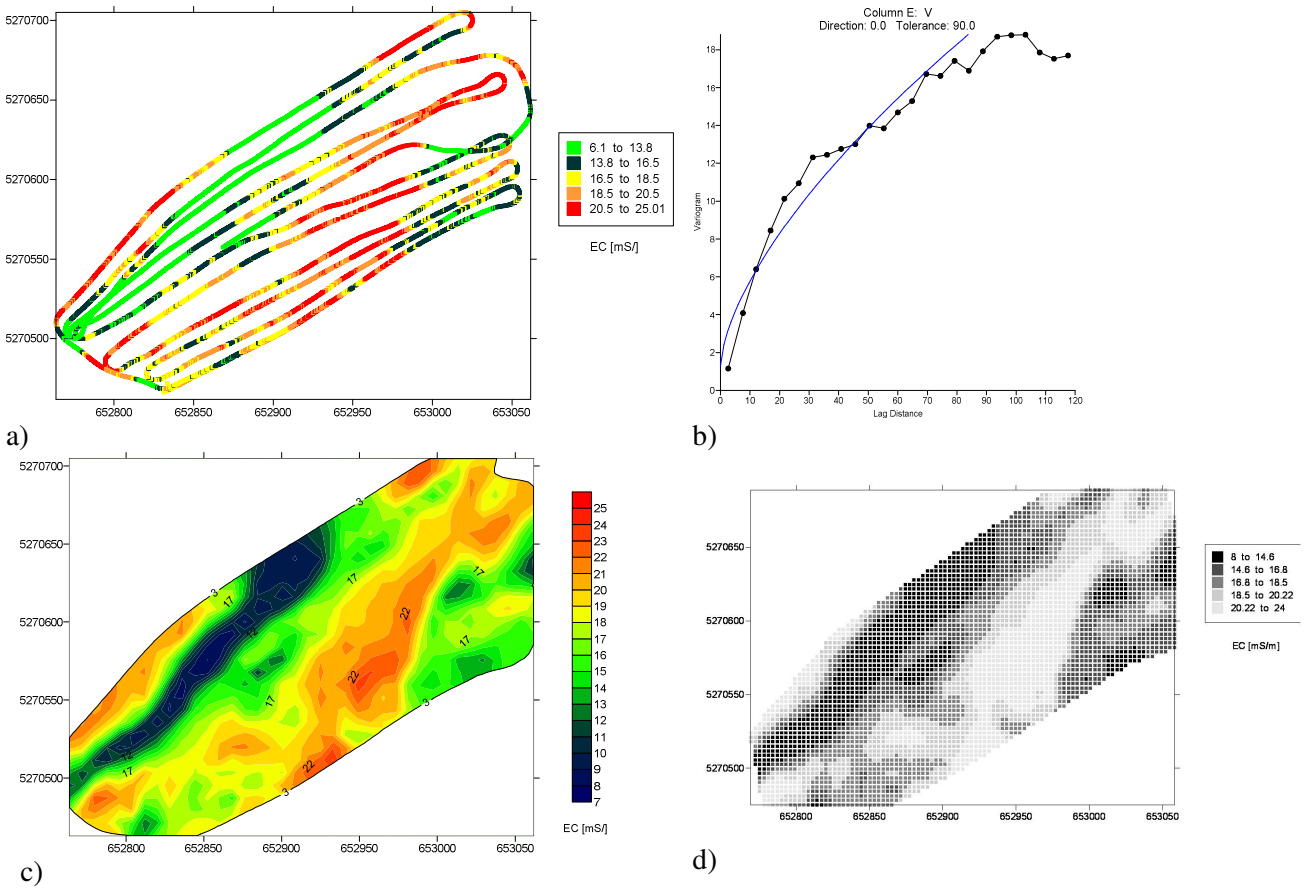
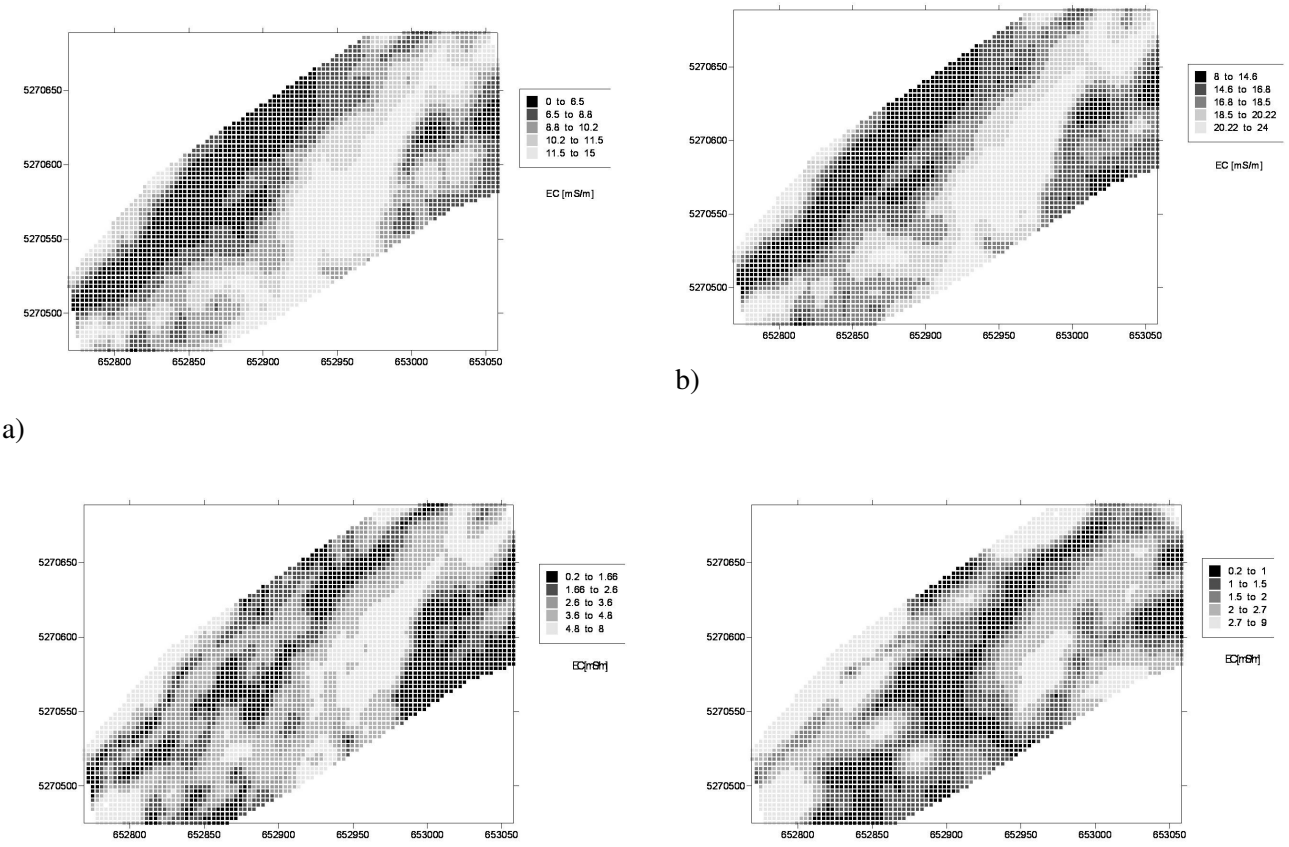


Figure 1. Work package A: a) plot of a raw EMI data set, in this case EM38V, b) corresponding variogram, c) interpolated map, d) discrete point values as precondition for cluster analysis

Using the variograms for each data set, n- interpolated maps could be generated to fill the data gaps between point values (in this case four maps). Step ii): after interpolation, one of the maps should be selected as matrix for discretization and blanking; usually the one with the smallest spatial expansion. In respect to the standard deviation of the interpolation, the map was blanked from uncertain data, in our case in the angle of the orthogonal map. Now the blanked map was discretized using the "mosaic" function of Surfer by choosing a convenient grid size and saved as ASCII file. The blanked data appears within the table as unrealistically high values – an imported attribute for selecting the excessive data from the other maps. Without blanking, all other maps were processed like the matrix map with the same grid coordinates and saved as ASCII files. Now all data sets have the same number of rows with exactly the same coordinates. After merging all ASCII files to one collective ASCII table, the excessive data could be deleted simply by sorting the unrealistically high values out from matrix table. At this point a discrete ASCII file with four different variables is prepared, ready for multivariate statistical analysis and also visible as classified post map (see Figure 2).



a) b)

c) d)

Figure 2: Visualization of four EC data sets with same grid coordinates – result of work package 1: a) EM38 horizontal mode, b) EM38 vertical mode, c) EM31 horizontal mode, d) EM31 vertical mode

The final aim of the next work package - statistical analysis and clustering - is to define different areas of the test site in respect to their properties by means of cluster analysis. Precondition for building useful clusters is the independency of the variables. Therefore, the next step of our approach is the comparison of the variables regarding their dependency. In this study, we apply a classical factor analysis using the statistical program SYSTAT 12. If the factor analysis detects a strong dependency between variables, it is recommended to use only one of them for clustering. In our example, only one of the EM38 variables is affected due to the similarity of the EM38H and EM38V data. Nevertheless, for demonstration in this study, both values were used for further processing. After selecting independent variables, the proper cluster algorithms were run. A cluster analysis is the assignment of a data set into subsets (*clusters*) so that data properties within the same cluster are similar. This method allowed the multidimensional comparison of data points and their classification – a well-established tool in social science, but not much applied in geosciences. Clustering can either be hierarchical or partitional (K-means clustering - partition n observations into k clusters). In this study, we use the K-means clustering only. In a K-means algorithm, each data point is compared with its proximate neighbor related to similarities and builds with its co-natural neighbor the first cluster group. This operation proceeds until the desired number of clusters is reached. The algorithm steps are: (i) choosing the number of clusters - *k*, (ii) randomly generating *k* clusters and determining the cluster centers, (iii) assigning each point to the nearest cluster center, (iv) recomputing the new cluster centers, (v) repeating the two previous steps until some convergence criterion is met - usually until the assignment hasn't changed (J. MacQueen, 1967). Again, we used the statistical program SYSTAT 12 and ran a K-mean cluster algorithm with four variables to three and four reasonable groups, respectively. The result was a new row in the collective ASCII table, the cluster each coordinate pair belongs to, again visible as classified post map (see Figure 3).

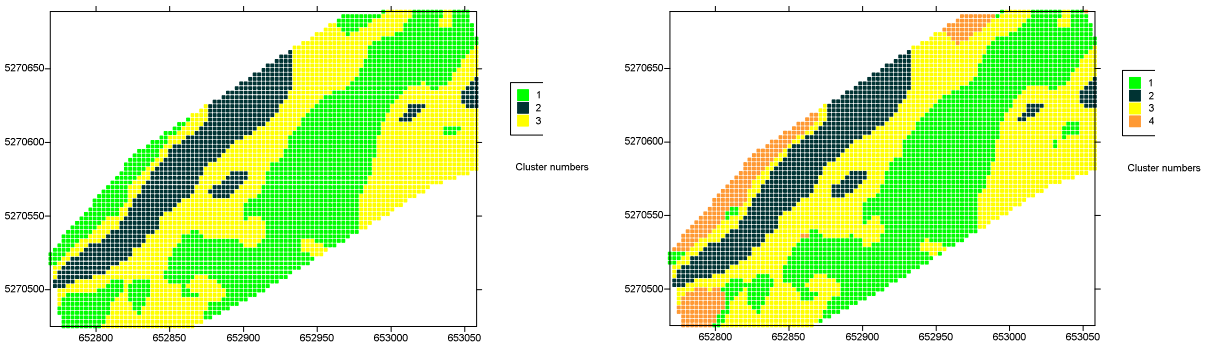


Figure 3 Results of cluster analysis: a) clustering into three partitions, b) clustering into four partitions; precondition for partitioning of sites by clustering are discrete and standardized soil parameters

Conclusion

For land management options, a sharp partitioning of areas from available multidimensional data sets is required. Clustering of field data is an appropriate tool for sharp partitioning of sites - independently from type and character of the database. Precondition is the existence of discrete and standardized parameters at predefined locations. Following the method used here, a direct and independent statistical comparison of properties and their partitioning of the investigated area is possible.

References

Beckett K (2008) 'Multispectral processing of high-resolution radiometric data for soil mapping' *Near Surface Geophysics* **6**(5) 281-287.

Dietrich P, Tronicke J (2009) Integrated analysis and interpretation of crosshole P- and S-wave tomograms: a case study. *Near Surface Geophysics* **7** (2), 101-109.

Dietrich P, Fechner T, Whittaker J, Teutsch G (1998) An Integrated Hydrogeophysical Approach to Subsurface Characterization. In 'Groundwater Quality: Remediation and Protection'. (Eds M Herbert, K Kovar) pp.513-520. (IAHS Publication)

Hartemink AE, McBratney A (2008) A soil science renaissance. *Geoderma* **148,** 123-129.

Hayley K, Bentley LR, Gharibi M, Nightingale M (2007) Low temperature dependence of electrical resistivity: Implications for near surface geophysical monitoring. *Geophys. Res. Lett.* **34,** L18402, doi:10.1029/2007GL031124.

Lin H (2003) Hydropedology: Bridging Disciplines, Scales, and Data. *Vadose Zone Journal* **2,** 1–11.

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', Vol. 1 (Eds LM Le Cam, J Neyman) pp. 281-297. (University of California Press, Berkeley).

Mendez-Barroso LA, Vivoni ER, Watts C, Rodriguez JC (2009) Seasonal and interannual relations between precipitation, surface soil moisture and vegetation dynamics in the North American monsoon region. *Journal of hydrology* **377**(1-2), 59-70.

Weller U, Zipprich M, Sommer M, Zu Castell W, Wehrhan M (2007) Mapping Clay Content across Boundaries at the Landscape Scale with Electromagnetic Induction. *Soil Sci. Soc. Am. J.* **71,** 1740–1747.