Collecting Spatial Data: Design and Sampling

Werner G. Müller

Department of Applied Statistics

Based on joint work with M. Hainy, H. P. Wynn, H. Wagner and H. Waldl
Spatial Data

• has the distinctive characteristic that, attached to every observation, we have 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.
Two Approaches for (Spatial) Data Acquisition

- **probability-based (design-based):** basically model-free methodology essentially oriented towards restoring unobserved data or quantifying population totals
  — *sampling theory*

- **model-based:** 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.
  — *optimum experimental design theory*
Connections

• Through the so-called superpopulation approach assuming:

$$y_i = \eta(x_i, \beta) + \varepsilon_i$$

• Usually aim is to balance the sampling design $\xi$ for keeping the anticipated mean-squared error low, see e.g. Benedetti et al., IS Review (2017).

$$E_{\xi}E_{\eta}(\hat{y} - y)^2$$

• Purposive sampling allows explicit minimization of a (design) criterion $\Phi$, see Wang et al., Spatial Statistics (2012) for a comparison.
Optimal Designs (for estimation)

Classical: select the inputs (and weights)

\[ \xi_N = \left\{ \begin{array}{l} p_1, p_2, \ldots, p_n \\ z_1, z_2, \ldots, z_n \end{array} \right\} \]

such that a prespecified criterion

\[ \max_{z_i, p_i} \Phi \left[ M \left( \xi_N \right) \right] \]

is optimized.

Well developed theory for standard (uncorrelated) regression based on Kiefer’s (1959) concept of design measures.
Design criterion motivated from Bayesian learning

\[ U(\xi) = E_{y|\xi} \Phi[\pi(\theta | y, \xi)] \]
\[ = \int_{y \in Y} \Phi[\pi(\theta | y, \xi)] \pi(y | \xi) dy \]

cf. Ferreira & Gamerman, Bayesian Analysis (2015),
is MC-estimated by

\[ \hat{U}(\xi) = \frac{1}{G} \sum_{i=1}^{G} \hat{\Phi}[\pi(\theta | y^{(i)}, \xi)] \]

Estimation of \( \Phi \) if likelihood is intractable:

**Approximate Bayesian Computation (ABC)**

(Marjoram et al. 2003).
Approximate Bayesian Computation (ABC)

- simulate $G$ data sets $\hat{Y}_i$ by a model with parameters $\theta_i$ drawn from a prior;
- Compare your simulated data with the (true) data $Y$ via a distance measure $d(Y, \hat{Y}_i)$;
- reject parameter $\theta_i$, which generate $d(Y, \hat{Y}_i) > \varepsilon$;
- the frequency distribution of the retained parameter approximates the posterior.
Approximate Bayesian Computation Design (ABCD)

- treat the sampling design $\xi$ as high-dimensional parameter (Simulation Based Bayesian Design, Müller (1999));
- in an outer loop generate $H$ “real” data sets $y(\xi)$;
- use ABC on $(\theta, \xi)$.

Advantages: (almost) any design criterion based on the posteriori is possible, evaluation of likelihood not necessary.

Example 1: Spatial Extremes

Data: maximum summer temperatures from 1895 to 2009 measured on 39 sites in the U.S. midwest region (115 observations per site), see Erhardt and Smith (2014).

Details in Hainy, Müller, Wagner, SERRA (2016)
Example 1 ctd.

Assuming a Schlather model, using tripletwise extremal coefficients as summary statistics, $G=H=1000$, ABC sample used as empirical prior.
Assuming a Schlather model, using tripletwise extremal coefficients as summary statistics, $G=H=1000$, ABC sample used as empirical prior, 2 initial points fixed.
Example 1 ctd.

Design Augmentation $n=3$ to $n=4$. 
Example 1 ctd.

Design Augmentation $n=4$ to $n=5$. 
References 1


• Hainy, M., Müller, W. G., Wynn, H. P., 2014. Learning functions and approximate bayesian computation design: ABCD. *Entropy* 16 (8), 4353-4374.


Two Modeling Paradigms

- **Gaussian Markov random fields**: discretely indexed; usually aggregate observations attached to lattices
  — *Spatial econometrics*

- **Gaussian fields**: continuously indexed; observing spatially varying processes
  — *Geostatistics*
Discretely indexed models

**Spatial Autoregressive (Error) model:**
Spatial dependence in the error term \( E[u_i,u_j] \neq 0 \)
\[
y = X\beta + u \quad \text{and} \quad u = \rho Wu + \varepsilon
\]
\[
y = \rho Wy + X\beta - \rho WX\beta + \varepsilon
\]
Spatial Lag model plus spatially lagged exogenous variables \( WX \).

VC matrix of the errors depends on the spatial process, e.g. SAR process: \( E[uu'] = \Omega(\rho) = \sigma^2[(I- \rho W)'(I- \rho W)]^{-1} \) or CAR process: \( E[uu'] = \Omega(\rho) = \sigma^2[I- \rho W]^{-1} \)

\( \varepsilon \) ... i.i.d. error term (mean = 0, variance = \( \sigma^2 \))
\( \rho \) ... spatial autoregressive parameter
\( W \) ... spatial weight matrix (standardized)
Kriging: data $y$ is observed at coordinates $z \in Z \subset \mathbb{R}$ as being generated by a random field:

$$y(z) = \eta(x(z), \beta) + \varepsilon(z)$$

Noise $\varepsilon$ is usually assumed to have zero mean, finite variances and a parametrized covariance

$$E[\varepsilon(z)\varepsilon(z')] = c(z, z'; \theta) = c(d, \theta).$$
An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach

Finn Lindgren and Håvard Rue
Norwegian University of Science and Technology, Trondheim, Norway

and Johan Lindström
Lund University, Sweden

[Read before The Royal Statistical Society at a meeting organized by the Research Section on Wednesday, March 16th, 2011, Professor D. M. Titterington in the Chair]
Example 2: Leukaemia Survival (NW England)


Response: count of survivals;

some prior knowledge – given error VC and estimate of $\rho=0.24$;

**aim: to detect and quantify relationships between spatial design criteria.**

Choose 3 out of 24 districts.
Discussion by M & Waldl, 2011
References 2