Analyse problems, not data: de-classifying spatial statistical methods

Peter J Diggle

(Faculty of Health and Medicine, Lancaster University, UK and Institute of Infection and Global Health, University of Liverpool, UK)

Joint work with Paula Moraga, Barry Rowlingson and Ben Taylor

Spatial statistical methods are now used routinely in a variety of scientific settings. Following Cressie (1991), a widely used classification of spatial statistical methods is in terms of data-formats as follows:

1. *geostatistical data* are responses Y_i associated with the values of a latent, spatially continuous stochastic process $S(x) : x \in A \subset \mathbb{R}^2$ at sample locations $x_i \in A$;

2. *lattice data* are responses Y_i associated with a pre-specified set of nominal locations x_i;

3. *point pattern data* are locations $x_i \in A$ that form a (partial) realisation of a stochastic point process.

The above classification based on data-formats is undeniably useful for developing and teaching statistical theory and methods. However, and especially when we extend our perspective from the purely spatial to the spatio-temporal, it can lead to an unhelpful detachment of statistical analysis strategies from the underlying scientific processes and questions that generated the data in the first place.

In the remainder of the talk, I will focus on applications of spatial and spatio-temporal statistical methods in epidemiology. In this context, each of the basic data-formats is usually supplemented by information on one or more spatially referenced covariates, or risk-factors. However, different risk-factors are typically measured at different, and potentially incommensurate, spatial resolutions. For example, in the UK social deprivation is defined at the level of census enumeration districts, some environmental characteristics such as land-use or elevation are available as gridded images, others such as weather data are recorded on a spatially discrete network of permanent monitoring sites, whilst residential locations are geo-coded as notional points with a resolution of 100 metres or less in urban settings, one or two kilometres in rural settings.

I will describe current work on using spatially continuous stochastic models as a framework for disease risk mapping taking account of multiple and diverse data-sources (Diggle, Moraga, Rowlingson and Taylor, 2013). I will argue that disease risk maps should be presented as predictive probability maps rather than as point-wise estimates and standard errors and that, in this context, predictive uncertainty typically dominates parameter uncertainty. One implication of this is that, whilst the use of Bayes' Theorem is crucial, the distinction between maximum likelihood and Bayesian parameter estimation may be relatively unimportant.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* New York : Wiley.

Diggle, P.J., Moraga, P., Rowlingson, B. and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science* (to appear)