Statistical inference and local spatial modelling

Living and working with uncertainty

Chris Brunsdon
Department of Geography,
University of Leicester,
Leicester, United Kingdom,
cb179@le.ac.uk

Abstract—**Uncertainty exists in all forms of spatial process modelling. Errors exist in data, the correct model form for the process is often uncertain and the actual process being modeled is itself random. This all suggests that there is a need to handle uncertainty in the modeling and data analysis process. One emerging approach to inference in this framework is that of Bayesian inference based on Monte Carlo Markov Chain simulations. Until the advent of these techniques, Bayesian approaches, although of theoretical interest were often computationally impractical. The simulation approach overcomes many of these problems and allows great flexibility in terms of questions that may be addressed. Here these methods will be reviewed, and in particular their application to spatial data analysis will be discussed, together with examples.**

Keywords: Bayesian data analysis; Monte Carlo Markov Chain;Geographically Varying Coefficients;

I. A Brief Review Of Bayesian Inference

Although Bayesian approaches have existed for a very long time, their use has greatly increased in recent years, largely due to the Monte Carlo Markov Chain (MCMC) approach, outlined for example in Gelfand and Smith (1990). Very briefly, if D is a set of data, and M a set of parameters describing a model then Bayes theorem states that

$$P(M \mid D) = \frac{P(D \mid M)P(M)}{P(D)} \qquad (1)$$

where P(a|b) denotes the conditional probability of *a* given *b*. The set of all possible values of **M** in a study constitutes the range of models under consideration. Essentially equation (1) provides a means of linking the probability of observing the data given a particular model, P(**D**|**M**), with the probability of that particular model **M** being the correct one given the data – this is the inferential step. The quantity P(**M**) is the probability that the particular model holds, prior to considering the data. This can be thought of as representing the beliefs of the analyst before the study has taken place – and can be thought of as a subjective probability. However, to maintain objectivity typically the probabilities of all possible **M**s are set to be equal. The quantity P(**D**) is the marginal probability of obtaining the observed data given any possible **M**.

This last term has often proved problematic, as **M** is often defined using continuous parameters, and computing the probability of **D** over all possible **M** values requires a multivariate integral to be evaluated, which is frequently impossible analytically. However, various random number simulation techniques (such as the Metropolis-Hastings algorithm (Metropolis and Ulam, 1949; Hastings, 1970) allow draws from the distribution of **M** to be simulated without needing to compute this integral.

This has made Bayesian analysis feasible in a number of situations that were previously impractical – and as a result of this arguable brought Bayesian analysis into the mainstream of current statistical practice. Two notable advantages of this framework are 1) flexibility in the specification of **M**; and 2) flexibility in hypotheses being tested. The first advantage arises from the fact that provided all **M**s under consideration yield a valid distributional form for P(**M**|**D**) it is generally possible to simulate draws from the distribution, regardless of whether its analytical properties are known. The second advantages can be seen by noting that hypothesis here relate to statements about the values of individual parameters of **M** and functions of these parameters. Estimating the posterior probabilities of these hypotheses is a matter of computing the proportion of simulations for which these statements are true.

II. Applications To Spatial Data Analysis

Having outlined the Bayesian approach in general, its application to the analysis of spatial data will now be considered – and in particular its use in the analysis of local effects in spatial models. As an example, spatially varying regression models will be considered. Here models will take the form

$$y_i = \beta_{0i} + \beta_{1i}x_i + \varepsilon_i \qquad (2)$$

where *i* indexes a set of geographical locations. Note that regression coefficients differ for each location – and that in this case **M** is the set $\{\beta_{0i}, \beta_{1i}; i=1,n\}$. If we expect the coefficient values to have some degree of spatial similarity, an appropriate prior model might be to model the β_{0i}'s as a stationary, isotropic spatial process with a given semivariogram – for example an exponential variogram model.

$$E\left(\left[\beta_{0i} - \beta_{0j}\right]^2\right) = b\left[1 - \exp\left(d_{ij}/a\right)\right] \qquad (3)$$

A similar prior distribution can be provided for the β_i's. Here d_{ij} is the distance between locations i and j and a and b can be specified, or themselves could be added to the parameters in **M**. Using MCMC it is possible to obtain estimates of the local parameters, in the manner of geographically weighted regression (GWR) (Brunsdon, Fotheringham and Charlton, 1996).

The advantages of this approach is that reliable estimates of the variance of the parameter estimates can be estimated from the simulations, as well as estimates of functions of the parameters. These functions can be quite general – for example an indicator function as to whether a given local coefficient is a local maximum – and posterior distributions for such functions can then be used to assess hypotheses.

In the talk a number of examples of Bayesian approaches to modeling spatial variability will be demonstrated, using the above model framework as well as some others. In addition to this suggestions for software for carrying out these kinds of simulation will be outlined.

REFERENCES

Brunsdon, C., Fotheringham, A.S. and Charlton, M.E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis.* 28(4), 281-298

Gelfand, A. E., and A. F. M. Smith. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association.* 85, 398–409

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika.* 57, 97–109.

Metropolis, N., and S. Ulam. (1949). The Monte Carlo method. *Journal of the American Statistical Association.* 44, 335–341.