SPATIAL DESIGN

Werner G. Müller
Department of Statistics, Vienna University of Economics, Austria

Keywords: Optimum design, random field, space-filling, covariogram estimation, entropy sampling.

Contents

1. Introduction
2. A statistical framework
3. Single purpose spatial designs
3.1. Optimum Designs for Trend Estimation (Uncorrelated Processes)
3.2 Exploratory Designs
3.3 Optimum Designs for Trend Estimation (Correlated Processes)
3.4 Optimum Designs for Spatial Prediction
3.5 Optimum Designs for Covariogram Estimation
4. Multipurpose spatial designs
4.1 Constrained and Compound Designs
4.2 Entropy Sampling
5. Relationships among design criteria
6. Conclusions and outlook
Acknowledgements
Glossary
Bibliography
Biographical Sketch

Summary

The aim of an investigator allocating spatial samples must be to gather the maximum information possible for her/his aims of analysis. In a model-oriented framework it is reasonable to guide this venture by employing principles of optimum design theory, a choice motivated in Section 2. In Section 3.1 the classical uncorrelated regression model is reviewed, followed by brief descriptions of cases when little is known about the form of the model (Section 3.2 exploratory designs) and the (for the usual spatial applications) very important correlated regression case (Section 3.3). Designs for spatial prediction and covariogram estimation conclude this section. Section 4 is devoted to the problem of how to combine designs for trend and covariogram estimation most properly and Section 5 reveals relationships among various methods. The paper is accompanied by an illustrative example: the redesign of a water-quality network.

1. Introduction

Spatial data occur in many fields such as agriculture, geology, environmental sciences, and economics. They have been recorded and analyzed probably as early as men started to make maps, however the origins of their statistical analysis as we understand it today must probably be attributed to the work of Matheron. Spatial data has the distinctive

characteristic that, attached to every observation, one has a set of coordinates that identifies the (geographical) position of a respective data collection site. The set of locations of those data collection sites (the so-called design) influences decisively the quality of the results of the statistical analysis. Usually in choosing the design the aim is to ensure continuous monitoring of a data generating process or to allow for point prediction of present or future states of nature.

Sampling theory and *optimum experimental design theory* are two large branches in theoretical statistics that have developed separately, though with considerable theoretical overlap, both of them providing methods for efficient site positioning. Whereas *sampling theory* is a basically model-free methodology essentially oriented towards restoring unobserved data, in *optimum design theory* the aim is to estimate the structure of the data generating process, e.g. the parameters of an assumed (regression) model or functions of these parameters.

In this article emphasis is on the latter branch but divergences and parallels between the two branches are pointed out whenever necessary. Principles from *optimum design theory* will be adhered to in presenting and developing methods specific for the solution of the spatial design problem, whereas complementary reviews of the model-free approaches to spatial design can be found in many other sources. Merging the terminologies from the different fields is unavoidable, which can be achieved by looking at the material from a random field perspective. It should be pointed out that the model-based approaches presented in this article are usually not very robust with respect to changes in the model assumptions. Thus in case one does not have a clear a priori picture about those assumptions the use of sampling methods - though in general not as efficient - might be the safer alternative.

Illustrative example: the redesign of a water-quality monitoring network.

The Südliche Tullnerfeld is a part of the Danube river basin in central Lower Austria and due to its homogeneous aquifer well suited for a model-oriented geostatistical analysis. It contains 36 official water quality measurement stations, which are irregularly spread over the region. A graphical representation of the sites (after rescaling to a unit of approximately 31 kms) on a 95 times 95 grid is given in Figure 1. The data set used in some calculations contains daily averages of chloride (Cl) concentrations in mg/l over the period 1992-1997 on all time points, for which at least one measurement was taken.

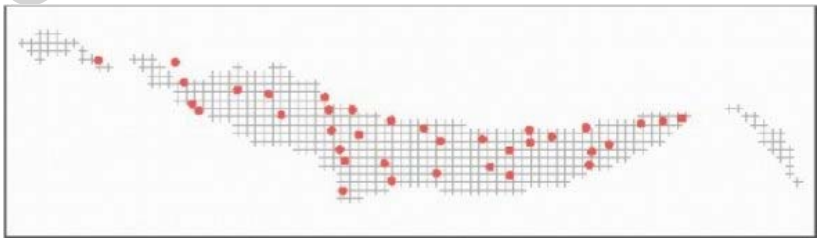

Figure 1: The water quality monitoring network in the Südliche Tullnerfeld; solid circles represent sites, grid points represent region.

2. A Statistical Framework

Principally, there are two perspectives on the generation of (spatial) data: it is considered to be either completely deterministic or it contains a stochastic component. Concentrating on the latter point of view the data y observed at coordinates $x \in \Re^2$ is regarded as being generated by a parameterized process (random field)
$y(x) = \eta(x,\beta) + \varepsilon(x)$,
where η denotes the deterministic and ε the stochastic part. Also $E[\varepsilon(x)\varepsilon(z)] = c(x,z;\theta)$ is assumed to follow a parameterized covariance structure. A common interpretation is that the long range variability (mean, trend) in the field is generated by η, whereas the short range fluctuations are due to ε. Thus they will be denoted as first and second order characteristics of the spatial process respectively.

It is evident (also from the analogy to the similar problem in time-series analysis) that there is no clear-cut distinction between first and second order characteristics. There exists an inherent impossibility to distinguish unequivocally the influences of local trends and spatial correlation. The more detailed the model for the trend is, the less of the systematic fluctuations of the field has to be ascribed to the spatial covariance model. The distinction is therefore a purely practical one and will be largely based upon interpretability and parsimony of the posited models.

In the stochastic framework the main purpose of statistical analysis is then the estimation of the parameters β and θ, identification of special events (such as appearances of maxima or minima), and prediction of y based upon these parametric models. When it comes to the question of where to position observation sites most efficiently, the natural consequence of the model-oriented point of view is the application of results from *optimum design theory*, the main concepts of which are briefly reviewed for a better understanding of the subsequent sections, where a special emphasis is put on the peculiarities of the spatial setting.

Contrasted to it can be the data-oriented (model-free) point of view: when $y(x)$ is considered to be deterministic, the aim is either interpolation to restore unobserved data or the calculation of overall field characteristics such as the mean. Adherents of this approach are better served by *sampling theory*, the applications of which for spatial problems can be found in numerous publications. Note, that the conclusions that can be reached under the two points of view can sometimes be confusing and contradicting. In restoring $y(\cdot)$, i.e. making predictions $\hat{y}(\cdot)$, from the model it may be advantageous to have correlated errors and one therefore needs less observations to reach a given precision than in the uncorrelated case. On the other hand correlated observations carry redundant information about η and one thus needs more of them to estimate the model as precisely as from uncorrelated observations.

Another distinction that needs to be made and has impact on the choice of techniques is whether one is primarily interested in first or second order characteristics of the observed field. In an applied study usually both is of interest and there remains the question of how to combine respective methods of design.

3. Single Purpose Spatial Designs

3.1. Optimum Designs for Trend Estimation (Uncorrelated Processes)

How can the information (on the parameter β) due to a design ξ be formally described in a real experiment? Assume the simplified (uncorrelated) version of the random field, i.e.

$$c(x,z;\theta) = \delta_{x,z}\ \sigma^2(x),$$

is considered. In the vicinity of the true β a linearization of η may be sufficiently exact and thus (almost) all the information about β (at least with respect to estimating linear functionals of it) from the experiment is contained in its so-called (standardized or average) Fisher information matrix

$$M(\xi,\beta) = \Sigma_x \partial\eta(x,\beta)\ \partial\eta^T(x,\beta)\xi(x),$$

where the vector $\partial\eta(x,\beta)$ contains the partial derivatives of η with respect to β evaluated at a prior guess for β (in the following the argument is dropped for simplicity) for every design point x and the measure $\xi(x)$ represents the design (here usually proportional to the number of observations to be taken at each site, so for an n-point design one may write ξ_n instead).

Now a certain criterion $\Phi[M]$ (e.g. the determinant of M leading to what is known as D-optimality) is optimized. Mostly this is done by employing numerical algorithms that come sufficiently close to the solution. The simplest of which consists of sequential addition of design points, maximizing a so-called sensitivity function $\phi(x,\xi_n)$, that is linked to the design criterion by an equivalence theorem (cf. Kiefer's celebrated pioneering work). E.g. for the D-criterion this sensitivity function is simply the ratio of the prediction variance with the estimated parameters to the prediction variance with the true parameters

$$\phi(x,\xi_n) = \sigma^{-2}(x)\ \mathrm{E}[(y(x) - \hat{y}(x))^2].$$

Evidently, minimization of this function can be used for thinning out existing networks, and an exchange type procedure offers a fruitful generalization to other purposes.

Illustrative example (continued).

In Figure 2 a D-optimum design ξ_{36} for a linear trend, i.e. $\eta(x,\beta) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, and the area of the Südliche Tullnerfeld is displayed. It is clear that most of the sites cluster at a few hot spots, which makes the design vulnerable with respect to changes in the assumptions.

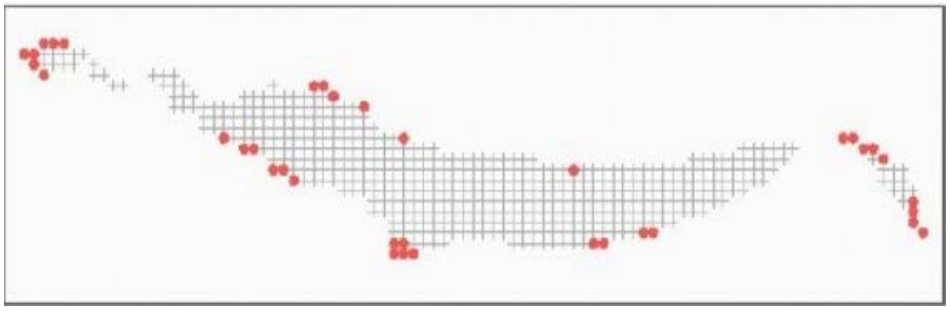

Figure 2: A D-optimum design for the Südliche Tullnerfeld.

-
-
-

Bibliography

Aldworth, J. and Cressie, N. (1999). Sampling designs and prediction methods for Gaussian spatial processes. *Multivariate Analysis, Design of Experiments, and Survey Sampling* (eds. S. Ghosh and J.N. Srivastava), 1—54, Dekker, New York. [A survey and an extensive simulation study on spatial sampling].

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Kluwer Academic Publishers, Dordrecht. [A comprehensive review of techniques for collecting spatial data from the sampling perspective].

Atkinson, A.C. and Fedorov, V.V. (1989). Optimum design of experiments. *Encyclopedia of Statistical Sciences*, supplement volume (eds. S. Kotz, C.B. Read, and N.L. Johnson), 107—114, Wiley, New York. [A brief exposition of the basics in optimal design theory].

Bellhouse, D.R. (1989). Spatial sampling. *Encyclopedia of Statistical Sciences*, supplement volume (eds. S. Kotz, C.B. Read, and N.L. Johnson), 581—584, Wiley, New York. [A brief exposition of the basics in sampling theory].

Bueso, M.C., Angula, J.M., Cruz-Sanjulián, J. and García-Aróstegui, J.L. (1999). Optimal Spatial Sampling Design in a Multivariate Framework. *Mathematical Geology*, 31(5):507—525. [A suggestion on how to extend existing approaches to the multivariate case].

Cressie, N. (1993). *Statistics for Spatial Data*, revised edition. John Wiley & Sons, New York. [Encyclopedic exposition of spatial statistics with some sections on design].

Cox, D.D., Cox, L.H. and Ensore, K.B. (1997). Spatial sampling and the environment: some issues and directions. *Environmental and Ecological Statistics*, 4:219—233. [A recent survey better on directions than on issues; with links to applications].

Fienberg, S.E. and Tanur, J.M. (1987). Experimental and sampling structures: Parallels diverging and meeting. *International Statistical Review*, 55:75—96. [A general discussion of the relationships between model- and sampling-based approaches].

Goos, P., Tack, L. and Vandebroek, M. (2001). Optimal designs for variance function estimation using sample variances. *Journal of Statistical Planning and Inference*, 92:233—252. [A recent paper with details on a design construction algorithm for the variance function estimation case].

John, P.W.M., Johnson, M.E., Moore, L.M. and Ylvisaker, D. (1995). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 44, 249—263. [Paper linking exploratory and optimum designs].

Kiefer, J. (1959). Optimal experimental designs (with discussion). *Journal of the Royal Statistical Society, Series B*, 272—319. [The at the time controversial pioneering paper in optimum design theory; reprinted in Kiefer, J. (1985). *Collected Papers*, Springer, New York].

Lu, Z.-Q., Berliner, L.M. and Snyder, C. (2000). Experimental design for spatial and adaptive observations. *Studies in the Atmospheric Sciences* (eds. L.M. Berliner, D. Nychka, and T. Hoar), Springer Lecture Notes in Statistics 144. [A recent article with emphasis on weather prediction application].

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246—1266. [Where it all began.]

Müller, W.G. (2000). *Collecting Spatial Data*, revised edition. Physica Verlag, Heidelberg. [The author's attempt of a more detailed survey on the applicability of design methods in spatial statistics].

Näther, W. (1985). *Effective Observation of Random Fields*. Teubner Texte zur Mathematik - Band 72. Teubner Verlag, Leipzig. [A pioneering monograph with a strong design flavor].

Pàzman, A. (2001). Statistical experiment and its optimum design. *EOLSS*. [A recent survey of design theory which should be readily available].

Ucinski, D. (1999). *Measurement Optimization for Parameter Estimation in Distributed Systems.* Technical University Press, Zielona Gora. [A monograph oriented to the design problem in the case of moving sensors].

Biographical Sketch

Werner G. Müller is associate professor at the Department of Statistics, Vienna University of Economics and Business Administration, Austria, where he teaches statistics and econometrics. It is also in these areas where his research interests focus, particularly on the topics of panel regression models, optimum experimental design and spatial statistics. His monograph "Collecting Spatial Data" (2001, Physica Verlag) bridges the gap between the latter two fields and proposes a number of techniques applied to air and water quality monitoring problems. Professor Müller has lived, taught and conducted research in Russia, Sweden and the United States.