A multivariate approach to site selection for comparative soil studies

S.C. Löhr[A,B], J. Hodgkinson[C] and S. Fraser[C]

[A] Discipline of Biogeosciences, Queensland University of Technology, Brisbane, QLD, Australia, Email s.loehr@qut.edu.au
[B] Institute for Sustainable Resources, Queensland University of Technology, Brisbane, QLD, Australia.
[C] CSIRO Exploration and Mining, Queensland Centre for Advanced Technologies, Pullenvale, QLD, Australia.

Abstract

Careful site selection is of great importance in comparative soil studies. Any conclusions based on the comparison of sites which have been subject to different genetic histories and processes are likely to be erroneous. This article proposes a multivariate site selection method based on Kohonen's self-organising maps. The method is designed to group similar samples and so permit the identification of sites that are suitable for comparative studies. A case study is presented to illustrate the method.

Key Words

Site selection, comparative study, multivariate, self-organising maps, soil water.

Introduction

An experimental approach can be effectively used to distinguish between causal and correlative relationships in the soil environment. The value of employing controlled experiments to study soil-forming processes has recently been reiterated by Bockheim and Gennadiyev (2009). In many cases, however, such an approach is impractical, primarily because of the long time-scales involved in most soil-forming processes. Where this is the case, comparative studies can provide similar information, provided that the sites to be compared are carefully selected. Examples of studies in which a comparative approach has been employed are common. They range from the use of chronosequences to determine the effect of soil age on a range of soil properties (Calero *et al.* 2009) to the comparison of soils formed in different parent materials to determine the effect of parent materials on soil organic carbon dynamics (Heckman *et al.* 2009). Indeed, much of our understanding of soil processes is derived from studies that employ a comparative rather than an experimental approach.

Site selection for comparative studies

Site selection is of great importance in comparative studies. This is because the soils need to be sufficiently similar, in terms of their genetic history and the processes currently operating within them, for their comparison to be valid. Ideally, the sites selected for comparative purposes will be identical, apart from the differences that can be attributed to the process or factor being studied (i.e. age, parent material, vegetation, fire history, landscape position, and climate). Given that no two soils are identical, this ideal situation will never be attained. Nevertheless, the conclusions based on the comparison of sites subject to different genetic histories and processes are likely to be erroneous. Thus, it is of great importance to select sites that are suitable for comparative studies. Unfortunately, this is often not tested, or only considered in a purely qualitative manner.

New site selection method

This work proposes a multivariate site selection method for comparative studies, based on the self-organising maps (SOM) data analysis method (Kohonen 2001). This method is designed to ensure that the selected sites are suitable for comparison. In other words, the method can be used to identify sites that are sufficiently similar so that their comparison is valid.

Case study

We present a case study in which we employ the method to select sites for a study on the effect of vegetation type on soil water composition. The study area lies within a coastal catchment in South-East Queensland (Australia), and extensive parts of the catchment have been cleared of native vegetation to establish exotic pine plantation since the 1950's. The study investigates the effect of vegetation (pine plantation *vs* remnant native forest) on the speciation of iron (Fe) in soil water. Soil water is collected *in situ*, using MacroRhizon (Eijkelcamp Agrisearch) micro-lysimeters installed at 50 cm depth. Thus, similar soil types are required for micro-lysimeter installation at the comparison sites.

Methods

Self organising maps

A detailed explanation of self organising maps is outside of the scope of this short contribution and we refer the readers to Kohonen (Kohonen 2001), or Astel *et al*. (2007) for an example of SOM analysis applied to a large environmental dataset. In brief, the self organising map can be considered a data visualisation and analysis tool. Although it is usually considered an exploratory tool, the method can be used to perform function fitting, prediction or estimation, clustering, pattern recognition and classification (Fraser and Dickson 2007). In this paper we employ SOM as a clustering tool because it allows the mixing of continuous and categorical input data types. In addition, samples with missing data values can be included in the analysis. In a previous study we have shown that SOM is able to group samples that have been affected by the same pedogenic processes, using only easily obtained data such as physical and chemical properties of soils and their topographic position (Löhr *et al.* 2010).

Selection of site variables

In this study we use a subset of the data from our previous study to identify three paired sites (six in total) for a comparative study of the effect of vegetation on the speciation of Fe in soil water samples.

In spite of the uniform geology and climate at all potential comparison sites, the sites are found at different landscape positions and within different soil types. The soils are characterised by distinct chemical and physical properties. Exotic pine plantations have been established relatively recently and are not expected to have affected bulk soil properties. Thus, it was deemed appropriate to select comparison sites based on similarities in landscape position and bulk soil properties (chemical and physical). Accordingly, the variables listed in Table 1 were included in the SOM analysis. The variable 'vegetation type' is excluded in order to permit clustering of similar sites with differing vegetation.

Table 1. Variables included in SOM analysis

Terrain	Gamma-ray spectrometry (GRS)	Soil chemistry	Soil physical properties	Clay mineralogy
Elevation, Slope, Curvature, Topographic Wetness Index	U, Th	1 M HCl-extractable: Na, Mg, Al, K, Ca, Mn, Fe and Zn	Organic carbon, pH, EC, Fe-concretion, clay fraction	Kaolinite, Vermiculite, Illite, Illite-Smectite

Results

After SOM analysis, the 120 samples were assigned to 56 best matching units (BMU's or clusters); between 1 and 6 samples were assigned to each cluster (by the SOM process). Paired sites for the comparative soil moisture study were selected from these and are shown in Table 2. Criteria for paired site selection were a) high overall similarity, as expressed by allocation to the same cluster and low q-error (a measure of the 'distance' of a sample from the cluster centroid) and b) different associated vegetation types (exotic pine plantation *vs* native vegetation). In order to independently verify that the selected sites are sufficiently similar for comparative purposes, the soil type at these locations was classified according to the Australian Soil Classification (Isbell 2002).

Discussion

Recent work has shown that the SOM approach is able to identify and group sites at which similar pedogenic and geochemical processes are operating (Löhr *et al.* 2010). These groupings remained meaningful even when the initial clusters (the best match units) were aggregated into larger groupings using k-means clustering. In the method proposed here, samples are not aggregated into large clusters, but retained in a greater number of small clusters. Thus, samples with moderate similarity remain in separate groups, helping to ensure that the paired samples display high overall similarity.

Nevertheless, the results demonstrate that the paired sites are not identical (Table 2). Sites 30 and 32, for instance, have substantially different extractable Fe and Mg concentrations. They are classified as different soil types and are likely to have formed as a result of different genetic processes. Sites 58 and 85, on the other hand, are classified as the same soil type. Although the sites have different slopes and an extractable Fe concentration that differs by a factor of two, their overall similarity cannot be disputed. The same is true of both sites in pair C. While these soils are classed as different soil types (mostly due to a greater degree of pedogenesis in site 21), the differences are minor.

Clearly, the successful use of the proposed site selection method depends on the data used to compare potential comparison sites. Apart from the terrain attributes, the data used here to select sites are based on analysis of the top 30 cm of soil only. This has resulted in paired sites at which the upper soil horizons are indeed similar. However, the use of exclusively surficial data is not sufficient to ensure selection of similar locations in all instances, as shown by site pair A. We suggest that the performance of the method in the case study can be improved by incorporating data of soil properties at greater depths into the analysis. Ideally, the comparison sites can be selected from the potential comparison sites grouped within a cluster after field comparison of these sites.

Employing a quantitative, data-driven approach for site selection has a number of advantages. In addition to minimising the possibility of invalid comparisons due to unsuitable site selection, the researcher can gain an understanding of the variability of the soil properties in the study area by comparing a number of potential sites using quantitative data. It is therefore possible to conduct a comparative study and include a consideration of the effects of subtle differences between the comparison sites, as well as spatial variability more generally.

Table 2. Properties of paired sites selected from 120 samples clustered into groups of similar sampling locations using SOM

Site parameter	PAIR A		PAIR B		PAIR C	
Site ID	30	32	58	85	21	41
BMU	16	16	20	20	55	55
q-error	1.6	1.2	2.7	2.4	1.4	2.0
Vegetation –plantation	Yes	No	Yes	No	Yes	No
Vegetation – native	No	Yes	No	Yes	No	Yes
Terrain curvature	0	-0.17	0	-0.31	0	0
Slope (%)	2.5	6.1	3.1	12.4	3.6	4.4
Elevation	31	17	51	31	31	28
TWI	10.6	9.7	9.1	8.3	10.0	8.4
K	7.18	3.61	14.59	18.16	10.18	16.16
Al	286	360	480	300	114	138
Ca	19.9	9.2	69.9	31.9	25.9	81.7
Fe	418.8	106.2	680	359.2	61.9	111.7
Mg	69.8	14.8	32	71.8	65.9	97.8
Mn	<0.4	<0.4	0.4	5.0	<0.4	0.4
Na	<4.4	<4.4	8.9	<4.4	10.4	14.0
Zn	<0.5	<0.5	6.4	4.19	<0.5	<0.5
Th (GRS)	1.7	1.9	2.6	2.9	2.0	2.3
U (GRS)	0.8	0.7	0.9	0.7	0.7	0.6
Clay fraction (%)	11.1	9.2	11.2	10.5	9.3	5.8
Vermiculite	2	1	3	2	0	0
Illite/Smectite	0	0	1	1	0	0
Illite	0	0	0	1	0	0
Kaolinite	1	1	3	1	0	0
Fe concretions (%)	1.4	1.3	0.4	1.5	5.4	0
pH	5.32	5.10	5.13	5.20	4.25	4.16
EC	6.4	9.6	18.3	10.5	32.1	33.3
Loss on ignition (%)	1.52	1.31	1.72	1.17	3.12	3.39
Soil type (Australian Soil Classification)	Mottled, Grey Kurosol	Humosesquic, Aeric Podosol	Mottled, Brown Kandosol	Mottled, Brown Kandosol	Humosesquic, Aeric Podosol	Acidic, Arenic Rudosol

Conclusion

We propose a multivariate approach to site selection for comparative soil studies. The method is based on the SOM data analysis method and is designed to ensure that selected sites are sufficiently similar so that their comparison is valid. Careful selection of the input variables is essential in order to a) exclude soil properties that may have been affected by the soil process of interest and b) include sufficient data to avoid clustering of sites which are dissimilar.

The case study showed the successful selection of two sets of comparison sites out of a total of 120 candidate locations. A third selected site pair were significantly different, and illustrated the importance of field validation of comparison sites. Nevertheless, a process-sensitive method such as the self-organising maps can prove a robust means of selecting locations suitable for comparative studies.

References

Astel A, Tsakovski S, Barbieri P, Simeonov V (2007) Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research* **41**, 4566-4578.

Bockheim JG, Gennadiyev AN (2009) The value of controlled experiments in studying soil-forming processes: A review. *Geoderma* **152**, 208-217.

Calero J, Delgado R, Delgado G, Martín-García JM (2009) SEM image analysis in the study of a soil chronosequence on fluvial terraces of the middle Guadalquivir (southern Spain). *European Journal of Soil Science* **60**, 465-480.

Fraser SJ, Dickson BL (2007) A new method for data integration and integrated data interpretation: self-organizing maps. In 'Exploration 07: Fifth Decennial International Conference on Mineral Exploration'. (Ed B Milkereit) pp. 907-910.

Heckman K, Welty-Bernard A, Rasmussen C, Schwartz E (2009) Geologic controls of soil carbon cycling and microbial dynamics in temperate conifer forests. *Chemical Geology* **267**, 12-23.

Kohonen T (2001) 'Self-Organizing Maps'. (Springer: Berlin)

Löhr SC, Grigorescu M, Hodgkinson JJ., Cox ME, Fraser SJ (2010). Iron occurrence in soils and sediments of a coastal catchment. A multivariate approach using self organising maps. *Geoderma*, doi:10.1016/j.geoderma.2010.02.025