11-2014

Optimal Location Design for Prediction of Spatial Correlated Environmental Functional Data

Mahdi Rasekhi
Shahid Chamran University, Iran, m-rasekhi@phdstu.scu.ac.ir

B. Jamshidi
Shahid Chamran University, Iran, b-Jamshidi@phdstu.scu.ac.ir

F. Rivaz
Shahid Beheshti University, Iran, rivaz@sbu.ac.ir

Optimal Location Design for Prediction of Spatial Correlated Environmental Functional Data

M. Rasekhi
Shahid Chamran University
Ahvaz, Iran

B. Jamshidi
Shahid Chamran University
Ahvaz, Iran

F. Rivaz
Shahid Beheshti University
Tehran, Iran

The optimal choice of sites to make spatial prediction is critical for a better understanding of really spatio-temporal data. It is important to obtain the essential spatio-temporal variability of the process in determining optimal design, because these data tend to exhibit both spatial and temporal variability. Two new methods of prediction for spatially correlated functional data are considered. The first method models spatial dependency by fitting variogram to empirical variogram, similar to ordinary kriging (univariate approach). The second method models spatial dependency by linear model co-regionalization (multivariate approach). The variance of prediction method was chosen as the optimization design criterion. An application to CO concentration forecasting was conducted to examine possible differences between the design and the optimal design without considering temporal structure.

Keywords: Spatio-temporal process, functional data, optimal design, ordinary kriging, total model, GenSA optimization

Introduction

A method for optimum choice of location to obtain better spatial prediction is needed. Ordinary geo-spatial prediction methods deal with scalar value for random variables in each coordinate (Cressie, 1993). Recently, Giraldo et al. (2011a, 2011b) have extended geo-spatial prediction methods for one-dimensional functional data (curves) based on the statistic proposed by Delicado et al. (2010). Delicado et al. constructed a statistic for spatial correlated functional data and proposed a new experimental variogram based on L_2 functional distance to function value spatial data. Giraldo et al. extended the ordinary kriging model. They also

used the total linear functional model for creating a new prediction method. In the next section, these models are introduced briefly.

Many researchers have investigated the problem of spatial sampling design e.g. Zhu & Stein (1999), Wikle & Royle (1999), Diggle & Lophaven (2006), Fuentes et al. (2007), etc. Here, variance of spatial prediction in extended ordinary kriging and Total model spatial prediction of function value data are chosen as criteria for obtained optimization. This is done by application of Xiang et al.'s (2012) optimization procedure, "Generalized Simulated Annealing for Efficient Global Optimization" or "GenSA," described in the section titled 'Optimization Procedure.' Following that, CO concentration data in Tehran weather pollution stations is introduced as spatial correlated functional data in seventeen stations. The proposed approach is illustrated in the section titled 'Application,' and for the possible differences between this design and the optimal design without considering temporal structure, the air monitoring network is redesigned based on the average data over time.

Prediction Procedures

Consider a functional spatial process $X = \{X_s(t) : s \in D \subseteq R^d, t \in [a,b]\}$, where functional variable X_s belongs to the separable Hilbert space H of square integrable functions defined on T for any $s \in D$. We assume second-order stationarity and isotropicity for each $t \in T$ in random process. Let $s_1, ..., s_n$ be the sites in D that we observe a realization $X_{s_1}, ..., X_{s_n}$ of the functional random process X_s.

Ordinary kriging for function value spatial data

Ordinary kriging to function value spatial data is extended by Giraldo e.al (2011a) as following model

$$\hat{X}_{s_0}(t) = \sum_{i=1}^{n} \lambda_i X_{s_i}(t); \quad \lambda_1, ..., \lambda_n \in R, t \in T \tag{1}$$

where $\hat{X}_{s_0}(t)$ is predicted function in location s_0. Modeling based on this prediction method needs some assumptions such as functional version of intrinsic stationarity and isotropicity (see Delicado et al., 2010). Empirical variogram for modeling spatial dependency is obtained founded on the following minimization variance considering unbiasedness condition

$$\min_{\lambda_1,\dots,\lambda_n} \int_T V\left(\hat{X}_{s_0}(t) - X_{s_0}(t)\right) dt . s.t \sum_{i=1}^{n} \lambda_i = 1 \tag{2}$$

where moment estimation method leads to

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \int_T \left(\chi_{s_i}(t) - \chi_{s_j}(t)\right)^2 dt \tag{3}$$

Therefore, ordinary variogram, including exponential, spherical and so on, is fit to (3) by least square method and scale value coefficients are obtained.

Total model prediction of function value spatial data

Giraldo et al. (2011b) create new predictor for function value spatial data based on functional total model (Ramsay & Silverman, 2005)

$$Y_i(t) = \beta_0(t) + \int_T X_i(\upsilon)\beta(\upsilon,t)d\upsilon + \varepsilon_i(t) \tag{4}$$

and multivariate spatial predictor (Ver Hoef & Cressi, 1993; Ver Hoef and Bari, 1998)

$$\begin{pmatrix} \hat{X}_1(s_0) \\ \vdots \\ \hat{X}_p(s_0) \end{pmatrix} = \begin{pmatrix} \lambda_{11}^1 & \cdots & \lambda_{1p}^1 & \cdots & \lambda_{n1}^1 & \cdots & \lambda_{np}^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \lambda_{11}^p & \cdots & \lambda_{1p}^p & \cdots & \lambda_{n1}^p & \cdots & \lambda_{np}^p \end{pmatrix} \begin{pmatrix} X_1(s_1) \\ \vdots \\ X_p(s_1) \\ \vdots \\ X_1(s_n) \\ \vdots \\ X_p(s_n) \end{pmatrix} \tag{5}$$

as follows

$$\hat{X}_{s_0}(v) = \sum_{i=1}^{n} \int_T \lambda_i(t,v) X_{s_i}(t) dt, \ v \in T \tag{6}$$

where $\lambda_n(t,v)$, ..., $\lambda_1(t,v)$ are functions $T \times T \rightarrow R$. Modeling based on this prediction need some assumption such as functional version of second-order stationarity and isotropicity that is proposed by Delicado et al. (2010). Coefficients of this model are found through following minimization variance with unbiasedness condition

$$\min_{\lambda_1(...),...,\lambda_n(...)} \int_T V\left(\hat{\chi}_{s_0}(v) - \chi_{s_0}(v)\right) dv \; s.t \; E\left(\hat{\chi}_{s_0}(v)\right) = E\left(\chi_{s_0}(v)\right) \qquad (7)$$

To solve equation (7), spatial dependency must be modeled with linear model co-regionalization (Wackernagel, 2003).

Optimization Procedure

Optimization is the process which one finds that value of a vector x, say, that maximizes or minimizes a given function f. The idea of optimization goes to the heart of statistical methodology, as it is involved in solving statistical problems based on least squares, maximum likelihood, posterior mode, and so on. Xiang et al. (2012) created global optimization procedure "GenSA" which is applicable for geo-statistical process. GenSA gives the lower and upper bound of geographical coordinates and finds optimum coordinates based on specific criteria. GenSA uses a distorted Cauchy-Lorentz visiting distribution, with its shape controlled by the parameter q_v

$$g_{q_v}\left(\Delta x(t)\right) \propto \frac{\left(T_{q_v}(t)\right)^{-\frac{D}{3-q_v}}}{\left(1 + (q_v - 1)\frac{\left(\Delta x(t)\right)^2}{\left(T_{q_v}(t)\right)^{-\frac{2}{3-q_v}}}\right)^{\frac{1}{q_v-1}+\frac{D-1}{2}}} \qquad (8)$$

Here t is the artificial time. This visiting distribution is used to generated a trial jump distance $\Delta x(t)$ of variable $x(t)$ under artificial temperature $T_{q_v}(t)$. The trial jump is accepted if it is downhill (in terms of the objective function). If the jump is uphill it might be accepted according to an acceptance probability. A generalized Metropolis algorithm is used for the acceptance probability

$$p_{q_v} = \min\left\{1, (1-q_a)\beta\Delta E\right\}^{\frac{1}{1-qa}} \qquad (9)$$

where q_a is a parameter.

The minimax of prediction model of variance was used as optimization criterion. Variance of mentioned predictions is calculated as follows

$$\sigma_{s_0}^2 = \int_T V\left(\hat{\chi}_{s_0}(t) - \chi_{s_0}(t)\right) dt \qquad (10)$$

Giraldo et al. (2011a) calculate variance of ordinary kriging, resulting in

$$\sigma_{s_0}^2 = \sum_{i=1}^{n} \lambda_i \int_T \Gamma_{s_i, s_0}(t) dt - \mu = \sum_{i=1}^{n} \lambda_i \gamma(h_i) - \mu \qquad (11)$$

where functional variogram $\Gamma(t)$ is defined by Delicado et al. (2010) and $\gamma(.)$ is the classical variogram fitted to empirical variogram. Giraldo et al. (2011b) calculate variance of Total model, resulting in

$$\sigma_{s_0}^2 = \int_T V\left(\hat{\chi}_{s_0}(v) - \chi_{s_0}(v)\right) dv = \sum_{i=1}^{n} Tr\left(\hat{C}_i^T Q_{ii} \hat{C}_i W\right) + 2\sum_{i<j} Tr\left(\hat{C}_i^T Q_{ij} \hat{C}_j W\right) (12)$$

More detail of this variance is provided in Giraldo et al. (2011a, 2011b). To calculate the predictions variance in any location s_0, a smoothing process is applied to the curves, which expands the curves and the functional parameters in terms of a set of Fourier basis functions. The number of Fourier basis is found by a Functional Cross-Validation procedure similar to the leave-one-out procedure (Cressie, 1993) introduced by Giraldo et al. (2011a).

Data

Transportation-related air pollution is one of Tehran's most important problems. One of the most hazardous air polluting agents is carbon monoxide (CO), often exceeding two or three times the average levels recommended by the World Health Organization (WHO). This gas is colorless, odorless, and tasteless, and its predilection to bind hemoglobin is 200-220 times more than that of oxygen. Thus, it can prevent oxygen transfer to tissues and cause tissue hypoxia. For these reasons, the demand for reliable data to assess progress in air quality has grown rapidly over the past decade. In fact, motivated by increasing air monitoring stations in Tehran, three newly designed sites are proposed for monitoring CO. The data set used here describes daily averages of carbon monoxide (in ppm) at 17 monitoring sites,

geographically distributed across Tehran. Following the air quality standards (NAAQS, http://www.epa.gov/air/criteria.html), daily CO concentration is measured as the daily 1-hour average concentration. The current analysis considers data collected in 2011.

Application

Positioning of Tehran air quality stations and curves of seventeen stations created by smoothing procedure (Ramsay & Silverman, 2005) with 31 Fourier basis is illustrated in Figure 1.
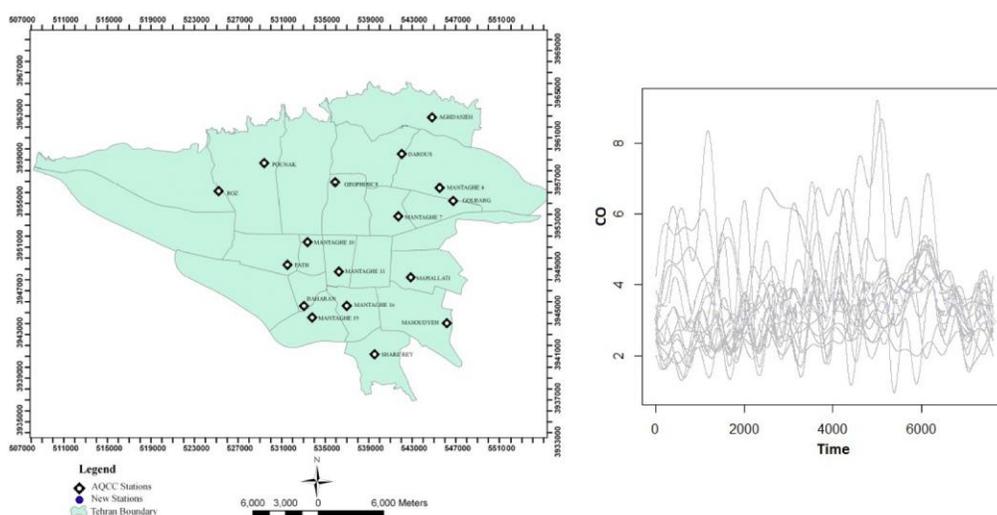

Figure 1. Position and CO concentration curves of seventeen air quality monitoring stations in Tehran

Considering variance of the mentioned prediction method in GenSA algorithm, three locations in Tehran map can be identified that minimize maximum predictions variance simultaneously. The possible differences between this design and the optimal design, without considering temporal structure, are surveyed, and the air monitoring network is redesigned based on averaged data over time. Figure 2 shows the optimal location based on Ordinary method for functional data vs. ordinary kriging on averaged time data (spatial data). It is worth mentioning that spherical

variogram (13) is chosen for modeling dependency structure, then parameters are estimated based on empirical variogram by applying least square method.

$$\gamma(\|h\|) = \begin{cases} \tau^2 + \sigma^2 & \|h\| > 1/\varphi \\ \tau^2 + \sigma^2 \left(\dfrac{3\varphi\|h\|}{2} - \dfrac{1}{2}\left(\varphi\|h\|\right)^3 \right) & \|h\| < 1/\varphi \\ 0 & o.w \end{cases} \qquad (13)$$
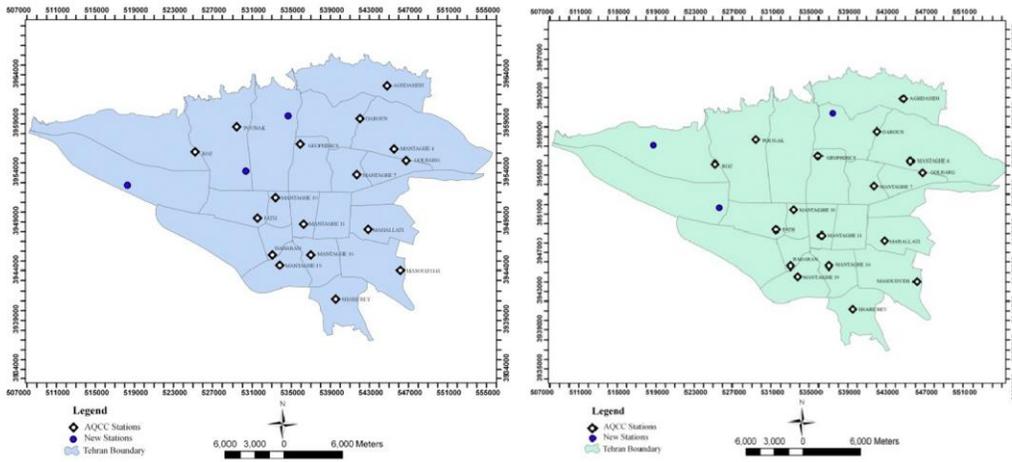

Figure 2. Optimal location for three monitoring sites based on kriging model for spatial data (left) and ordinary model for spatial functional data (right)

Figure 3 shows optimal location founded on Total prediction model for functional data vs. ordinary kriging on averaged time data (spatial data). Dependency structure is modeled with a linear model co-regionalization with exponential variogram (14)

$$\gamma(\|h\|) = \begin{cases} \tau^2 + \sigma^2 \left(1 - \exp(-\varphi\|h\|)\right) & \|h\| > 0 \\ 0 & o.w \end{cases} \qquad (14)$$
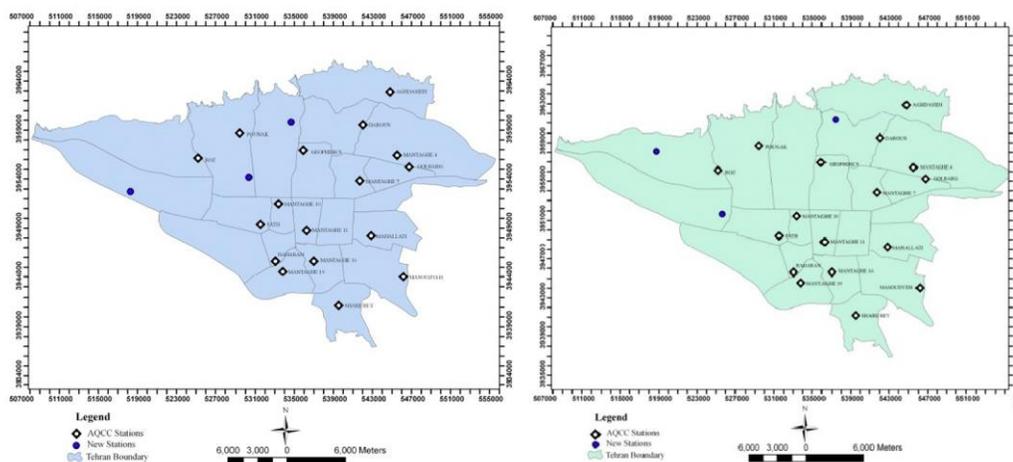
461

Figure 3. Optimal location for three monitoring sites based on kriging model for spatial data (left) and Total model for spatial functional data (right)

Results

Figures 2 and 3 show the optimal design based on ordinary model and Total model for the spatially correlated functional data.

Conclusion

Although the modeling of spatial dependency based on the two proposed functional models is different, both tend to locate new monitoring stations nearer from existing stations than in the non-functional version of ordinary kriging. Thus, considering time on spatial data affects location sampling. In other words, maximum variance of functional predictions of the three locations is global minimized closer to other stations, but optimal design based on averaged data over time (spatial data) tends to fill the space (Figures 2 and 3).

References

Cressie, N. (1993). *Statistics for spatial data.* New York: John Wiley and Sons.

Delicado, P., Giraldo, R., Comas, C., & Mateu, J. (2010). Statistics for spatial functional data: Some recent contributions. *Environmetrics*, *21*, 224-239.

Diggle, P., & Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, *33,* 53-64.

Fuentes, M., Chaudhuri, A., & Holland, D. M. (2007). Bayesian entropy for spatial sampling design of environmental data. *Environmental and Ecological Statistics, 14*, 323-340.

Giraldo, R., Delicado, P., & Mateu, J. (2011a). Ordinary kriging for function-value spatial data. *Journal of Environmental and Ecological Statistics, 18*, 411-426.

Giraldo, R., Delicado, P., & Mateu, J. (2011b). Geostatistics with infinite dimensional data: A generalization of cokriging and multivariable spatial prediction. *Matematica: ICM-ESPOL, 9*, 16-21.

Ramsay, J., & Silverman, B. (2005). *Functional data analysis* (2nd ed.). Berlin: Springer.

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

Ver Hoef, J., & Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology, 25*(2), 219-240.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction With Applications* (3rd ed.). Berlin: Springer.

Wikle, C. K., & Royle, J. A. (1999). Space-Time Dynamic Design of Environmental Monitoring Networks. *Journal of Agricultural, Biological, and Environmental Statistics*, *4*, 489-507.

Xiang, Y., Gubian, S., Suomela, B., & Hoeng, J. (2013). Generalized Simulated Annealing for Global Optimization: The GenSA Package. *The R Journal, 5*(1), 13-29. Retrieved at http://journal.r-project.org/archive/2013-1/xiang-gubian-suomela-etal.pdf

Zhu, Z., & Stein, M. L. (1999). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference, 134*, 583-603.