Experiments with fast sequential simulation for assessment of geospatial algorithms

Peter Doucette[1], *John Dolloff*[1], *Alexander Spizler*

National Geospatial-Intelligence Agency, 7500 GEOINT Drive, Springfield, VA

[1]contractor

Approved for public release 14-376

Abstract

A simple, reliable, and fast sequential simulation (FSS) technique based upon the exponential variogram model is demonstrated. FSS is shown to be significantly faster and simpler than standard implementations of sequential Guassian simulation, but with a corresponding tradeoff of modeling flexibility. This tradeoff may be adequate for many analysis scenarios in GIScience, which may in turn motivate broader community appeal. Monte Carlo experiments with FSS are demonstrated with evaluating conflation matching performance for geographic information system (GIS) features.

Keywords: sequential simulation, conflation.

1. Introduction

Geostatistical techniques for modeling and simulating uncertainty in spatial data (Cressie, 1993; Goovaerts, 1997) have been applied to limited extents within the broader GIScience community. This may be due in part to a general lack of user awareness and/or the relative utility of algorithm implementation (Brown and Heuvelink, 2007).

Simulation of uncertainty is a general strategy to assess its effects when propagating through a "black box" algorithm, particularly when analytical methods are impractical. The main drawback is high computational expense. Sequential simulation methods are devised as an approximate, but faster alternative compared to computing an exact solution from, e.g., $\sqrt{\sum_z}$, where $\sum_z$ is a full *nxn* covariance matrix for a scalar *z* that corresponds to a 2D random field, and *n* is the number of points in the field. Sequential Gaussian simulation (SGS) is a standard implementation found in popular geostatistical packages, e.g., *gstat* (Pebesma, 2004), and *GSLIB* (Deutsch and Journel, 1992).

This paper demonstrates application of the *fast sequential simulation* (FSS) method (Dolloff and Doucette, 2014) to a general topic of interest in GIScience: conflation of GIS feature layers. FSS is shown to be significantly faster than SGS found in *gstat* or *GSLIB* for unconditional simulations. This is achieved by exploiting properties particular to the exponential function for the variogram model. Although FSS is comparatively less general in application, its dramatic speed gain, along with its simplicity, may motivate more awareness of uncertainty propagation among a broader spectrum of users.

2. Fast Sequential Simulation (FSS)

The following description of FSS is condensed from Dolloff and Doucette, 2014. The simulation algorithm used is described briefly here for a scalar *z*. For the *k*-th row and *l*-th column of a *K*x*L* grid:

$$z(k+1,l+1) = r \cdot z(k+1,l) + s \cdot z(k,l+1) - r \cdot s \cdot z(k,l) + u(k,l)\,, \qquad (1)$$

where $r = e^{-\delta l/T_x}$ and $s = e^{-\delta k/T_y}$, δk and δl are the specifiable grid spacings (in linear units of meters) for the rows and columns (*y* and *x* directions) respectively, and T_x and T_y are the specifiable spatial correlation distance constants (meters). The term u is a random sample of Gaussian white noise of the form, $random\_N(0,\sigma_u)$, where $\sigma_u^2 = (1-r^2)(1-s^2)\sigma_z^2$, and σ_z^2 is the specifiable variance applicable to all *z* across the grid. The corresponding normalized correlation function (correlogram) is both separable and exponential, i.e., the correlation between two arbitrary grid points separated by Δk rows and Δl columns is equal to $r^{\Delta l}s^{\Delta k}$, or directly in terms of distances Δx and Δy in meters, $e^{-\Delta x/T_x}e^{-\Delta y/T_y}$.

The above $z(k,l)$ across the grid correspond to a realization of a mean-zero Gaussian 2D random field. Note that equation (1) can be easily extended to simulate multi-variate perturbations. The simplest such case corresponds to uncorrelated horizontal perturbations *x* and *y*, where equation (1) is implemented twice: once when *z* corresponds to *x* perturbations, and once when *z* corresponds to *y* perturbations. The spatial correlation distance constants and variance can also be specified independently for *x* and *y* perturbations. Extensions of equation (1) to multi-variate perturbations which are correlated between components as well as to non-homogeneous perturbations are also relatively straightforward.

Figure 1 (*left*) shows time performance comparisons among five different techniques for unconditional simulation of stationary and isotropic fields. Cholesky and SQRTM involve computing a solution for $\sqrt{\sum_z}$, which becomes prohibitive for large *n*. Open

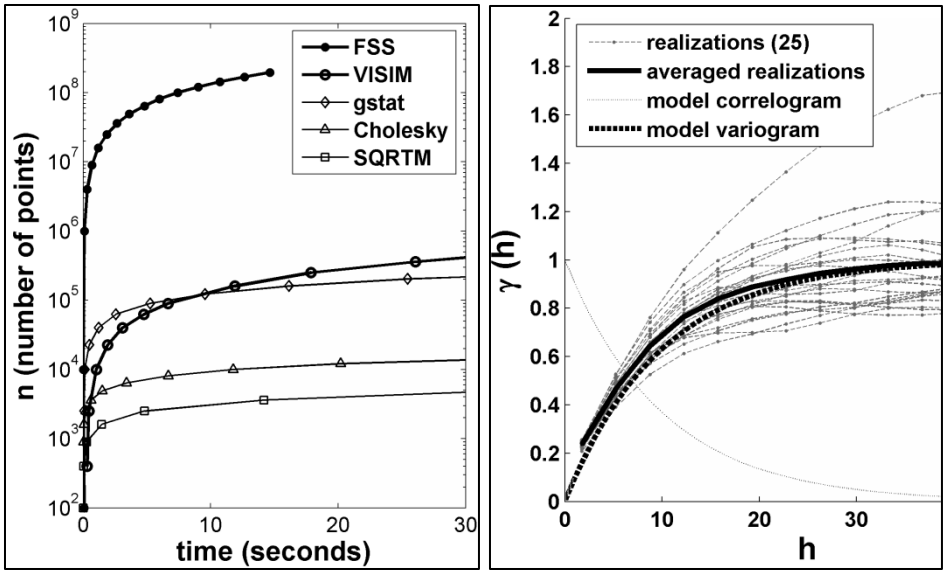

Figure 1: (***left***) speed comparisons among five methods for unconditional simulation of square grids (adapted from Dolloff and Doucette, 2014); (***right***) FSS variogram model recovery for 25 realizations of a 100x100 point grid, with $T_x = T_y = 10$; $\sigma_x = \sigma_y = 1$; $\delta_x = \delta_y = 1$.

source SGS implementations of *VISIM* (Hansen, 2011) and *gstat* (Pebesma, 2014) were ~1 order of magnitude faster than Cholesky. FSS was ~4 orders of magnitude faster than Cholesky. Figure 1 (*right*) demonstrates the variogram model recovery for FSS by averaging 25 realizations for a 100x100 grid.

3. Experiments with Data Conflation

The classic conflation problem involves matching the individual features from two GIS layers in order to transfer attributes between them. To quantify the matching performance of an automated conflation algorithm using real data sets requires a costly manual truthing process. However, simulation of feature layers allows for automation of this the process, which can provide useful insights (Doucette, et al., 2013).

3.1. Simulating polylines

Figure 2 shows a sample FSS realization applied to TIGER (open data from the US Census Bureau) road centerlines for Spartan Village on the MSU campus. Depending upon the simulation model parameters used, it is possible to introduce spurious intersections within a given realization that do not conform to the reference topology. These must be detected and filtered prior to conflation testing.

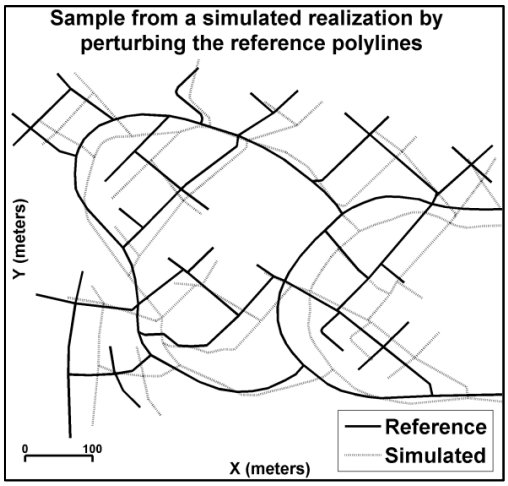

Figure 2: FSS unconditional realization (light) for Spartan Village TIGER roads (dark) using $T_x=T_y=500m$; $\sigma_x=\sigma_y=25m$; $\delta_x=\delta_y=50m$.

To simulate various realistic scenarios, it is possible to condition the simulated features in a variety of ways, including topologically. For example, providing for 1-to-1, 1-to-none, 1-to-many, and many-to-many feature matching; allowing only 4-way intersections; matching only primary versus secondary roads, etc.

Figure 3 shows results from a Monte Carlo (MC) analysis for GIS feature matching. The conflation algorithm tested was *RoadMatcher* (Vivid Solutions, 2005), which is open source. The test data were TIGER road centerlines covering a portion of Ingham County MI, centered on the MSU campus (*right*). The (*left*) pane shows matching performance as a function of coordinate perturbation for 3 different search distances (*SD*) used by *RoadMatcher*. Each *SD* curve was generated by averaging results from 10 realizations per plotted point over a perturbation range of 5:5:50m. This required a total of 100 simulations per *SD* Curve.

The *SD* plots reveal that matching performance peaked at *SD* = 100m regardless of the coordinate perturbations out to σ = 50m in *x* and *y*. The local spatial density of the roads likely influences this result. This is also somewhat evident in the heat map (right pane), which spatially shows the match error rates per road feature. I.e., darker roads indicate higher matching errors. For example, the interstate highways reveal a consistently high matching error rate likely due to the proximity of the dual lanes. Other isolated instances of high match error roads could be flagged for further investigation.

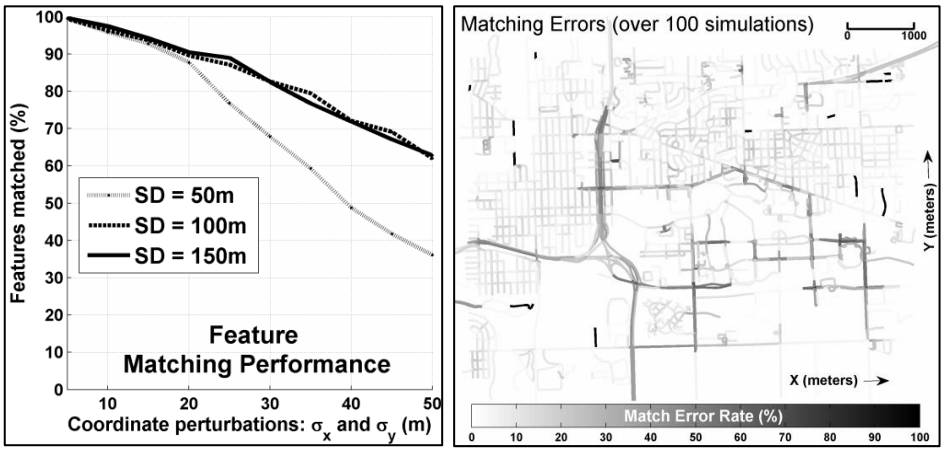

Figure 3: Feature matching (1-to-1) performance from MC analysis with *RoadMatcher* and TIGER road centerlines from a 7kmx5km area of Ingham County, MI. (***left***) matching performance as a function of coordinate perturbation and search distance (*SD*) with $T_x=T_y=500m$, and $\delta_x=\delta_y=50m$; (***right***) heat map showing match error rate per individual feature counted over 100 simulations with *SD=100m*. All simulations were unconditional, stationary, isotropic, with no nugget.

The goal of this type of MC analysis is to provide the user of conflation algorithms the ability to assess performance for parameter tuning. Furthermore, a matching performance heat map can expose potential problem features or regions for conflation.

3.2. Simulation grid density and model recovery

The 4,536 points (node and shape) that composed the reference polylines (Figure 3 (*right*)) were perturbed by simple bilinear interpolation from FSS grid realizations. Although bilinear interpolation is less robust than Kriging for sparse data, it is more feasible, and less computationally expensive for dense data. Figure 4 shows the effect of different grid spacing (δ) and distance constant (T_x) w.r.t. the model variogram when interpolating the *x*-coordinate perturbations for the polylines of Figure 3 (*right*). As δ increases (sparser samples), the experimental variograms over-smooth. As T_x increases (more gradual curve rise), over-smoothing becomes less pronounced as expected.

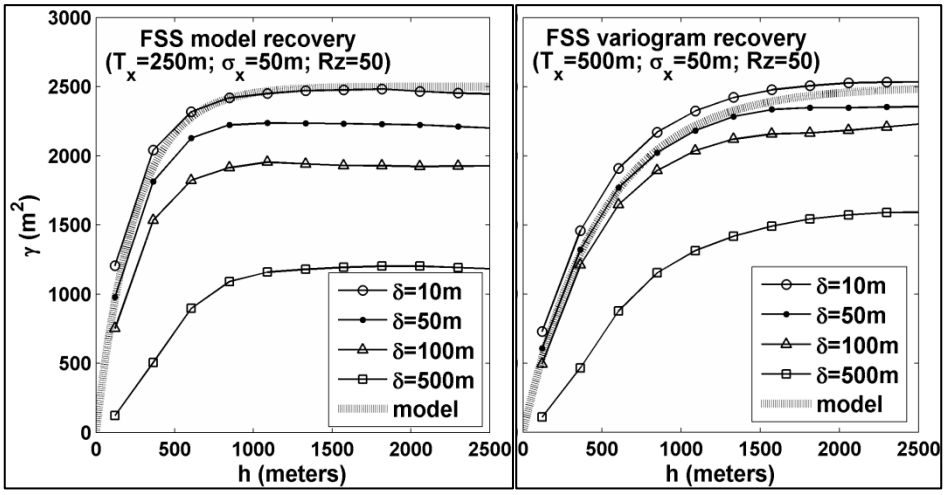

Figure 4: Experimental variograms for *x*-coordinate perturbations of polylines (averaged over *Rz=50 realizations*) for different grid spacing (δ). (***left***) $T_x=250m$; (***right***) $T_x=500m$.

4. Conclusion

Geostatistical techniques for simulating uncertainty have seen limited application within the broader GIScience community, which may be due to a lack of practical implementations. This paper demonstrated the *fast sequential simulation* (FSS) method to be significantly faster, albeit less general, than comparable open source implementations of SGS. If the user is able to trade off generality for simplicity and speed gain in order to simulate denser grids, then more motivation may be induced across a broader user base for certain applications.

FSS was applied to the problem of evaluating the polyline matching performance of a GIS data conflation algorithm for parameter tuning. For purposes of demonstration, only unconditional simulation was demonstrated, and variogram model parameters were empirically derived. While this approach may be reasonable for some evaluation scenarios, others may require conditional simulations (spatial, topological, and attributional), with variogram model parameters derived from truthed data layers.

References

Brown, J. and Heuvelink, G. (2007), "The Data Uncertainty Engine (DUE): A software tool for assessing and simulating uncertain environmental variables". *Journal of Computers & Geosciences*, 33(2), pp. 172-190.

Cressie, N. (1993), *Statistics for Spatial Data*. Wiley, New York.

Deutsch, C. and Journel, A. (1998), *GSLIB: Geostatistical Software Library and User's Guide* (2nd ed.). Oxford University Press, New York.

Doucette, P., Dolloff, J., et al., (2013), "Evaluating Conflation Methods using Uncertainty Modeling". *Proc. of SPIE*, vol. 8747.

Dolloff, J. and Doucette, P. (2014), "The Sequential Generation of Gaussian Random Fields for Applications in the Geospatial Sciences". (submitted) *ISPRS Int. Journal of Geo-Information*.

Goovaerts, P., (1997), *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.

Hansen, T.M. (2011), mGstat (version 0.99). http://mgstat.sourceforge.net/mGstat.pdf

Pebesma, E.J. (2014), Package 'gstat' (version 1.0-19)

http://cran.r-project.org/web/packages/gstat/gstat.pdf

Vivid Solutions (2005), RoadMatcher (ver 1.4).

http://www.vividsolutions.com/products.asp?catg=spaapp&code=roadmatcher