Discussion of the paper by J.M. Ver Hoef and E.E. Peterson

Sujit K. Sahu

School of Mathematics,

University of Southampton

1 Introduction

The flow of water in streams and rivers poses a unique problem in defining association between under water monitored quantities at any two sites. The usual methods of using Matérn covariance functions for the random quantities measured above water do not work since the flow of the water and movement of creatures such as fish in both upstream and downstream directions must be allowed to influence the association appropriately. This very original and impressive paper develops and illustrates some new moving average models for stream networks. New covariance models are presented based on the stream distance rather than the Euclidean distance. We begin our discussion by raising some questions on the models developed. Finally, we conclude by considering possible extensions of the variance component models to other inferential settings.

2 Variance Component Models

The paper cleverly constructs a variance component model corresponding to Equation (10):

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\mathbf{T}\boldsymbol{\beta} + \mathbf{z_u}(\mathbf{s}) + \mathbf{z_d}(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1}$$

where $z_u(\mathbf{s})$ and $z_d(\mathbf{s})$ are underlying independent tail-up and tail-down processes with previously defined covariance functions $C_u(\cdot)$ and $C_d(\cdot)$, respectively, and $\mathbf{x}(\mathbf{s})$ are location specific covariate values. Below Equation (10) authors also discuss the possibility of adding another component accounting for unmeasured covariates, $z_o(\mathbf{s})$ say, (where the suffix o stands for omni-directional) that could be related due to underlying bedrock characteristics. This additional component can also serve many other purposes, for example, we may want to model characteristics of connected streams, rivers and lakes at the same time. Wider segments of the rivers and the lakes connecting the upstream and downstream will require the use of the term $z_o(\mathbf{s})$ since the random observation at any location can depend on that from any other location, not just the ones upstream or downstream. This gives rise to the general model

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\mathbf{T}\boldsymbol{\beta} + \mathbf{z_u}(\mathbf{s}) + \mathbf{z_d}(\mathbf{s}) + \mathbf{z_o}(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{2}$$

The additional term, $z_o(\mathbf{s})$, may, however, render one or both of $z_u(\mathbf{s})$ and $z_d(\mathbf{s})$ non-significant since the omni-directional term may capture all the dependence. The data alone may not be rich enough to separate out the directional dependences. This parallels a very common problem in spatial statistics, see for example Section 2.3.2 in Banerjee *et al.* (2004) on assessing anisotropy using directional variograms. They remark that "Directional variograms from data generated under a simple isotropic model will routinely exhibit differences

of magnitude seen in Figure 2.9(a)". One possible solution can be the introduction of weights for various components, analogous to the discussion in Section 2.2.1. Suitable prior covariance structure for the $z_o(\mathbf{s})$, $z_u(\mathbf{s})$ and $z_d(\mathbf{s})$ processes may also help in identifying them.

A further drawback of the above formulation is the a-priori assumption of independence of the tail-up, tail-down and omni-directional components. It is not hard to imagine applications where these cannot be assumed independent, and there can be confounding effects between the three components, e.g. the same fish (creature) can travel both up stream and down stream and 'horizontally' as well. In such cases a multivariate specification must be provided. There are well known problems of multivariate spatial specifications and either a separable model or a linear model of co-regionalization can be specified, see for example, Gelfand *et al.* (2004) and the references therein, including Ver Hoef and Barry (1998).

3   Extension to the Space-time data

The authors discuss the possibility of extending the model to space-time data. Indeed, the model representation in Equation (2) can easily do that:

$$Y(\mathbf{s}, \mathbf{t}) = \mathbf{x}(\mathbf{s}, \mathbf{t})^{\mathbf{T}} \boldsymbol{\beta}_{\mathbf{t}} + \mathbf{z}_{\mathbf{u}}(\mathbf{s}, \mathbf{t}) + \mathbf{z}_{\mathbf{d}}(\mathbf{s}, \mathbf{t}) + \mathbf{z}_{\mathbf{o}}(\mathbf{s}, \mathbf{t}) + \epsilon(\mathbf{s}, \mathbf{t}), \qquad (3)$$

where the spatial processes at a particular time point are extended to spatio-temporal processes indexed by time point t, $(t = 1, 2, \ldots)$. A careful choice of the dynamic processes is necessary for model description, identification, estimation, and prediction. The covariate process $\mathbf{x}(\mathbf{s}, \mathbf{t})$ may depend on time and may need to be modeled as well; see for example

Huerta *et al.* (2004) and Sahu *et al.* (2007) where meteorological variables such as temperatures are modeled simultaneously with ozone concentration levels. The process $\boldsymbol{\beta}_t$ can be assumed to be $\boldsymbol{\beta}_t = \rho\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t$ where $\boldsymbol{\eta}_t$ are independent Gaussian random variables. The parameter ρ can be assumed to be: (a) zero for independence of $\boldsymbol{\beta}_t$'s, (b) one for random walk, and (c) some non-zero value in the interval $(-1, 1)$ corresponding to auto-regressive processes. The tail-up ($z_u(\mathbf{s}, \mathbf{t})$), tail-down ($z_d(\mathbf{s}, \mathbf{t})$) and omni-directional ($z_o(\mathbf{s}, \mathbf{t})$) processes can be assumed to be independent over time as a simple starting model. Complex, multivariate space-time interaction can be built up by joint modelling of the three processes; Chapter 8 in Banerjee *et al.* (2004) is an excellent starting point for this sort of modeling. The pure error process $\epsilon(\mathbf{s}, \mathbf{t})$ is usually assumed to be independent in space and time, providing the so called 'nugget effect'.

4   Extension to the Generalized Linear Models

The first stage Gaussian models described so far will not be appropriate for discrete data. It is also very common to observe presence-absence data for species or chemicals in a stream network. In those cases the Gaussian distribution assumption for the $Y(\mathbf{s}, \mathbf{t})$ must be replaced by an appropriate member of the exponential family of distributions, and the model specification (3) will now be written as:

$$g(E(Y(\mathbf{s}, \mathbf{t}))) = \mathbf{x}(\mathbf{s}, \mathbf{t})^{\mathbf{T}}\boldsymbol{\beta}_{\mathbf{t}} + \mathbf{z}_{\mathbf{u}}(\mathbf{s}, \mathbf{t}) + \mathbf{z}_{\mathbf{d}}(\mathbf{s}, \mathbf{t}) + \mathbf{z}_{\mathbf{o}}(\mathbf{s}, \mathbf{t}), \tag{4}$$

for a suitable link function $g(\cdot)$. Process assumptions made on the second and subsequent stages can remain the same except for the nugget effect $\epsilon(\mathbf{s}, \mathbf{t})$ which will no longer be there, although there are computational reasons for keeping the term.

The likelihood function for these variance components models will not be tractable for estimation purposes. The Bayesian computation methods based on Markov chain Monte Carlo (MCMC) techniques can be used as an alternative. However, the Bayesian methods will require specification of prior distributions for all the parameters and hyper-parameters. Once a MCMC method has been successfully implemented, it is a relatively straightforward task to perform predictions using posterior predictive distributions.

5   Additional References

Banerjee, S., Carlin, B. P., Gelfand, A. E. (2004) Hierarchical modeling and analysis for spatial data. Chapman& Hall/CRC.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004) Nonstationary Multivariate Process Modeling through Spatially Varying Coregionalization (with discussion). *Test*, **13**, 1–50.

Huerta, G., Sanso, B., and Stroud, J. R. (2004), "A spatiotemporal model for Mexico City ozone levels," *Journal of the Royal Statistical Society*, Series C, **53**, 231–248.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007). High Resolution Space-Time Ozone Modeling for Assessing Trends. *Journal of the American Statistical Association*,

102, 1221–1234.

Ver Hoef, J. M. and Barry, R. P. (1998) Constructing and fitting models for co-kriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, **69**, 275-294.