Geostatistical inference under preferential sampling: Background and motivation

John Paige

Department of Statistics, University of Washington Seattle, WA, 98195, USA

As discussed in Diggle et al. [2010], geostatistics considers stochastic processes varying in space. In some applications the process, S, might be measured at a set of discrete locations, say $X = \{x_1, x_2, ..., x_n\}$ (often a subset of $\mathbb{R}^2$, as in Diggle et al. [2010]). In others, S might be measured as a set of averages or a quantile function over different regions. Here, we consider the following model for measured observations, $Y_i \in \mathbb{R}$, of the spatial process:

$$Y_i = \mu + S(x_i) + Z_i, \tag{1}$$

where $i \in \{1, ..., n\}$, μ is the mean of the measured process, and the $Z_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$ can be interpreted as independent fine-scale measurement errors. Note that in the above model we assume $E[S] = 0$, since the mean of right-hand side can be changed with the μ parameter. We further assume $S(X) = \{S(x_1, ..., S(x_n)\}$ is multivariate normal with $\text{Cov}(S(x), S(x'))$ as a function of $|x - x'|$. Note that assuming a constant mean is not restrictive, since we can model μ as a function in space using a generalized linear regression framework, letting $E[Y_i] = \mu_i = \vec{X}_i'\vec{\beta}$, where errors are correlated due to spatial correlations in S.

Correlations in S are modelled under the assumption of stationarity and isotropy in the case of Diggle et al. [2010]. Stationarity in combination with isotropy means that, in addition to the form of the covariance in values of S being a function of distance alone, the variance of the spatial process becomes

$$\text{Var}(S(x) - S(x')) = \sigma^2 - \sigma^2\rho(|x - x'| \ |\vec{\theta}). \tag{2}$$

The positive definite function $\rho(\cdot|\vec{\theta})$ represents a correlation depending on distance conditional on the parameters $\vec{\theta}$. The variance in Eq. (2) is called a variogram function. Diggle et al. [2010] uses the Matérn correlation function,

$$\rho(u \mid \phi, \kappa) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}(u/\phi)^\kappa K_\kappa(u/\phi),$$

where $\kappa > 0$ is a smoothness parameter, $\phi > 0$ is the spatial scale parameter, and $K_\kappa(\cdot)$ is the modified Bessel function of the second kind of order κ. As noted in Diggle et al. [2010], the Matérn class of covariance functions is very flexible, containing exponential correlation as a special case when $\kappa = 0.5$, and is commonly used in geostatistics. One difficulty when using a Matérn covariance model is that ϕ and κ are sometimes difficult to estimate jointly, especially κ.[Diggle et al., 2010].

There are several advantages to modeling the variogram in a spatial model. It provides a simple interpretation relating correlations and variance in S over space as a function of distance. In addition, it is easy to estimate assuming stationarity and under certain distributional assumptions of the sample locations. A point cloud method of estimating empirical variogram works by plotting the distance between observations y_i and y_j versus $(y_i - y_j)^2/2$, which is the method of moments estimator for the variance based on two observations assuming the y_i have zero expectation [Diggle et al., 2010]. If the y_i are not mean zero, we can simply subtract off μ from Eq. (1). Additionally, we could partition the possible distances between observation locations into sufficiently fine bins, and estimate the variance between differences of observation values when the corresponding distances are in any given bin. Of course, if the bins are too fine, then our variogram estimates within each distance bin would be poor. More advanced methods for estimating variograms are given in Chilès and Delfiner [1999, section 2.2], Cressie [1985], and Cressie [1991, section 2.4].

In classical geostatistical analysis, it is assumed that the observation locations are independent of the measured spatial field [Diggle et al., 2010]. However, in the case of certain datasets, this might not be the case. Consider a tornado chaser trying to measure the wind speed of the tornado. It would be a terrible idea to try to infer average wind speeds in a region from the tornado chaser's wind speed data. This is because the data they collect is much more likely to be in locations and times where the wind speeds are abnormally high. This is what Diggle et al. [2010] refers to as *preferential sampling*.

In order to account for preferential sampling, Isaaks and Srivastava [1988] and Srivastava and

Parker [1989] proposed a non-ergodic variogram estimator as an alternative to classical estimators that they claimed was more robust to nonstationary and preferential data. However, Curriero et al. [2002] found that the non-ergodic estimators 'possess no clear advantage' over the traditional estimators, and in fact performed worse in the cases they studied. Schlather et al. [2004] note that if S is stationary, then $M_k(h) \coloneqq E[S(x)^k | x, x+h \in X]$ is constant if the sampling process is non-preferential, since the expectation does not depend on $x+h$. However the expectations might not be constant under preferential sampling. Schlather et al. [2004] then defines tests for preferentiality (assuming that S is stationary) based on $M_1(h)$ and $M_2(h)$ using simulations under models assuming non-preferentiality.

In addition to analyzing variations in observation *values* over space, it is common to study the patterns in the observation *locations*, as in Diggle et al. [2010]. Point processes are stochastic processes used to model random distributions of countably many points throughout a spatial domain (here we assume a bounded subset of $\mathbb{R}^2$) [Gelfand et al., 2010]. Common models for point processes include homogeneous or inhomogeneous Poisson processes, Cox processes, and Markov point processes [Gelfand et al., 2010]. A homogeneous Poisson process with rate (intensity) $\lambda > 0$ is a point process such that for bounded Borel sets $B, B' \subset \mathbb{R}^2$ the following conditions are satisfied:

1. the number of points in a bounded Borel set, B, is a random variable given by $N(B) \sim Pois(\lambda \mu(B))$, where μ is the Lebesgue measure, and
2. $N(B) \perp\!\!\!\perp N(B')$ when $B \cap B' = \emptyset$

An inhomogeneous Poisson process is the same, except it has a rate $\lambda(x)$ that varies in space so that

$$N(B) \sim Pois\left(\int_B \lambda(x)\, dx\right)$$

Cox processes [Cox, 1955] (also known as doubly stochastic Poisson processes) are generalizations of inhomogeneous Poisson processes with the following properties:

1. Random rate $\Lambda = \{\Lambda(x) : x \in \mathbb{R}^2\}$ is a nonnegative stochastic process, and

2. In any fixed realization, $\Lambda(x) = \lambda(x) : x \in \mathbb{R}^2$, the point process is an inhomogeneous Poisson process with rate $\lambda(x) \geq 0$.

In the case of

$$\Lambda(x) = \exp\{\alpha + \beta S(x)\}, \tag{3}$$

where $S(x)$ is the stochastic Gaussian process defined in Eq. (1), the resulting point process is known as a log-Gaussian Cox process (lGCP) [Diggle et al., 2010].

Just as the variogram can measure the level of correlation in observation values through space, the K-function can be helpful when analyzing the level of correlation in observation locations. The K-function is defined as $K(s) = \lambda^{-1}E[N_0(s)]$ for homogeneous point processes, where $E[N_0(s)]$ denotes the expected number of points with distance s from another point. Under complete spatial randomness (CSR) (*i.e.* if it is consistent with a homogeneous Poisson process model) the K function has the form $K(s) = \pi s^2$, but when the empirical K function is above or below πs^2, the data is indicative of clustering or repulsion respectively. The K function can be used in Monte Carlo goodness of fit tests for point process models [Gelfand et al., 2010, Section 18.3]. It can also be used to test whether the data follows complete spatial randomness (CSR) or whether there is clustering or repulsion among the observation locations. The K function for lGCPs has the form:

$$K(s) = \pi s^2 + 2\pi \int_0^s \gamma(u) u \, du \tag{4}$$

where $\gamma(u) = \exp\{\beta^2\sigma^2\rho(u;\phi,\kappa)\} - 1$ is the covariance of Λ, $\text{Cov}(\Lambda(x), \Lambda(x+u))$ under the parameterization of Eq. (3).

The main contribution of Diggle et al. [2010] is the authors' use of a lGCP model for sample locations in conjunction with the spatial process model in Eq 1 for data being measured. They give a relatively simple parameter fitting procedure by maximizing a tractable Monte Carlo likelihood and give examples of the dangers that could occur under a 'naive' geostatistical analysis that assumed non-preferentiality. Further, they give new tests for preferentiality and compute the results of a Monte Carlo goodness of fit based on the K-function under their proposed model.

References

JP Chilès and P Delfiner. Geostatistics: Modeling spatial uncertainty. *New York: Wiley*, 1999.

David R Cox. Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164, 1955.

Noel Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, 17(5):563–586, 1985.

Noel Cressie. *Statistics for spatial data*. New York: Wiley, 1991.

Frank C Curriero, Michael E Hohn, Andrew M Liebhold, and Subhash R Lele. A statistical evaluation of non-ergodic variogram estimators. *Environmental and Ecological Statistics*, 9(1):89–110, 2002.

Peter Diggle, Raquel Menezes, and Ting-li Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 59(2): 191–232, 2010.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

EH Isaaks and R Mohan Srivastava. Spatial continuity measures for probabilistic and deterministic geostatistics. *Mathematical geology*, 20(4):313–341, 1988.

Martin Schlather, Paulo J Ribeiro, and Peter J Diggle. Detecting dependence between marks and locations of marked point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):79–93, 2004.

R Mohan Srivastava and Harry M Parker. Robust measures of spatial continuity. In *Geostatistics*, pages 295–308. Springer, 1989.