

Statistics and spatial heterogeneity

R.M. Lark

Centre for Mathematical and Computational Biology
Rothamsted Research, Harpenden, Herts AL5 2JQ

Novel micro-scale techniques to collect data on soil offer an exciting opportunity to study important soil processes at the spatial scales of key agents (e.g. microbes) and structural features (e.g. pores). However, these techniques also present us with challenges if their potential is to be fully realized. The aim of this paper is to outline these challenges, and some possible solutions, and to illustrate with some recent results from collaborative studies at Rothamsted.

First, it is useful to recall how micro-scale soil observations differ from those that we commonly make on soil properties in conventional soil sampling and analysis.

1. A micro-scale soil data set is typically an exhaustive set of observations on a particular *sample unit* (which may be an aggregate, a section, an exposed surface etc.) For example, if we scan an aggregate by XRMT, then we have data corresponding to any point in the aggregate. This is by contrast to conventional soil data sets which are typically observations on samples (e.g. cores) drawn from a wider domain (e.g. a field).
2. A micro-scale soil datum is generally obtained by an imaging process (or analogue) such that it corresponds to a discrete unit of support (e.g. voxel), but may be correlated with neighbouring data due to convolution effects of imaging and any subsequent post-processing (in addition to any actual correlated variation in the underlying properties of the soil). By contrast a classical soil sample is defined on a support (such as a core) but this is generally simple and easy to characterize.
3. In most cases there is a large gap between the scale of the domain of a micro-scale soil data set and the regions with which we are ultimately concerned. Micro-scale soil measurement is ultimately justified if it permits better management of the soil at field, catchment or landscape scale. By contrast the scale of a classical soil sample is almost always the domain of interest (scale is not to be confused with support, a set of cores from random locations across a field is a field-scale data set).

As a consequence of these considerations:

- i. The analysis of data on a particular sample unit will (usually) not be concerned with *estimation*, but with adequate **characterization** of the structure of the variation within the unit so as to predict or explain soil behaviour at the scale of the unit, or larger.
- ii. Data sets will typically be **large**, and systematically structured (grids or arrays). This offers an opportunity to use some sophisticated methods to analyse **complex non-stationary spatial variation over multiple scales**.
- iii. Appropriate **sampling strategies** are needed so that micro-scale measurements can be effectively deployed to address problems in soil management. Because of considerations under (1) above, sampling problems will generally be concerned

with how to select sample units for (for example) scanning from the wider domain of interest, but some micro-scale measurement technologies (e.g. probes, laser ablation) may present sampling problems for their direct deployment.

- iv. The analysis of micro-scale data may offer **particular technical challenges** (such as denoising, thresholding or deconvolution).

Let z_u be a soil property measured at micro-scale on an array of n locations $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^T$ where \mathbf{x}_i contains the co-ordinates in d -dimensions of the i th observation (e.g. a particular voxel). If we want to predict the aggregate behaviour of the soil at the scale of the whole array, then the mean value of z_u over this array is adequate information only when the behaviour depends linearly on z_u . This is why, for example, some workers have computed variograms of micro-scale soil data to characterize their spatial correlation structure. However, the variogram only completely characterizes the spatial-dependent variation of z_u if the data can be regarded as a realization of a multivariate normal random variable $Z_u(\mathbf{X})$ which has a stationary distribution. This is a strong assumption. I shall therefore discuss

- A. How the spatial variation of soil data might be **better characterized**, using approaches such as alternative spatial models, or multiple point geostatistics. I will discuss the challenges that implementing these ideas will pose.
- B. How **non-stationarity in soil variables**, over different spatial scales, can be analysed by wavelet transforms. This will be illustrated with analyses of non-stationary covariation of different forms of carbon in soil.

The challenge of deploying micro-scale soil measurements to address questions about soil behaviour and management at field to landscape scale will also be discussed. Specifically:

- C. I shall show how **appropriate spatial sampling schemes** can be used to investigate soil variation over contrasting scales by nested methods of analysis, and discuss how this approach might be implemented for structured sampling of soil for micro-scale measurements.

In addition, I shall discuss how :

- D. Wavelet analysis methods might be used to tackle some of the technical problems in analysis of micro-scale data, such as the identification of underlying structure in noisy imagery
- E. We can avoid some of the inferential problems ('data mining') that have recently caused difficulties in other areas of science that exploit technology to generate large data sets (specifically neurological studies on MRI brain scans).