METMAGRASPA
Torino 2014

Dynamic Spatial Sampling

M. Bohorquez[1], R. Giraldo[2] and J. Mateu[3]

[1,2] *Department of Statistics, National University of Colombia, Bogota. E-mail:* [1] *mpbohorquezc@unal.edu.co -* [2] *rgiraldoh@unal.edu.co;*
[3] *Department of Mathematics, Universitat Jaume I, Spain. E-mail: mateu@mat.uji.es;*

Abstract. *This paper pretends to give new tools for dynamic spatial sampling designs to find the optimal estimation and the optimal spatial prediction, based on the variation of spatial dependence structure in both cases, discrete and continuous time. In order to model the time series of the spatial covariance parameters, the measurement error and the bias caused by the estimation are included in the formulation of state space models. A discussion of useful properties and techniques to estimation and forecasts in several scenarios is presented. The methodology is applied to a network of quality air in the Bogotá city.*

Keywords. *Environmental Science; Geostatistics; Optimal sampling; Spatial planning*

1 Introduction

A spatio-temporal process is a stochastic process $\{Z(\boldsymbol{s},t) : (\boldsymbol{s},t) \in D_s \times D_t\}$, where $D_s \times D_t$ is the spatio-temporal index set. We note that $D_s \times D_t \subseteq \mathbb{R}^d \times \mathbb{R}$ with $\mathbb{R}^d$ for the spatial bit and $\mathbb{R}$ for the temporal one. When D_s is continuous, the process is called a geostatistical process. Let $Z(\boldsymbol{s}_{i_{t_k}}, t_k)$ be the process at location $\boldsymbol{s}_i$ at time t_k, with $t_k \in D_t$, $k = 1,...,T$, $k \in \mathbb{N}$, $\boldsymbol{s}_{i_{t_k}} \in D_s$ and

$$Z_{S_{t_k}} = (Z(\boldsymbol{s}_{1_{t_k}}, t_k), Z(\boldsymbol{s}_{2_{t_k}}, t_k), ..., Z(\boldsymbol{s}_{n_{t_k}}, t_k))', \quad S_k \subset \mathbb{R}^d \tag{1}$$

the random vector at n_{t_k} spatial locations $S_{t_k} = \{\boldsymbol{s}_{1_{t_k}}, ..., \boldsymbol{s}_{n_{t_k}}\}$ at each time point t_k. Our approach fits models in continuous space and time, based on the observations of (1) which we denote as

$$z_{S_{t_k}} = (z(\boldsymbol{s}_{1_{t_k}}, t_k), ..., z(\boldsymbol{s}_{n_{t_k}}, t_k))' \tag{2}$$

n_{t_k} is allowed to change at each t_k according to any statistical quality criteria or some technical or economical constraint. We assume that while spatial configuration does not change the series are measured at the same point times and therefore, they have the same length, even though it is not a requirement. Our goals are, on the one hand, to determine the set of spatial locations, that ensures the optimal spatial mean estimation $\mu_{t_{T+1}}$ at the future time t_{T+1}, and on the other hand to determine the set $S_{t_{T+1}}$ of spatial locations to optimal spatial prediction of the random vector $\boldsymbol{Z}_{S^0_{t_{T+1}}}$ at some set $S^0_{t_{T+1}}$ of interest places.

An optimal sampling design is the one that finds the best combination predictor-design or estimator-design, according to the optimization of a criterion previously established [4]. We have T spatial vectors $\boldsymbol{z}_{S_{t_1}}, ..., \boldsymbol{z}_{S_{t_T}}$ observed at each set of spatial locations $S_{t_k} = \{\boldsymbol{s}_{1_{t_k}}, ..., \boldsymbol{s}_{n_{t_k}}\},\ k = 1, ..., T$. Now we have to find the set of locations $S_{t_{T+1}} = \{\boldsymbol{s}_{1_{T+1}}, ..., \boldsymbol{s}_{n_{t_{T+1}}}\}$ that optimizes the corresponding unbiased mean estimator $\hat{\mu}_{S_{t_{T+1}}}$ for the spatial process at time point t_{T+1}. Once we have $\hat{\mu}_{S_{t_{T+1}}}$, the next step is to add this new measure to the series to build the vector $\left(\hat{\mu}_{S_{t_1}}, ...\hat{\mu}_{S_{t_T}}, \hat{\mu}_{S_{t_{T+1}}}\right)$ and so on. In this way, we can use updated information for the next design. We are not optimizing temporal points as we are assuming that temporal instants over which the process is observed, have been previously determined, although not necessarily regularly spaced. In addition, we assume that the sampling locations can be changed at the future time t_{T+1}. A similar procedure used for the optimal estimation of the mean of the process, can be used to optimize the prediction of the random vector $\boldsymbol{Z}_{S^0_{t_{T+1}}}$ at the future time t_{T+1} in a set of interest sites. The updated prediction at each place $\boldsymbol{s}^0_{i_{t_{T+1}}}$, $i = 1, ..., n^0_{t_{T+1}}$ is obtained by applying ordinary kriging at time point t_{T+1}. Now, we detail the proposal for the optimal mean estimation.

2 Methodology

2.1 Optimal Spatial Sampling for the Mean Estimation at the future time $T+1$.

To consider the spatial covariance structure and take advantage of the fact that the most common case is to have a large number T of temporal observations in a few locations, we use the time series of the variance of the Generalized Least Squares (GLS) estimator for the spatial mean and find the set of spatial locations that minimizes its forecast at time $T+1$. The GLS estimator $\hat{\mu}_{S_{t_k}}$, for the mean of the spatial process based on the observed vector at time point t_k at the set of locations $S_{t_k} = \{\boldsymbol{s}_{1_{t_k}}, ..., \boldsymbol{s}_{n_{t_k}}\}$, see (2) is given by, $\hat{\mu}_{S_{t_k}} = \boldsymbol{1}'\Sigma^{-1}_{S_{t_k}}\boldsymbol{z}_{S_{t_k}}/\boldsymbol{1}'\Sigma^{-1}_{S_{t_k}}\boldsymbol{1},;\ t_k \in D_t$ and its variance $Var_{S_{t_k}}(\hat{\mu}_{S_{t_k}})$ takes the form

$$Var_{S_{t_k}}(\hat{\mu}_{S_{t_k}}) = \left(\boldsymbol{1}'\Sigma^{-1}_{S_{t_k}}\boldsymbol{1}\right)^{-1} = \left(\sum_{i=1_{t_k}}^{n_{t_k}} \sum_{j=1_{t_k}}^{n_{t_k}} \boldsymbol{C}_{t_k}\left(\boldsymbol{s}_{i_{t_k}}, \boldsymbol{s}_{j_{t_k}}|\Theta_{t_k}\right)\right)^{-1} \tag{3}$$

$\Sigma_{S_{t_k}}$ is the covariance matrix of the spatial process at time t_k at the set of locations S_{t_k}, i.e. $\Sigma_{S_{t_k}} = \left(Cov(Z(\boldsymbol{s}_{i_{t_k}}, t_k), Z(\boldsymbol{s}_{j_{t_k}}, t_k))\right)$, $i, j = 1_{t_k}, ..., n_{t_k}$, and $Cov(Z(\boldsymbol{s}_{i_{t_k}}, t_k), Z(\boldsymbol{s}_{j_{t_k}}, t_k))$ is given by a known valid spatial covariance model $\boldsymbol{C}_{t_k}$ with parameter vector Θ_{t_k}, and we have the following possibilities: (a.) $\boldsymbol{C}\left(\boldsymbol{s}_{i_{t_k}}, \boldsymbol{s}_{j_{t_k}}|\Theta\right)$: $\boldsymbol{C}$ and Θ are the same for all t_k and (b.) $\boldsymbol{C}\left(\boldsymbol{s}_{i_{t_k}}, \boldsymbol{s}_{j_{t_k}}|\Theta_{t_k}\right)$: $\boldsymbol{C}$ is constant but the parameter Θ depends on each t_k, Θ_{t_k}. The case $\boldsymbol{C}_{t_k}\left(\boldsymbol{s}_{i_{t_k}}, \boldsymbol{s}_{j_{t_k}}|\Theta_{t_k}\right)$ when both the model for $\boldsymbol{C}$ and the parameter vector Θ vary with each t_k, corresponds to a very unstable process. A more realistic assumption is that the spatial covariance structure changes but after longer intervals of time. Under spatial second-order stationarity, we have $\boldsymbol{C}(\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}|\Theta)$, $\boldsymbol{C}(\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}|\Theta_{t_k})$ and $\boldsymbol{C}_{t_k}(\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}|\Theta_{t_k})$ respectively.

A natural design at time t_{T+1} goes through the minimization of $arg\min_{S_{t_{T+1}} \subset D'_s} Var_{S_{t_{T+1}}}(\hat{\mu}_{S_{t_{T+1}}})$. This procedure assures that $\hat{\mu}_{S_{t_{T+1}}}$ has minimum variance. Due to the continuity of D_s, it is not possible to evaluate the criterion in all its subsets and besides, it does not make sense to take sites extremely close. So, the criterion only is evaluated in a finite number of available sets in $D'_s \subset D_s$. D'_s must be built according to some knowledge of region conditions, possibility to access and maybe economical criteria. In other case, the best option is the evaluation of the criterion on a fine regular grid. If the pair $(\boldsymbol{C}_{t_k}, \Theta_{t_k})$ does not change for any t_k, there is a unique design of size n_{t_k} for all $t_k \in D_t$. This methodology can also be used when it is not necessary or possible to move all sampling locations but only a few. For example, suppose there is a daily mobile station, and that this station is at location $\boldsymbol{s}_{1_{t_T}}$ at time t_T. The variance of the mean estimation with a new location $\boldsymbol{s}_{1_{t_{T+1}}}$ but keeping $\boldsymbol{s}_{2_{t_T}},...,\boldsymbol{s}_{n_{t_T}}$, is given by

$$\begin{aligned} Var_{S_{t_{T+1}}}(\hat{\mu}_{S_{t_{T+1}}}) &:= \left(\sum_{i_{t_T},j_{t_T}=2_{t_T}}^{n_{t_T}} \boldsymbol{C}_{t_{T+1}}\left(\boldsymbol{s}_{i_{t_T}},\boldsymbol{s}_{j_{t_T}}|\boldsymbol{\Theta}_{t_{T+1}}\right) + \sum_{j_{t_T}=2_{t_T}}^{n_{t_T}} \boldsymbol{C}_{t_{T+1}}\left(\boldsymbol{s}_{1_{t_{T+1}}},\boldsymbol{s}_{j_{t_T}}|\boldsymbol{\Theta}_{t_{T+1}}\right)\right)^{-1} \\ &:= \left(cte + \sum_{j_{t_T}=2_{t_T}}^{n_{t_T}} \boldsymbol{C}_{t_{T+1}}\left(\boldsymbol{s}_{1_{t_{T+1}}},\boldsymbol{s}_{j_{t_T}}|\boldsymbol{\Theta}_{t_{T+1}}\right)\right)^{-1} \end{aligned} \tag{4}$$

where $S_{t_{T+1}} = \{\boldsymbol{s}_{1_{t_{T+1}}},\boldsymbol{s}_{2_{t_T}},...,\boldsymbol{s}_{n_{t_T}}\}$ and $\boldsymbol{s}_{1_{t_{T+1}}} \in D_{s_{1_{T+1}}} \subset D_s$ with $D_{s_{1_{T+1}}}$ the set that contains the N possible locations for $\boldsymbol{s}_{1_{t_{T+1}}}$. Consequently, the design criterion is simplified to select the spatial location $\boldsymbol{s}_{1_{T+1}}$ which minimizes (4), that is, $arg\min_{\boldsymbol{s}_{1_{T+1}} \in D_{s_{1_{T+1}}}} Var_{S_{t_{T+1}}}(\hat{\mu}_{S_{t_{T+1}}})$. Our proposal considers the general case where $\boldsymbol{C}_{t_k}$ depends on t_k and there is the option to move locations at each of the future time points. But, if locations will not be moved until the point time t_{T+m+1}, and in the current locations the measures are going to be taken at times $t_{T+1},...,t_{T+m}$, the criterion can be modified to a global measure such as the total variance $Tr\left((Var(\mu_{t_{T+1}}),...,Var(\mu_{t_{T+m+1}}))\right)$. Note that even if the process is in continuous time, only has sense to include in the optimization the point times with measures. A known time series model or a function, $\theta_t = f(t)$ allows to carry out forecasts $m-step$ ahead of future value $\boldsymbol{\Theta}_{T+1},...,\boldsymbol{\Theta}_{T+m+1}$ and compute any of the criteria considered before.

2.2 Estimation and forecasting of $\hat{V}ar(\hat{\mu}_{T+1})$

In practical cases, the covariance model is unknown and it has to be estimated from the data. In this case the designs are only suboptimal. We propose, first to model the parameters of dependence structure at each time point observed and then model the time series of these parameter estimators, in order to find the forecasts to compute the variance the future time, so we are able to find the optimal spatial configuration that optimize $\hat{\mu}_{t_{T+1}}$. We suppose the same covariance model for $k = 1,...,T,T+1$ time points, except for the parameter vector $\boldsymbol{\Theta}_{t_k}$, $\boldsymbol{C}(\boldsymbol{s}_{i_{t_1}} - \boldsymbol{s}_{j_{t_1}}|\boldsymbol{\Theta}_{t_1}),...,\boldsymbol{C}(\boldsymbol{s}_{i_{t_T}} - \boldsymbol{s}_{j_{t_T}}|\boldsymbol{\Theta}_{t_T})$. The methods for estimation of $\boldsymbol{C}(.)$ are the same for any of the goals, estimation or prediction. Several options have been proposed such as maximum likelihood, least squares and composite likelihood. In addition, the dependence of the parameter vector of spatial covariance model on t can be in either two ways: stochastic or deterministic. The forecasting method depends if the process is in discrete or continuous time, and the use of Box-Jenkins ARIMA models or state space models. The analysis of time series observed at irregular points can be handled very efficiently with continuous time models. As these time series are built with estimations of covariance parameters, we must consider that data contains measurements errors and bias, although we expect that the noise-to-signal ratio be small. In order to involve the measurement error in parameter estimation of the time-series models, we use the approach given by [3], that details the estimation procedure and the required constraints to ensure identifiability. In this paper, we assume that measurement error is an additive white noise process, so, the identifiability is ensured. Nevertheless, in presence of additive white noise, the forecast obtained with stationary and invertible ARMA models are strongly affected only when the parameters are near to the unit circle, in other case the influence of measurement error on forecast is very small. So, in the cases where measurement error affects notoriously the forecast, the best approach is modelling through the state space models to extract the signal to generate forecasts.

2.3 Real Data Analysis

We analyze network data for air quality in Bogotá city. The data correspond to consecutive hours from May 13, 2013 at 1:00 a.m. to May 13, 2014 at 12:00 a.m. There are 10 stations that monitor hourly particulate matter up to 10 micrometers in size (PM10), stations 1:10 and one of them is a mobile air quality monitor, station 3Mb, see figure 1a. There are considerable differences between the sectors of the city, due to traffic and industries. There is evidence of non-constant variance, so we apply to the observations the Box-Cox transformation with $\lambda = -0.1$ and we use median polish to model the trend. We fit a generalized Cauchy covariance, that is a very flexible model and is given by

$$C(\boldsymbol{s}_i - \boldsymbol{s}_j) = \sigma^2 \left(1 + \left(\frac{|\boldsymbol{s}_i - \boldsymbol{s}_j|}{b}\right)^{\gamma}\right)^{\nu} \quad b > 0,\ 0 < \gamma \leq 2,\ \nu > 0 \tag{5}$$

where b is a scale parameter, γ is a shape parameter and ν parameterizes the long-memory dependence. From the empirical variograms in the figure 1b., we consider that there is no reason to suppose discontinuity at the origin, since there is no jump in $|\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}| = 0$. So, we keep the nugget parameter fixed and equal to zero. Regarding the selection of the covariance model, at first, we choose the parameter model $\gamma = 2$ because of the shape of the short-lag terms of empirical variograms, then this parameter is held fixed and we run several maximizations on a grid of ν values, and the estimation is restricted to the parameters σ^2 and b at each time point. The reason for this procedure is that the estimation of γ and ν, at least here, cause some numerical instability. So, the fitted model for spatial covariance at each time t_k, $k = 1, \ldots, 73$ is the particular case of (5),

$$C(\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}) = \sigma_t^2 \left(1 + \left(\frac{\|\boldsymbol{s}_{i_{t_k}} - \boldsymbol{s}_{j_{t_k}}\|}{b_t}\right)^2\right)^{-3/2} \tag{6}$$

The model used for time series σ_t^2 is a continuous autoregressive process, $CAR(1)$. As the process is sampled at equally spaced intervals of length one hour and in presence of measurement error, the state space model, turns in to an $ARMA(1,1)$ model with parameter estimated $\hat{\Theta} = (0.8940, 0.4562)$. The time series of b_t has constant mean and there is no significative autocorrelation, so the forecasting for t_{T+1} is the mean 7.012243. The red points in the map, figure 1a. are the optimal spatial locations for the next 7 days of the mobile station.

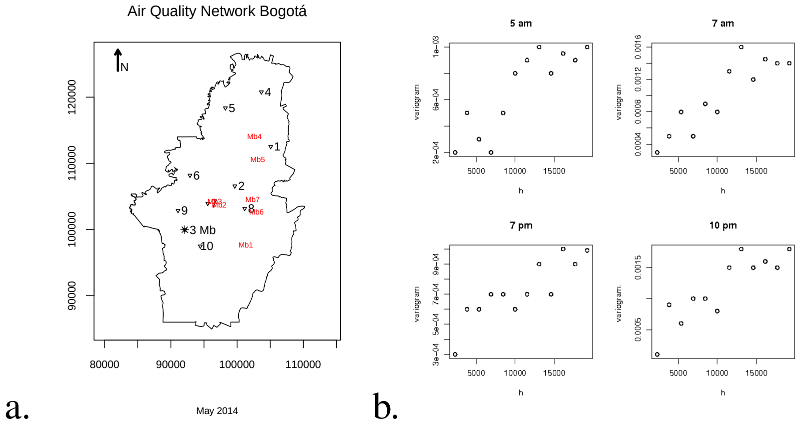

Figure 1: a. Air quality, Bogota, 2013–2014. b. Some semivariograms obtained on May 13, 2014

References

[1] Angulo, J., Bueso, M., and Alonso, F. (2000). A study on sampling design for optimal prediction of space–time stochastic processes *Stochastic Environmental Research and Risk Assessment, Springer* **14**, 412–427.

[2] Durbin, J., and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press. New York.

[3] Lee, H., and Shin, W. D. (1997) Maximum likelihood estimation for arma models in the presence of ARMA errors *Communications in Statistics - Theory and Methods* **26**, 1057–1072.

[4] Müller, W. (2007) *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Springer Verlag. Berlin.

[5] Wikle, C. K., and Royle, J. A. (1999) Space: Time Dynamic Design of Environmental Monitoring Networks. *Journal of Agricultural, Biological, and Environmental Statistics, International Biometric Society* **4**, 489–507.

[6] Zimmerman, D. (2006) Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction *Environmetrics, Wiley Online Library* **17**, 635–652.