SPATIAL REGRESSION MODELS USING INTER-REGION DISTANCES IN A NON-RANDOM CONTEXT

Nicolas Christou (nchristo@stat.ucla.edu)
Department of Statistics
University of California, Los Angeles
8130 Mathematical Sciences Building
Los Angeles, CA 90025

Gary Simon (gsimon@stern.nyu.edu)
Department of Statistics and Operations Research
New York University
44 West Fourth Street
New York, NY 10012-1126

Abstract

This paper considers spatial data $z(\boldsymbol{s}_1), z(\boldsymbol{s}_2), \cdots, z(\boldsymbol{s}_n)$ collected at n locations, with the objective of predicting $z(\boldsymbol{s}_0)$ at another location. The usual method of analysis for this problem is kriging, but here we introduce a new signal-plus-noise model whose essential feature is the identification of hot spots. The signal decays in relation to distance from hot spots. We show that hot spots can be located with high accuracy and that the decay parameter can be estimated accurately. This new model compares well to kriging in simulations.

Key words and phrases: Hot spot; Kriging; Spatial prediction; Variogram.

1 Introduction

In this paper we deal with spatial data obtained at a single time. Such data occurs in mining, agriculture, atmospheric science, ecology, epidemiology, hydrology, meteorology, waste disposal, and so on. Often the goal of such a study is a prediction at an unsampled location.

Let $\boldsymbol{Z} = (z(\boldsymbol{s}_1), z(\boldsymbol{s}_2), \cdots, z(\boldsymbol{s}_n))'$ be the vector of the observed values at locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$. The objective is to predict the unobserved value $z(\boldsymbol{s}_0)$ at location $\boldsymbol{s}_0$ which is not one of $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$. These data may involve spatial correlation which cannot be ignored.

Kriging, a term introduced by Matheron (1963), is a very popular method to solve the problem of spatial prediction. It was first used in mining data. It assumes a random field expressed through a variogram or covariance function, and correct estimation of the variogram (or covariance function) is very crucial. The model assumption (see Cressie (1991)) is $Z(\boldsymbol{s}) = \mu + \delta(\boldsymbol{s})$ where $\delta(\boldsymbol{s})$ is a zero mean stochastic term with variogram $2\gamma(\cdot)$. If we assume intrinsic stationarity then $E(Z(\boldsymbol{s}+\boldsymbol{h})-Z(\boldsymbol{s})) = 0$ and the variogram is defined as $\mathrm{Var}(Z(\boldsymbol{s}+\boldsymbol{h})-Z(\boldsymbol{s})) = 2\gamma(\boldsymbol{h})$ This can be written as $\mathrm{Var}(Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})) = E(Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s}))^2$ and thus the method of moments estimator for the variogram can be used (also called the classical estimator, Cressie (1991))

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{N(\boldsymbol{h})} \sum_{N(\boldsymbol{h})} (Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j))^2 \tag{1}$$

where the sum is over $\mathrm{N}(\boldsymbol{h})$ such that $\boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}$. Kriging minimizes the mean squared error of prediction $E[z(\boldsymbol{s}_0) - \hat{z}(\boldsymbol{s}_0)]^2$, where $\hat{z}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i z(\boldsymbol{s}_i)$; that is, the predictor assumption is a weighted average of the sample values, and $\sum_{i=1}^{n} w_i = 1$ to ensure unbiasedness.

In this paper we assume that spatial data come from a signal plus independent noise. We replace spatial covariance structure with signal plus noise. Data of the signal-plus-noise kind may exist because of *"hot spots,"* a term very popular in epidemiology. For example a nuclear accident location will be a hot spot and will possibly cause thyroid cancer up to a certain distance from it. Other examples from epidemiology are Lyme disease, where the proximity to rivers and similar environments will increase the chance of lyme disease, thyroid goiter disease where the lack of iodine increases the chance of occurence (the further a person is from the sea the greater the risk), tuberculosis for which a hot spot may exist at a poor neighborhood.

In a study about Lyme disease by Magnarelli et al. (1993) and Magnarelli (1995) it was found that most of the patients were from central and southeastern Connecticut. This is because in those areas foci (hot spots) for Lyme borreliosis (Lyme disease) exist. Ticks and blood specimens were collected from white-tailed deer (odocoileus virginianus) and analyzed for Borrelia burgdorferi, the etiologic agent of Lyme disease. Another example from epidemilogy is the scrub typhus in north Queensland, Australia, reported by McBride et al. (1999). All cases reported were soldiers who had visited a training area in north Queensland. These are not the only examples from epidemilogy. The list of diseases and hot spots is quite extensive.

This new method has a number of very useful advanages. It exploits a non-random structure for the expected value, while leaving the noise component as statistically independent errors, it identifies hot spots among the data points, and allows the estimation of useful parameters, such as the number and locations of hot spots and the decay parameter.

2 Proposed Model - Estimation Technique

A hot spot is a location or region with high activity. As we move away from the hot spot the rate decays, related to the distance from the hot spot. We propose a model that uses exponential decay, and the decay rate is a parameter to be estimated.

2.1 The Model

Let $s_1, s_2, ..., s_n$ be the spatial locations. The proposed spatial regression model, the response at location s_i has the following form

$$z(\boldsymbol{s}_i) = \beta_0 + \beta_1 e^{-Bdist_{iH_1}} + \cdots + \beta_k e^{-Bdist_{iH_k}} + \epsilon_i \tag{2}$$

where

- $\beta_0, \beta_1, \cdots \beta_k$ regression coefficients
- B decay parameter
- dist_{iH_j} distance from data point i to hot spot j
- ϵ_i independent error term with standard deviation σ.

Of course $\beta_0, \beta_1, \cdots, \beta_k, B$ and σ are parameters to be estimated, as is k, the number of hot spots. We assume that the hot spots are among the points $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$. What we observe are the $z(\boldsymbol{s}_1), \cdots, z(\boldsymbol{s}_n)$ values. Given those values, and assuming the existence of hot spots, we are proposing the estimation technique described next.

2.2 Estimation Technique

To estimate and fit a model to the observed $z(\boldsymbol{s}_i)$ values, we want to create variables which are also functions of the distance between the data points. If indeed the data are generated by a number of hot spots then the variables which are most related to the $z(\boldsymbol{s}_1), \cdots, z(\boldsymbol{s}_n)$ are the ones which are essentially indicators for the hot spots. More specifically let d_{ij} be the Euclidean distance between spatial locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. We define the $n \times 1$ vector $\boldsymbol{X}_i$ as having j^{th} entry

$$(\boldsymbol{X}_i)_j = e^{-Bd_{ij}} \quad i = 1, \cdots, n \text{ and } j = 1, \cdots, n \tag{3}$$

where n is the number of spatial locations. For example the vector $\boldsymbol{X}_{13}$ will have the following form $\boldsymbol{X}_{13} = (e^{-Bdist_{13,1}}, \cdots, 1, \cdots, e^{-Bdist_{13,n}})'$ Therefore we can construct n of those vectors. Next we regress the response vector $\boldsymbol{Z}$ on the predictors $\boldsymbol{X}$ and other covariates that may be relevant. This model is overspecified in that there are too many predictors, so we use stepwise regression. The carriers entering into the model will also be identifiers of possible hot spots. If there are hot spots, then the predictors matching the hot spots will be the ones selected.

2.2.1 Estimation of Parameters

We consider here the case where one hot spot exists, and that the location of this hot spot is known. Given this information we employ the following procedure based on maximum likelihood estimation. Later we will present the more realistic case where the location of the hot spot is not known. We will show subsequently that we can, with very high probability, determine the correct number of hot spots and identify their locations. The model for the one hot spot case is $z(\boldsymbol{s}_i) = \beta_0 + \beta_1 e^{-Bd_{iH}} + \epsilon_i$, with normal ϵ_i. This has four unknown parameters, β_0, β_1, B, and σ^2. We estimate these parameters by maximum likelihood. If B were known, we would need only the ordinary slope and intercept estimates for a simple regression. We make an initial guess at B and then find ordinary least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. The guess at B is updated through Newton-Raphson and the process is iterated to convergence. The above technique is used to estimate B when the location of the hot spot is known. In general we do not know this location. One way to locate the hot spot is to go from data point to data point and repeat the above procedure as if that data point were the hot spot. Finally we should pick as the hot spot the one that gives the highest R^2. This process is not as painful as it sounds, as most of the data points are hopelessly unrealistic as choices for a hot spot.

Now, one can ask the following question: Among n data points what is the probability that the i^{th} data point is correctly selected as the hot spot? Next we will give an answer to this question, at least for a simple geometry.

2.3 Probability of Selecting the Correct Hot Spot

Let us consider here a five-point layout (one point in the center with the other four points forming a square around it). Suppose that the true hot spot is at data point 1 (northwest corner). We generate a data vector $\boldsymbol{Z}= (z(\boldsymbol{s}_1), \cdots, z(\boldsymbol{s}_5))'$=$\boldsymbol{U}_1+\boldsymbol{\epsilon}$, where

$\boldsymbol{U}_1 = (1, e^{-Bd_{21}}, e^{-Bd_{31}}, e^{-Bd_{41}}, e^{-Bd_{51}})'$. Here, we set $\beta_0 = 0$ and $\beta_1 = 1$, so that the model is $z(\boldsymbol{s}_i) = e^{-Bd_{i1}} + \epsilon_i$. The vector $\boldsymbol{U}_1$ is non-random and it is created just for notational convenience. Now, we create five potential carrier vectors (as in (3)) $\boldsymbol{X}_1$ through $\boldsymbol{X}_5$, where $(\boldsymbol{X}_i)_j = e^{-Bd_{ij}}, i, j = 1, \cdots, 5$. These potential carriers will all be based on a guess at B. For the moment we will postpone the question as to whether we correctly estimated B, and we will assume that B is known. Our results utilize the value of B. Trying a different B would only slightly alter the findings. We would like to make a first calculation to determine whether we correctly identify data point 1 as the hot spot. We will select data point 1 as the true hot spot using stepwise regression if the correlation of $\boldsymbol{Z}$ with $\boldsymbol{X}_1$ is the largest. We use the following relationship

$$P[\text{best correlation is not with } \boldsymbol{X}_1] =$$

$$P[\{\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_2})\} \cup \{\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_3})\} \cup$$

$$\{\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_4})\} \cup \{\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_5})\}]$$

In the right hand side of the previous equation the four events may not be disjoint, so this expression becomes

$$P[\text{best correlation is not with } \boldsymbol{X_1}] \leq$$

$$P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_2})] + P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_3})] +$$

$$P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_4})] + P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_5})]$$

Therefore a lower bound for the desired probability is

$$P[\text{best correlation is with } \boldsymbol{X_1}] = \text{P[select point 1 as hot spot]} \geq$$

$$1 - \{P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_2})\ ] + P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_3})\ ] + P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_4}\ )] + P[\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) < \mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_5})\ ]\} \tag{4}$$

Examine just the first of these probability calculations. This should establish a pattern which will let us solve for the others.

Observe that $\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_1}) = \dfrac{\left[\sum_{i=1}^{5}(z_i - \bar{z})(x_{i1} - \bar{x}_1)\right]^2}{\left\{\sum_{i=1}^{5}(z_i - \bar{z})^2\right\}\left\{\sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2\right\}}$

Similarly, $\mathrm{Corr}^2(\boldsymbol{Z}, \boldsymbol{X_2}) = \dfrac{\left[\sum_{i=1}^{5}(z_i - \bar{z})(x_{i2} - \bar{x}_2)\right]^2}{\left\{\sum_{i=1}^{5}(z_i - \bar{z})^2\right\}\left\{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2\right\}}$

The condition $\mathrm{Corr}^2(\mathbf{Z}, \mathbf{X_1}) < \mathrm{Corr}^2(\mathbf{Z}, \mathbf{X_2})$ is equivalent to

$$\frac{\left[\sum_{i=1}^{5}(z_i - \bar{z})(x_{i1} - \bar{x}_1)\right]^2}{\left\{\sum_{i=1}^{5}(z_i - \bar{z})^2\right\}\left\{\sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2\right\}} < \frac{\left[\sum_{i=1}^{5}(z_i - \bar{z})(x_{i2} - \bar{x}_2)\right]^2}{\left\{\sum_{i=1}^{5}(z_i - \bar{z})^2\right\}\left\{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2\right\}}$$

Because $\sum_{i=1}^{5} \bar{z}(x_{i1} - \bar{x}_1) = 0$ the above inequality can be written as

$$\frac{\left[\sum_{i=1}^{5} z_i(x_{i1} - \bar{x}_1)\right]^2}{\sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2} < \frac{\left[\sum_{i=1}^{5} z_i(x_{i2} - \bar{x}_2)\right]^2}{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2}$$

Finally, we can substitute $z_i = U_{i1} + \epsilon_i$ to get this:

$$\frac{\left[\sum_{i=1}^{5}(U_{i1} + \epsilon_i)(x_{i1} - \bar{x}_1)\right]^2}{\sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2} < \frac{\left[\sum_{i=1}^{5}(U_{i1} + \epsilon_i)(x_{i2} - \bar{x}_2)\right]^2}{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2}$$

In simplest notation we want $P[C^2/V_1 < D^2/V_2] = P[C^2/V_1 - D^2/V_2 < 0]$, where C, D are jointly normal, or $P([C/\sqrt{V_1} - D/\sqrt{V_2}][C/\sqrt{V_1} + D/\sqrt{V_2}] < 0)$. If we use the transformation $u = C/\sqrt{V_1} + D/\sqrt{V_2}$, and $v = C/\sqrt{V_1} - D/\sqrt{V_2}$ we can write this probability as $P(uv < 0)$, which can be expanded as

$$P(uv < 0) = P(u > 0 \cap v < 0) + P(u < 0 \cap v > 0) \tag{5}$$

The quantities $C/\sqrt{V_1} \pm D/\sqrt{V_2}$ have means

$$\frac{\sum_{i=1}^{5} U_{i1}(xi1 - \bar{x}_1)}{\sqrt{\sum_{i=1}^{5}(x_{i1} - \bar{x}_1)^2}} \pm \frac{\sum_{i=1}^{5} U_{i1}(x21 - \bar{x}_2)}{\sqrt{\sum_{i=1}^{5}(x_{i2} - \bar{x}_2)^2}}$$

and variances

$$2\sigma^2 \pm \frac{2\sigma^2}{\sqrt{V_1 V_2}} \sum_{i=1}^{5}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$$

It can be shown easily that the covariance between u and v is zero. As u and v are normal, they are therefore independent. Now $P(uv < 0) = P(u > 0 \cap v < 0) + P(u < 0 \cap v > 0) =$

$P(u > 0)P(v < 0) + P(u < 0)P(v > 0)$. Finally we have

$$P(uv < 0) = \Phi\left(\frac{\mu_u}{\sigma_u}\right)\left[1 - \Phi\left(\frac{\mu_v}{\sigma_v}\right)\right] + \left[1 - \Phi\left(\frac{\mu_u}{\sigma_u}\right)\right]\Phi\left(\frac{\mu_v}{\sigma_v}\right) \tag{6}$$

Therefore we can compute a lower probability bound of correctly identifying data point 1 as the true hot spot. Of course we need to compute also the probability of falsely identifying hot spots. The previous result, as given by equation (6) gives us only lower bound probabilities of identifying the true hot spot. For the points which are not the true hot spots, we desire upper bound probabilities, since again we cannot find exact probabilities. For example, one can ask the question: What is the upper bound of the probability of falsely identifying data point 2 as the true hot spot, given that the true hot spot is at data point 1? The following computation procedure will give us these probability upper bounds. The probability that data point 2 is falsely selected as the true hot spot is

$$P[\text{data point 2 is selected}] =$$

$$P[\{\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_2}) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_1})\ \} \cap \{\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_2}) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_3})\ \} \cap$$

$$\{\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_2}) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_4})\ \} \cap \{\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_2}) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_5})\ \}]$$

We know that because the true hot spot is at data point 1 the most difficult event in the right side of the previous equation is $\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X}_2) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X}_1)$. Therefore the probability of selecting data point 2 as the true hot spot, is less or equal than the probability that the correlation of $\boldsymbol{Z}$ with $\boldsymbol{X}_2$ will be greater than the correlation of $\boldsymbol{Z}$ with $\boldsymbol{X}_1$.

$$P[\text{data point 2 is selected}] \leq P[\mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_2}) \geq \mathrm{Corr}^2(\boldsymbol{Z},\boldsymbol{X_1})\ ]$$

This probability has already been computed using equations (4) and (6).

We can calculate these probabilities for each pair $(\boldsymbol{Z},\boldsymbol{X}_i)$ and thus establish a lower bound for the probability of correctly selecting the true hot spot and upper bound for the probability of falsely identifying spurious hot spots. These probability bounds, along with a simulation study, will be shown next.

2.3.1 Simulations

To examine the performance of the lower and upper bounds, as described above, we run simulations for the five-point layout. The true hot spot is at data point 1, (north-west corner).

We computed the probability bounds for various values of the true R^2 and B. We vary the true R^2 from 95% down to 50%, and the decay parameter from 0.005 up to 0.040. The variance of the error term is then determined from the equation below (7). The so-called true R^2 is the squared correlation between the population versions of $\boldsymbol{Z}$ and $\boldsymbol{X}$, where $\boldsymbol{X}$ has entries $e^{-Bd_{1h}}, \cdots, e^{-Bd_{5H}}$ and is given by

$$\text{True} R^2 = \frac{\sigma^2_{\text{signal}}}{\sigma^2_{\text{signal}} + \sigma^2_{\text{error}}} \tag{7}$$

The model for the data generation is $z(\boldsymbol{s}_i) = \beta_0 + \beta_1 e^{-Bd_{i1}} + \epsilon_i$. Setting $\beta_0 = 0, \beta_1 = 1$ we get $z(\boldsymbol{s}_i) = e^{-Bd_{i1}} + \epsilon_i$. The variance of the signal is the variance of the values $e^{-Bd_{11}}, \cdots, e^{-Bd_{51}}$. The probability bounds from these simulations are shown in Table 1. We observe that the lower and upper bound probabilities generally reassur us about the probability of finding the correct hot spot.

2.4 Test for a Hot Spot

Suppose that we tentatively identify a hot spot at location H. In the model $z(\boldsymbol{s}_i) = \beta_0 + \beta_1 e^{-Bd_{iH}} + \epsilon_i$ we would hope to reject the hypothesis $H_0 : B = 0$, versus the alternative $H_a : B > 0$. The power of the test will be

$$\begin{aligned} \text{Prob[reject } B = 0] &= \\ \text{Prob[hot spot correctly identified]} \times & \\ \text{Prob[reject } B = 0 \mid \text{correctly identified]} + & \end{aligned}$$

			Lower Bound	Probability Upper Bounds			
R^2	B	σ^2	Point 1	Point 2	Point 3	Point 4	Point 5
0.95	0.005	0.040	0.99983	0.00000	0.00017	0.00000	0.00000
0.90	0.005	0.057	0.99321	0.00000	0.00678	0.00000	0.00000
0.80	0.005	0.086	0.94450	0.00197	0.04992	0.00197	0.00165
0.70	0.005	0.113	0.84492	0.01710	0.10497	0.01710	0.01592
0.60	0.005	0.141	0.68009	0.05375	0.16050	0.05375	0.05191
0.50	0.005	0.172	0.45579	0.10996	0.21637	0.10996	0.10792
0.95	0.010	0.061	0.99999	0.00000	0.00001	0.00000	0.00000
0.90	0.010	0.088	0.99795	0.00001	0.00203	0.00001	0.00000
0.80	0.010	0.132	0.96585	0.00242	0.02774	0.00242	0.00156
0.70	0.010	0.173	0.87437	0.01871	0.07260	0.01871	0.01560
0.60	0.010	0.216	0.71065	0.05624	0.12546	0.05624	0.05141
0.50	0.010	0.265	0.48379	0.11271	0.18343	0.11271	0.10737
0.95	0.015	0.073	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.015	0.106	0.99931	0.00001	0.00066	0.00001	0.00000
0.80	0.015	0.158	0.97639	0.00288	0.01620	0.00288	0.00164
0.70	0.015	0.208	0.89131	0.02026	0.05227	0.02026	0.01589
0.60	0.015	0.259	0.72943	0.05858	0.10153	0.05858	0.05187
0.50	0.015	0.317	0.50155	0.11528	0.16001	0.11528	0.10788
0.95	0.020	0.080	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.020	0.116	0.99970	0.00002	0.00026	0.00002	0.00000
0.80	0.020	0.174	0.98131	0.00323	0.01040	0.00323	0.00182
0.70	0.020	0.228	0.90045	0.02140	0.04020	0.02140	0.01655
0.60	0.020	0.284	0.74023	0.06028	0.08631	0.06028	0.05290
0.50	0.020	0.348	0.51210	0.11713	0.14461	0.11713	0.10903
0.95	0.025	0.084	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.025	0.122	0.99982	0.00002	0.00013	0.00002	0.00001
0.80	0.025	0.183	0.98368	0.00344	0.00740	0.00344	0.00204
0.70	0.025	0.240	0.90543	0.02205	0.03309	0.02205	0.01737
0.60	0.025	0.299	0.74646	0.06125	0.07687	0.06125	0.05417
0.50	0.025	0.366	0.51837	0.11819	0.13483	0.11819	0.11043
0.95	0.030	0.087	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.030	0.126	0.99987	0.00002	0.00007	0.00002	0.00001
0.80	0.030	0.189	0.98490	0.00353	0.00577	0.00353	0.00227
0.70	0.030	0.248	0.90829	0.02234	0.02884	0.02234	0.01819
0.60	0.030	0.309	0.75022	0.06167	0.07101	0.06167	0.05543
0.50	0.030	0.378	0.52223	0.11864	0.12866	0.11864	0.11182
0.95	0.035	0.088	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.035	0.129	0.99989	0.00002	0.00005	0.00002	0.00001
0.80	0.035	0.193	0.98558	0.00355	0.00484	0.00355	0.00249
0.70	0.035	0.252	0.91004	0.02239	0.02625	0.02239	0.01893
0.60	0.035	0.315	0.75261	0.06175	0.06734	0.06175	0.05657
0.50	0.035	0.386	0.52473	0.11873	0.12475	0.11873	0.11307
0.95	0.040	0.090	1.00000	0.00000	0.00000	0.00000	0.00000
0.90	0.040	0.130	0.99990	0.00002	0.00004	0.00002	0.00001
0.80	0.040	0.195	0.98600	0.00352	0.00428	0.00352	0.00267
0.70	0.040	0.256	0.91118	0.02232	0.02462	0.02232	0.01956
0.60	0.040	0.319	0.75419	0.06165	0.06501	0.06165	0.05752
0.50	0.040	0.391	0.52640	0.11862	0.12225	0.11862	0.11412

Table 1: Probability lower bound for data point 1 being identified as the hot spot, and probability upper bounds for data points 2,3,4,5 being falsely identified as the hot spots for five-point layout. The true hot spot is at the north-west data point 1.

$$\sum_i \text{Prob[hot spot falsely identified at point } i] \times$$
$$\text{Prob[reject } B = 0 \mid \text{falsely identified]} \geq$$
$$\text{Prob[hot spot correctly identified]} \times$$
$$\text{Prob[reject } B = 0 \mid \text{correctly identified]} \qquad (8)$$

As the second term is likely to be small, we expect that (8) gives a very good lower bound.

The test uses the estimate $\hat{B}$ which was found earlier, and the limiting variance, which can be obtained from Fisher's information matrix.

2.4.1 Simulations to Obtain Asymptotic Variances

Simulations are run for data in the layout of the state of North Carolina (see Figure 1), with one point for each of the 100 counties (shown in Cressie (1991)). A hot spot is assumed at location $\boldsymbol{s}_{25}$. The simulation model used for the data generation is the following

$$z(\boldsymbol{s}_i) = \beta_0 + \beta_1 e^{-Bd_{i,25}} + \epsilon_i \qquad (9)$$

We set $\beta_0 = 0, \beta_1 = 4$, and we use various combinations of the true R^2 and B. The results of these simulations are shown on Table 2, with each line representing one run. All t-statistics are significant at the 1% level. In every case, the null hypothesis $H_0 : B = 0$ is rejected, given that the hot spot is correctly identified.

R^2	B	$\hat{B}$	$se_{\hat{B}}$	t	R^2	B	$\hat{B}$	$se_{\hat{B}}$	t
0.95	0.025	0.0252	0.000000780	28.49	0.95	0.035	0.0356	0.000001321	30.94
0.90	0.025	0.0252	0.000001637	19.72	0.90	0.035	0.0358	0.000002792	21.44
0.80	0.025	0.0254	0.000003654	13.27	0.80	0.035	0.0363	0.000006300	14.45
0.70	0.025	0.0255	0.000006221	10.22	0.70	0.035	0.0367	0.000010836	11.15
0.60	0.025	0.0256	0.000009612	8.27	0.60	0.035	0.0371	0.000016926	9.03
0.50	0.025	0.0258	0.000014318	6.82	0.50	0.035	0.0377	0.000025530	7.46
0.95	0.030	0.0304	0.000001028	29.96	0.95	0.040	0.0407	0.000001667	31.54
0.90	0.030	0.0305	0.000002168	20.75	0.90	0.040	0.0411	0.000003531	21.85
0.80	0.030	0.0308	0.000004870	13.98	0.80	0.040	0.0416	0.000007989	14.73
0.70	0.030	0.0311	0.000008345	10.78	0.70	0.040	0.0422	0.000013777	11.36
0.60	0.030	0.0315	0.000012984	8.73	0.60	0.040	0.0427	0.000021577	9.20
0.50	0.030	0.0318	0.000019498	7.21	0.50	0.040	0.0434	0.000032644	7.60

Table 2: t-statistics for testing $B = 0$, for North Carolina layout. All are significant at the 1% level.

3 Kriging Vs. Proposed Method

We wanted to be fair in comparing kriging with the proposed method. We first generated simulated data assuming that one hot spot exists and then data by the random field model friendly to kriging. The first type of simulated data favors the proposed method while the second type favors kriging. For all data the geography of the state of North Carolina is used (Figure 1) where there are 100 data points, one for each county.

3.1 One Hot Spot Model

For this simulation, we assume that there is one hot spot, selected at the southeast corner of the state (data point 25). For simulations assuming existence of one hot spot at location 25, the response at data point i is given by (9) above. We also need to choose the decay parameter B and the variance of the independent error terms

(ϵ's). We vary B from 0.025 up to 0.040 in steps of 0.005. The variance of the error term σ^2 is related to the true R^2 of the model (see equation (7)). The true R^2 goes from 95% down to 90% and then down to 50% in steps of 10%. All the parameters used in the data generation are shown on Table 3.

Parameters for data generation	
Proposed method	Kriging (estimates)
$\beta_0 = 0$ $\beta_1 = 4$ True $R^2 = 0.50, 0.60, 0.70, 0.80, 0.90, 0.95$ $B = 0.25, 0.30, 0.35, 0.40$	range ($\hat{\alpha} = 40 - 200$) nugget ($\hat{c_0} = 0.05 - 0.50$) sill ($\hat{c_0} + \hat{c_1} = 0.40 - 1.00$)

Table 3: Parameters for data generation. Note that the kriging parameters shown here are estimates, while the parameters for the proposed method are the true ones.

Given these values, what method can best fit these data, kriging or the proposed method? To answer this question we need to estimate the variogram for the kriging system and construct the carriers for the proposed method.

3.1.1 Kriging Estimates, Data Generated by Hot Spot Model

For the kriging system the exponential variogram is used

$$2\gamma(\boldsymbol{h};\boldsymbol{\theta}) = \begin{cases} 0, & \boldsymbol{h} = \boldsymbol{0} \\ c_0 + c_1(1 - exp(-\frac{\|\boldsymbol{h}\|}{\alpha})), & \boldsymbol{h} \neq \boldsymbol{0} \end{cases} \tag{10}$$

$\boldsymbol{\theta} = (c_0, c_1, \alpha)'$, where $c_0 \geq 0$, $c_1 \geq 0$, and $\alpha \geq 0$.

The above variogram parameters are assessed from the sample variogram, by minimizing the weighted sum of squares proposed by Cressie (1985)

$$\sum_{i=1}^{K}\left\{\frac{2\hat{\gamma}(\boldsymbol{h}(k))}{2\gamma(\boldsymbol{h}(k);\boldsymbol{\theta})}-1\right\}^2 |N(\boldsymbol{h}(k)| \tag{11}$$

3.1.2 Proposed Method

The proposed method calls for the construction of the independent variables $\boldsymbol{x}_i$, $i = 1, 2, ..., n$, as in (3). We need to estimate the B in (3) using the Newton-Raphson iterative process. We choose a starting value for B. We then regress the vector $\boldsymbol{Z}$ on the 100 predictors (one at a time) and we select the one with the highest R^2. After the variable selection we proceed with the Newton-Raphson estimation of B and the other parameters of the model, β_0, β_1, and σ^2.

3.1.3 Comparison

The two methods are compared using the Predicted Sum of Squares (PRESS) criterion PRESS $= \sum_{i=1}^{n}(z(\boldsymbol{s}_i) - \hat{z}(\boldsymbol{s}_i))^2$, where $\hat{z}(\boldsymbol{s}_i)$ is the predicted value at locaton $\boldsymbol{s}_i$ using the other $n-1$ values. As we mentioned earlier, we generate 100 data points. We omit one data point at a time and we estimate it using the remaining 99 data points. This is followed for both kriging and the proposed method. For example, after the omission of data point 1 we use data points at locations $\boldsymbol{s}_2, \cdots, \boldsymbol{s}_{100}$ to estimate the variogram. After the estimation of the variogram we predict $z(s_1)$ using a weighted average of the values $z(\boldsymbol{s}_2), \cdots, z(\boldsymbol{s}_{100})$.

Similarly, for the proposed method, we use the 99 data values to first locate the hot spot, estimate B, β_0, β_1, and σ^2 and then predict $z(\boldsymbol{s}_1)$. This procedure is followed for each data point. At the end we have available the actual (observed) data and the predicted data under kriging and the proposed method.

For each method and for every random sample generated we compute the predicted sum of squares. Then the ratio $PRESS_{\text{kriging}}/PRESS_{\text{proposed}}$ is computed.

In Table 4 we present the simulation results and the comparison between the two methods when the true hot spot is located at the edge of the state of North Carolina (data point 25). We generate 100 samples, each of size 100 (number of counties). The ratio represents the average predicted sum of squares of kriging (for the 100 random samples), divided by the average predicted sum of squares of the proposed method (for the same 100 samples). A ratio of the two predicted sum of squares greater than one indicates that the proposed method outperforms kriging, while a ratio of less than one is in favor of kriging. The proposed method outperforms kriging in the vast majority of the random samples (these results are not shown because we would need dozens of pages to present them!). Instead we present the average of these results. In all of them we observe that the proposed method outperforms kriging when the true R^2 is either low or high. For example when the true R^2 is 95%, $B = 0.025$, the variance of the signal is 0.4379, and the variance of the error terms (using equation (7)) is $\sigma^2_{error} = 0.0230$, we observe that the ratio of the two PRESS's is 1.3797. For this combination the proposed method is a big winner. As the true R^2 becomes smaller the ratio of the two methods is getting closer to one, which means that kriging improves over the proposed method, but never outperforms the proposed method. This is true for all the values of B. The improvement of kriging relative to the proposed method as the signal gets weaker occurs because the sample mean is a better predictor than any regression predictor. It is known that an increase of the nugget effect, leads kriging to become more like a simple average (see Isaaks and Srivastava (1989)). Therefore, when the error terms are very strong (which means weak signal), kriging can challenge the proposed method. We also ran simulations with low R^2 $(20\% - 30\%)$ but the ratio of the two PRESS's is never below 1.

		Decay Parameter B			
		0.025	0.030	0.035	0.040
True R^2	95 %	1.3797	1.3787	1.3850	1.3096
	90 %	1.2670	1.2780	1.2844	1.2372
	80 %	1.1755	1.1774	1.1863	1.1606
	70 %	1.1397	1.1400	1.1374	1.1422
	60 %	1.1166	1.1179	1.1192	1.1270
	50 %	1.0801	1.0921	1.0957	1.0962
σ^2_{signal}		0.4379	0.3677	0.3174	0.2808

Table 4: Ratio of kriging over proposed method predicted sum of squares for the one hot spot case at the edge of the state, location 25.

Figure 1: Map showing the 100 counties of North Carolina, numbered in alphabetical order. County names and distances are given in Cressie (1991). 1 inch=138 miles.

3.2 Simulated Data-Covariance Function Known

In the previous section we assumed that the data are generated by a hot spot and that they come from a signal plus independent error term. This assumption favors the proposed method. As a matter of fairness, we generate simulated data assuming a random field with specified covariance function. One would expect that kriging will perform much better.

The 100 county seats of the state of North Carolina are used again as our spatial locations. The spatial variables $z(s_1), z(s_2), \cdots, z(s_n)$ have covariance matrix $\boldsymbol{\Sigma}$, based on the exponential covariance function

$$\text{Cov}(\boldsymbol{h}; c_0, c_1, \alpha) = \begin{cases} c_0 + c_1, & \boldsymbol{h} = \boldsymbol{0} \\ c_1 exp(-\frac{\|\boldsymbol{h}\|}{\alpha}), & \boldsymbol{h} \neq \boldsymbol{0} \end{cases} \tag{12}$$

The Cholesky decomposition discussed by Cressie (1991) can be used to write $\boldsymbol{\Sigma} = \boldsymbol{LL'}$, where $\boldsymbol{L}$ is a lower triangular matrix. This facilitates the creation of data with this covariance structure. In simulations, we used $\alpha = 50$ miles and $\alpha = 100$ miles; nugget $c_0 = 0.5, 1.0, 1.5, 2.0$; sill $c_0 + c_1 = 1, 2, 3, 4$. After the data are generated as described above, the estimation of the variogram is needed. In order to be fair to both methods we fit both the exponential and the spherical variogram models for the kriging calculation. Since the data are generated using an exponential covariance function, fitting only with the exponential variogram this would result in favoring kriging. We should not assume, that in reality kriging knows the true covariance function. Therefore it is very reasonable to fit both the exponential and the spherical variogram models. For each combination of the parameters we generate 100 random samples to compare kriging with the proposed method using again the predicted sum of squares criterion. The results are shown on Table 5. When the average (100 samples) predicted sum of squares of kriging over the average of the predicted sum of squares of the proposed method is less than one, kriging outperforms the proposed method, and when the ratio is above one, the proposed method is a winner.

Although the covariance function is known, the proposed method performs better than kriging in some cases. Of course kriging on average outperforms the proposed method, and this is not a big surprise. However the performances of kriging and the proposed method are close, as the entries in Table 5 are close to one. Especially when kriging fits the spherical variogram incorrectly, the ratios of the two predicted sum of squares are closer to one. Even in some cases the proposed method outperforms kriging. On the other hand, when data are generated by a hot spot model, we observe

		$\alpha = 50$ miles							
		kriging fits (correct) exponential variogram				kriging fits (incorrect) spherical variogram			
		$c_0 = 0.5$	$c_0 = 1.0$	$c_0 = 1.5$	$c_0 = 2.0$	$c_0 = 0.5$	$c_0 = 1.0$	$c_0 = 1.5$	$c_0 = 2.0$
c_1	0.5	0.9503	0.9079	0.9293	0.9385	0.9824	0.9828	0.9764	0.9499
	1.0	0.9020	0.9238	0.9355	0.9143	1.0183	0.9843	0.9868	0.9860
	1.5	0.9127	0.9299	0.9351	0.9485	1.0097	1.0085	0.9754	0.9924
	2.0	0.9166	0.9045	0.9354	0.9344	0.9571	1.0173	0.9759	0.9875
		$\alpha = 100$ miles							
		kriging fits (correct) exponential variogram				kriging fits (incorrect) spherical variogram			
		$c_0 = 0.5$	$c_0 = 1.0$	$c_0 = 1.5$	$c_0 = 2.0$	$c_0 = 0.5$	$c_0 = 1.0$	$c_0 = 1.5$	$c_0 = 2.0$
c_1	0.5	0.9444	0.9541	0.9586	0.9447	0.9448	0.9546	0.9634	0.9513
	1.0	0.9264	0.9372	0.9554	0.9540	0.9399	0.9379	0.9760	0.9549
	1.5	0.9079	0.9306	0.9548	0.9485	0.9012	0.9235	0.9416	0.9447
	2.0	0.9019	0.9224	0.9348	0.9539	0.9096	0.9416	0.9126	0.9608

Table 5: Ratio of kriging over proposed method predicted sum of squares when the covariance structure is known.

ratios of the two predicted sum of squares to be around $1.35 - 1.40$, an indication that the proposed method is a clear winner in those simulations.

4 An Example

We compare the proposed method with kriging using the southwest of England unemployment data. For this data set the percentage of the total workforce unemployed in January, 1967, (see Cliff and Ord (1973)) in the 37 employment areas in the southwest of England is used (see Figure 3).

4.1 Kriging - Proposed Method

For the kriging calculation we first construct the sample variogram (see Figure 2). Based on the appearance of this sample variogram we fit the linear variogram with estimated parameters $\hat{c}_0 = 7.5$ and $\hat{b} = 0.1$:

$$2\hat{\gamma}(\boldsymbol{h}) = \begin{cases} 0, & \boldsymbol{h} = \boldsymbol{0} \\ 7.5 \ + \ 0.1\|\boldsymbol{h}\|, & \boldsymbol{h} \neq \boldsymbol{0} \end{cases} \tag{13}$$

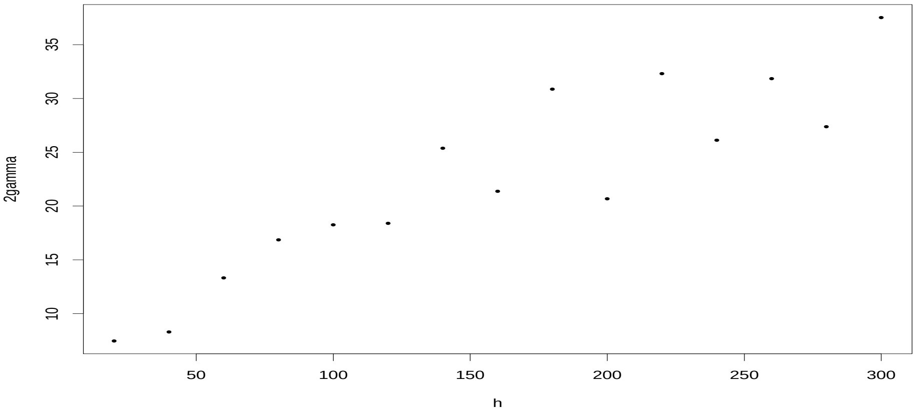

Figure 2: The variogram estimator $2\hat{\gamma}$ for the southwest of England unemployment data. Horizontal axis in miles.

For the proposed method we estimate the decay parameter $\hat{B} = 0.007$ to construct the predictors. The two methods are compared using the predicted sum of squares (PRESS). We found that the proposed method gives a smaller PRESS (PRESS kriging=6.4465, PRESS proposed=5.8549). We should mention here that even though the best variogram fitted to these data is the linear (see Figure 2), we also fitted the exponential and the spherical variograms. This was done because one may fit the wrong variogram to the data. Again we observe that the proposed method outperforms kriging. Using the exponential variogram (PRESS kriging=6.6383), and using the spherical variogram (PRESS kriging=7.0828).

4.2 Conclusion

Figure 3 shows the hot spot predictor which enters the model at the 5% level of significance. We observe that only one predictor entered the model (number 13). Data point 13 is near the city of Bristol and near the data points that have low unemployment rates compared to the other ones. We can claim that there is a hot spot in the neighborhood of the data point 13. The hot spot is probably Bristol, where one expects to find low unemployment (more jobs near a big city). This result is consistent with that of Cliff and Ord (1973). They regressed the unemployment rate on the cartesian coordinates (x_1, x_2) and they found that the linear surface falls from the southwest to the northeast parts of the map and reflects the higher levels of unemployment in the extreme southwest, where the economy is heavily reliant upon tourism and mining. Our method has the advantage of identifying the hot spot.

5 Final Remarks

In solving the spatial prediction problem different approaches can be taken. Popular methods are kriging and trend-surface analysis (not discussed in this paper). The proposed new method assumes the existence of hot spots and we have shown that

Figure 3: Possible hot spot, data point 13, for the southwest of England unemployment data. 1 inch=59 miles.

these are reliably located. The debate on which method is best can be endless. The decomposition of the process $z(\boldsymbol{s}_i)$ into large-scale variation plus smaller-scale variation cannot be specified uniquely. Trend-surface prediction decomposes $z(\boldsymbol{s}_i)$ into large-scale variation plus white noise. Ordinary kriging prediction relies on a random field that decomposes $z(\boldsymbol{s}_i)$ into constant mean plus spatially correlated error term with variogram $2\gamma(\cdot)$. The new model is similar to trend-surface analysis, but with simple and easily-interpreted structure.

References

[1] Burden, R.L., and Faires, J.D. (1993). Numerical Analysis, fifth edition. PWS Pub. Co., Boston 768p.

[2] Cliff, A.D. and Ord, J.K. (1973). Spatial Autocorrelation.

[3] Cressie, N. (1990). The Origins of Kriging. *Mathematical Geology 22*, 239-252.

[4] Cressie, N. (1991). Statistics for Spatial Data. John Wiley, New York, 900p.

[5] Hogg, R.V., and Craig, A.T. (1995). Introduction to Mathematical Statistics, fifth edition. Prentice Hall, Englewood Cliffs, New Jersey, 564p.

[6] Isaaks, E.H. and Srivastava, R.M. (1989). Applied Geostatistics. Oxford University Press, New York, 561p.

[7] Magnarelli, L.A., Anderson, J.F., and Cartter, M.L. (1993). Geographic distribution of white-tailed deer with ticks and antibodies to Borrelia burgdorferi in Connecticut. *Yale Journal of Biology Medicine 66(1)*, 19-26.

[8] Magnarelli, L.A., Denicola, A., Stafford, K.C., and Anderson, J.F. (1995). Borrelia burgdorferi in an urban environment: white-tailed deer with infected ticks and antibodies. *Journal of Clinical Microbiology 33(3)*, 541-544.

[9] McBride, W.J., Taylor, C.T., Pryor, J.A., and Simpson, J.D. (1999). Scrub typhus in north Queensland. *Medicine Journal 170*, 318-320.

[10] Sen, A. and Srivastava, M. (1990). Regression Analysis: Theory, Methods, and Applications. Springer-Verlag, New York 347p.

[11] Zimmerman, D.L. and Zimmerman, M.B. (1991). A Comparison of Spatial Semivariogram Estimators and Corresponding Ordinary Kriging Predictors. *Technometrics 33*, 77-91.