2022 SISMID Module 9
Lecture 8: Mapping for Point Data

Jon Wakefield and Lance Waller

Departments of Statistics and Biostatistics
University of Washington

Overview

Overview of Point Data Mapping

We now suppose that, rather than having counts arising from areas, the responses are measured at spatial points.

The conceptual model is that the data could be measured anywhere, but we have observations from a set of sampled locations.

It is not the locations of the observations that is of interest, but rather construction of the underlying surface (disease mapping, on which we focus), or examining associations between the response and covariates (in a spatial regression setting).

The sampling of locations may be:

- via random sampling – for example we could take a simple random sample of cases and controls (or a stratified version, e.g., by urban/rural) and determine their residential addresses, or
- deterministic – for example, we may sample at pre-specified locations (e.g., on a grid) in air, water or soil.

Overview of Point Data Mapping

Point data may consist of:

- Binary disease indicators, e.g. disease status.
- Multiple level disease indicators, e.g. strains of a particular infectious disease, or cause of under-5 death.
- Continuous measurements, e.g. pollutant concentrations.

We will not consider multiple level data[1], but the models we describe for binary and continuous data can be extended to this case, by an appropriate choice of (multinomial) likelihood.

For both discrete and continuous measurements we may wish to visually assess spatial dependence by examining the residual response surface, i.e., after adjusting for covariates.

[1] For an application to multinomial data, see Tarr *et al.* (2018); in this paper spatial surfaces were examined for three different lineages of Escherichia coli O157:H7 infections

Point Data for Continuous Outcomes

Modeling Spatial Dependence for Continuous Data

We first consider continuous outcomes: these are of interest in their own right, but also form a stepping stone to the binary data case.

We illustrate methods using the famous Meuse data set, which gives locations and top soil heavy metal concentrations (in ppm), along with a number of soil and landscape variables, collected in a flood plain of the river Meuse, near the village Stein in the South of the Netherlands.

Heavy metal concentrations are bulk sampled from an area approximately 1.5km (north-south) by 2.5km (east-west); there are 155 measurements.

We may be interest in modeling the concentrations as a function of covariates (spatial regression), or predict the concentration surface, which requires modeling the residual surface (mapping).

We will model and predict a surface for zinc concentration; the concentrations at the sampling sites are mapped in Figure 1.

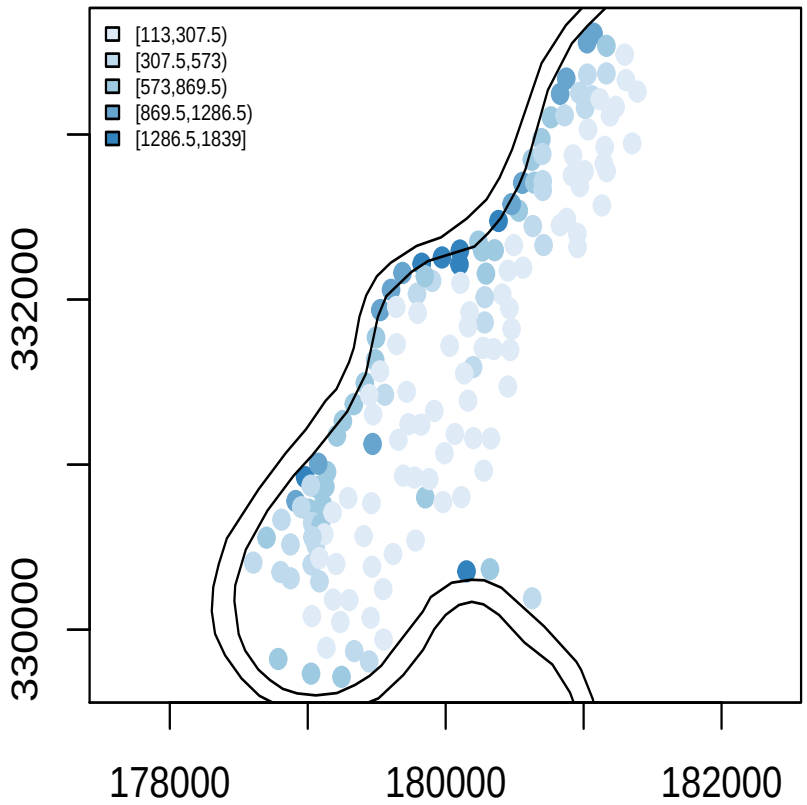

Figure 1: Levels of zinc at 155 sampling sites.

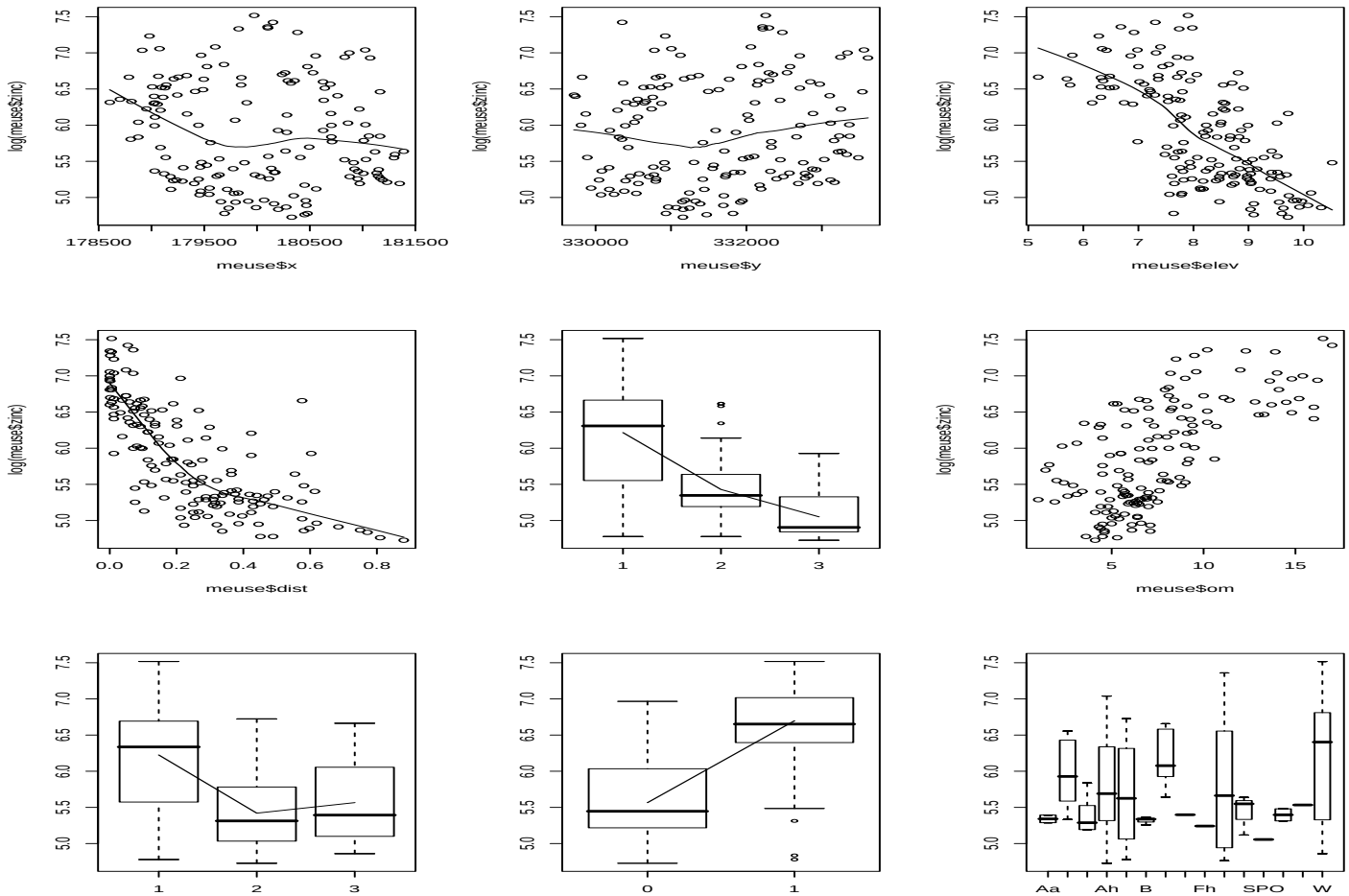

Figure 2: Log zinc versus potential explanatory variables: *x*-coord, *y*-coord, elevation, distance from river, soil type, organic matter, flooding frequency class, lime class, landuse class.

Modeling Spatial Dependence for Continuous Data

For continuous responses $Y_i = Y(\boldsymbol{s}_i)$ (which may have been transformed to produce data that are approximately normally distributed) measured at location $\boldsymbol{s}_i$ a plausible starting model is:

$$Y(\boldsymbol{s}_i) = \beta_0 + \boldsymbol{x}(\boldsymbol{s}_i)\boldsymbol{\beta}_1 + S(\boldsymbol{s}_i) + \epsilon_i,$$

where $S(\boldsymbol{s}_i)$ are terms with spatial structure, ϵ_i are independent random error terms at location $\boldsymbol{s}_i$, for $i = 1, \ldots, n$.

In terms of the spatial component, we might either assume that $S(\boldsymbol{s}_i)$ is

- deterministic, for example, a spline model in space, or
- stochastic, so that the collection $S(\boldsymbol{s}_i)$, $i = 1, \ldots,$, are a collection of random variables.

Modeling Spatial Dependence for Continuous Data

Expressing the model in this way we may view $Y(\boldsymbol{s})$ as a process.

This is useful conceptually and allows us to think about:

- Inference for the parameters of the model, based on Y_i, $i = 1, \ldots, n$.
- Prediction of the observable response (as opposed to the mean response) at unobserved locations $\boldsymbol{s}$.

Modeling Spatial Dependence for Continuous Data

A possible model for the error terms is:

$$\begin{aligned} \boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n] &\sim \mathrm{N}_n(0, \boldsymbol{I}\sigma_\epsilon^2) \\ \boldsymbol{S} = [S(\boldsymbol{s}_1), \ldots, S(\boldsymbol{s}_n)] = [S_1, \ldots, S_n] &\sim \mathrm{N}_n(\boldsymbol{0}, \boldsymbol{\Sigma}) \end{aligned}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix, and $\boldsymbol{\Sigma}$ is an $n \times n$ variance-covariance matrix containing diagonal elements (variances)

$$\Sigma_{ii} = \sigma_s^2$$

and off-diagonal elements Σ_{ij}.

For example, we may assume the exponential covariance model:

$$\Sigma_{ij} = \sigma_s^2 \rho^{d_{ij}} = \sigma_s^2 \exp(-\phi d_{ij}) \tag{1}$$

where $d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$ is the distance between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

This is an example of a spatial Gaussian process (GP) is a model for data $S(\boldsymbol{s})$ such that any collection of points $S(\boldsymbol{s}_i)$, $i = 1, \ldots, n$, follows a multivariate normal distribution.

The Semi-Variogram

A key element of the model is the form of the spatial dependence, but how can we gain clues to the correct form/examine the appropriateness of a particular model?

We recap on the semi-variogram, as introduced in the time series section.

A method for assessing the form and extent of the spatial dependence is via examination of the variogram.

Consider a stochastic process $Y(\boldsymbol{s})$ and let

$$\gamma(\boldsymbol{s}, \boldsymbol{s}') = \text{cov}\{Y(\boldsymbol{s}), Y(\boldsymbol{s}')\} = \text{E}[\{Y(\boldsymbol{s}) - \mu(\boldsymbol{s})\}\{Y(\boldsymbol{s}') - \mu(\boldsymbol{s}')\}],$$

denote the autocovariance function of $Y(\boldsymbol{s})$.

The Semi-Variogram

We write

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + e(\boldsymbol{s}),$$

where $\mu(\boldsymbol{s})$ is the deterministic component and $e(\boldsymbol{s})$ the stochastic component.

Definition: A process $e(\boldsymbol{s})$ is second-order stationary if E$[e(\boldsymbol{s})]$ is constant, for all $\boldsymbol{s}$, and $\gamma(\boldsymbol{s}, \boldsymbol{s}')$ depends only on $\boldsymbol{s} - \boldsymbol{s}'$.

If, further, $\gamma(\boldsymbol{s}, \boldsymbol{s}')$ depends only on $||\boldsymbol{s} - \boldsymbol{s}'||$, then the process is called isotropic, otherwise it is anisotropic.

For a residual process any non-zero constant has been absorbed into $\mu(\boldsymbol{s})$.

The Autocorrelation Function

There is a fundamental difficulty with trying to decompose $Y(\boldsymbol{s})$ into the trend and the stochastic component in a single series because the two are unidentifiable without further assumptions.

Bartlett (1964) showed that a a process with constant intensity and clusters of points is not distinguishable from a process with independent points from a randomly varying intensity (an example of a Cox process).

Is it serial dependence in the residuals, or a high-order polynomial trend for example?

The Autocorrelation Function

Since we have no replicates we need to assume some form of similarity of the data generating process across the study region.

For a second-order stationary random process, and letting $\boldsymbol{h}$ represent a 2D vector, the autocovariance function is

$$\text{cov}\{Y(\boldsymbol{s}), Y(\boldsymbol{s}+\boldsymbol{h})\} = C(\boldsymbol{h}) = \text{cov}\{e(\boldsymbol{s}), e(\boldsymbol{s}+\boldsymbol{h})\},$$

so that $C(0)$ is the variance of $Y(\boldsymbol{s})$ for all $\boldsymbol{s}$.

The autocorrelation function is defined as

$$\rho(\boldsymbol{h}) = \frac{C(\boldsymbol{h})}{C(0)}.$$

The Autocorrelation Function

The semi-variogram is defined, for a process $e(\boldsymbol{s})$ and $\boldsymbol{h} \geq 0$ by

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\text{var}\left[e(\boldsymbol{s}) - e(\boldsymbol{s}+\boldsymbol{h})\right] = \frac{1}{2}\text{E}\left[\{e(\boldsymbol{s}) - e(\boldsymbol{s}+\boldsymbol{h})\}^2\right].$$

If the process displays dependence then if $\boldsymbol{s}$ and $\boldsymbol{s}'$ are 'close' the variables measured at these points are likely to be positively correlated, and so will tend to be similar, so that the variance of the difference is relatively small.

The Autocorrelation Function

Recall that for a second-order stationary process, $\mathrm{E}[e(\boldsymbol{s})] = 0$ for all $\boldsymbol{s}$ and $\mathrm{cov}\{e(\boldsymbol{s}), e(\boldsymbol{s}+\boldsymbol{h})\}$ only depends on $\boldsymbol{h}$ (which implies constant variance).

For a second-order stationary smooth process

$$\begin{aligned}\gamma(\boldsymbol{h}) &= \frac{1}{2}\left\{\mathrm{E}[e(\boldsymbol{s})^2] + \mathrm{E}[e(\boldsymbol{s}+\boldsymbol{h})^2] - 2\mathrm{E}[e(\boldsymbol{s})e(\boldsymbol{s}+\boldsymbol{h})]\right\} \\ &= \sigma_e^2\{1-\rho(\boldsymbol{h})\},\end{aligned}$$

where $\mathrm{var}(e) = \sigma_e^2$.

Note that here the residuals $e(\boldsymbol{s})$ are the spatial terms we labeled $S(\boldsymbol{s})$ in earlier models.

The Autocorrelation Function

The semi-variogram is also well-defined for an intrinsically stationary process for which $\mathrm{E}[e(\boldsymbol{s})] = 0$ and for which

$$\mathrm{E}[(e(\boldsymbol{s}) - e(\boldsymbol{s}+\boldsymbol{h}))^2] = 2\gamma(\boldsymbol{h}).$$

As $\boldsymbol{h}$ increases then for observatons far apart in space

$$\gamma(\boldsymbol{h}) \to \mathrm{var}\,[e(\boldsymbol{t})] = \sigma_e^2,$$

which (recall) is assumed constant.

Variograms

Recall the semi-variogram is given by

$$\gamma(\boldsymbol{h}) = \frac{1}{2}\text{var}\left[\, e(\boldsymbol{s}) - e(\boldsymbol{s}+\boldsymbol{h}) \,\right].$$

If there is spatial dependence then points close together will tend to be similar and so the variance of the pairwise difference will be small, but will increase as the distance increases.

At some distance the points will become independent, and the semi-variogram will take on the variance of the process, this is called the sill, and the distance at which this occurs, the range.

Variograms

Often we would like to include an additional variance for "measurement error".

Let

$$e(\boldsymbol{s}) = S(\boldsymbol{s}) + \epsilon,$$

with the ϵ independent error terms with

$$\text{var}(\epsilon) = \sigma_\epsilon^2$$

and $S(\boldsymbol{s})$ spatially dependent error terms with

$$\text{var}[S(\boldsymbol{s})] = \sigma_s^2$$

and

$$\text{cov}[S(\boldsymbol{s}), S(\boldsymbol{s} + \boldsymbol{h})] = \sigma_s^2 \rho(\boldsymbol{h}).$$

Variograms

In this case the form of the variogram is

$$\begin{aligned}\gamma(\boldsymbol{h}) &= \frac{1}{2}\left\{\mathrm{E}[e(\boldsymbol{s})^2] + \mathrm{E}[e(\boldsymbol{s}+\boldsymbol{h})^2] - 2\mathrm{E}[e(\boldsymbol{s})e(\boldsymbol{s}+\boldsymbol{h})]\right\} \\ &= \sigma_\epsilon^2 + \sigma_s^2\{1-\rho(\boldsymbol{h})\},\end{aligned}$$

where $\mathrm{var}(e) = \sigma_\epsilon^2 + \sigma_s^2$.

The value of the semi-variogram close to the origin is of particular interest, since it determines the degree of smoothness of the process.

A discontinuity at the origin is called the nugget effect, and is often attributable to measurement error or small-scale spatial variation (we never directly observe dependence at a distance less than the two nearest sampling points).

Variograms

To illustrate the different aspects of the spatial dependence model, consider the semi-variogram

$$\gamma(d) = \sigma_\epsilon^2 + \sigma_s^2[1 - \rho(d)],$$

where

- σ_ϵ^2 is the nugget variance,
- σ_s^2 is the spatial variance.

Variograms

The correlation function is often taken to be then Matérn, which we parameterize as (Brown, 2014),

$$\rho(d) = \frac{1}{\Gamma(\kappa)2^{\kappa-1}} \left(\frac{d\sqrt{8\kappa}}{\phi} \right)^{\kappa} K_{\kappa} \left(\frac{d\sqrt{8\kappa}}{\phi} \right),$$

with

- $\Gamma(\cdot)$ the gamma function,
- K_{κ} is a modified Bessel function of the second kind of order κ,
- ϕ is a range parameter (the distance at which the spatial correlation falls to 0.14),
- κ is a shape parameter (also called the order); this parameter determines the analytic smoothness of $S(\boldsymbol{s})$, with $S(\boldsymbol{s})$ being $\lceil \kappa - 1$ times mean-squared differentiable[2]. Hence, larger values of κ give smoother realizations.

Figure 3 displays this semi-variogram with $\sigma^2_{\epsilon} = 0.1$, $\sigma^2_{s} = 0.5$, $\kappa = 2$, $\phi = 1$.

[2] $\lceil z$ denotes the smallest integer greater than or equal to z

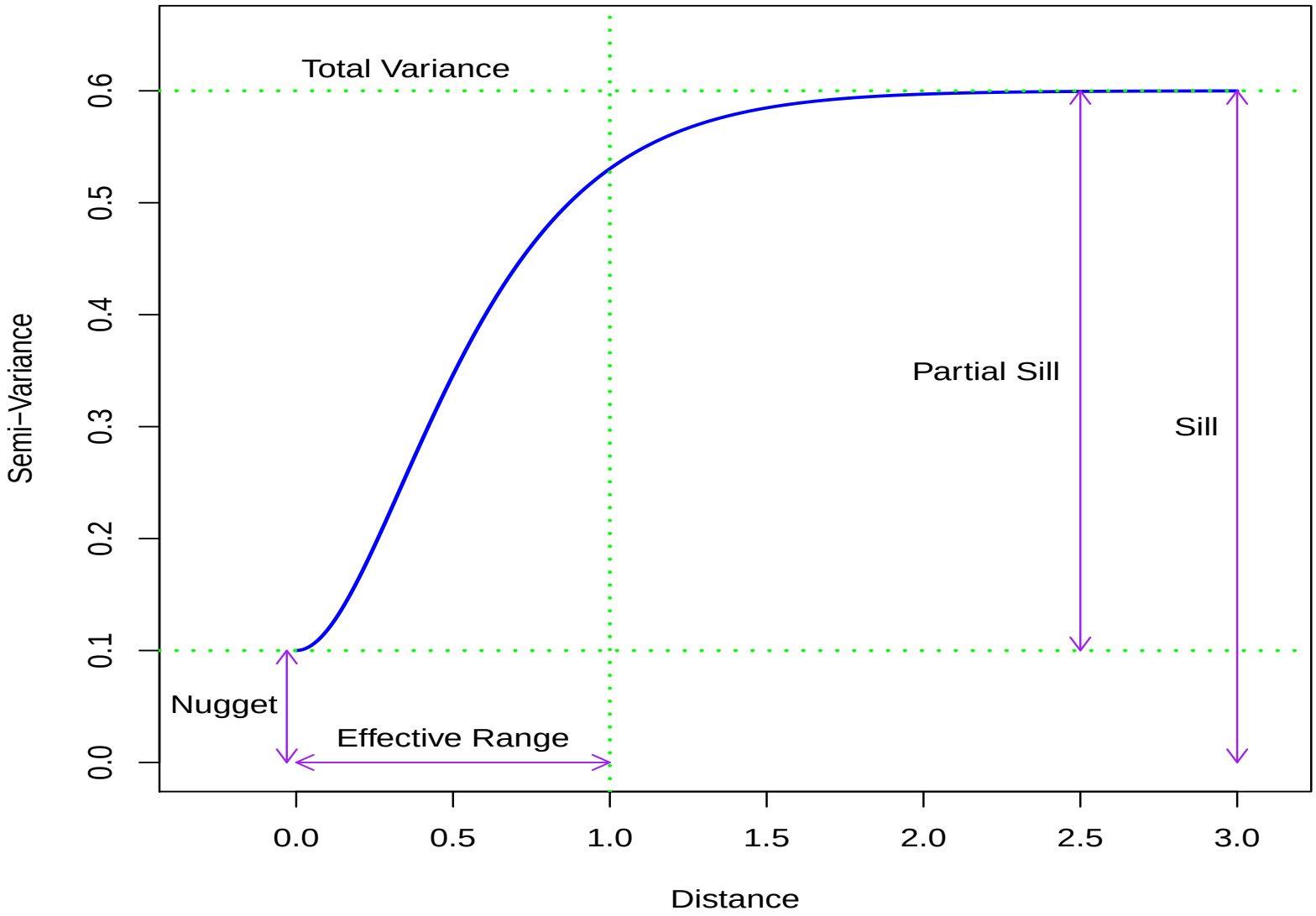

Figure 3: Theoretical Matérn semi-variogram. The sill (total variance) is $\sigma_\epsilon^2 + \sigma_s^2$, the partial sill is σ_s^2, the nugget is σ_ϵ^2 and the effective range is ϕ.

Variograms

Notes:

- Choosing $\kappa = 0.5$ gives the exponential correlation function

$$\rho(d) = \exp(-d/\phi).$$

- Taking the limit $\kappa \to \infty$ gives the Gaussian correlation function

$$\rho(d) = \exp[-(d/\phi)^2],$$

  which is not recommended since it can lead to an ill-conditioned covariance matrix, see the discussion of Diggle *et al.* (1998).
- Careful on interpretation of results from different implementations, since a number of different parameterizations are in circulation.
- Often the data are not informative enough for reliable estimation of κ, and so it is fixed.

Empirical Variograms

We gives examples of the variogram for log zinc measured close to the Meuse river.

In Figure 4 the variogram cloud (an example of an empirical variogram) is plotted and consists of, for all pairs of data points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ the contributions

$$(e_i - e_j)^2/2$$

plotted against distance

$$||\boldsymbol{s}_i - \boldsymbol{s}_j||.$$

The resultant plot in Figure 4 looks a mess, which is not uncommon. . . .

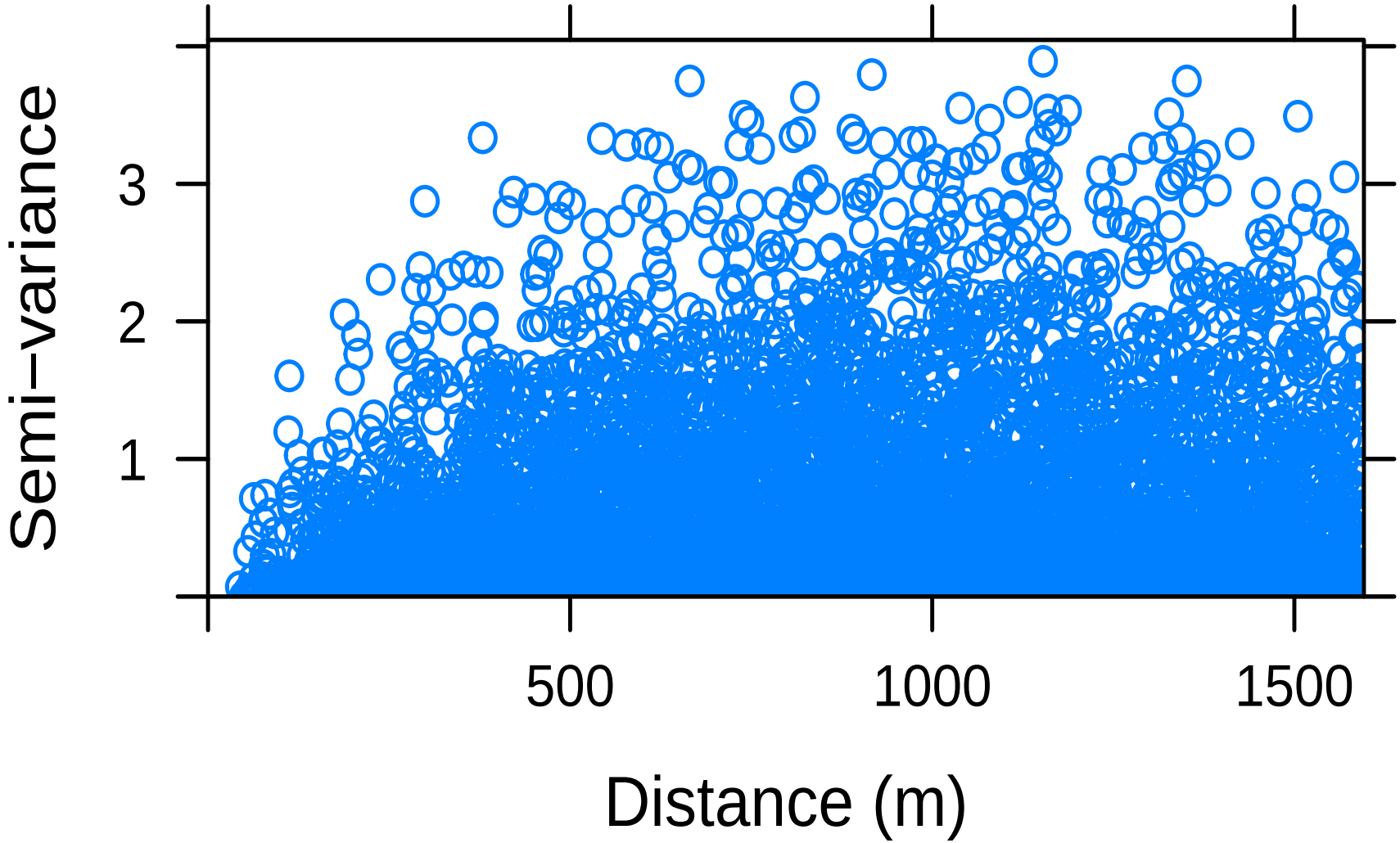

Figure 4: Variogram cloud for the log zinc data.

Empirical Variograms

- Often it is useful to look at the binned variogram in which distances are binned, with the contributions within each bin averaged.
- Specifically, if d is the midpoint of a bin:

$$\widehat{\gamma}(d) = \frac{1}{2|N_d|} \sum_{i,j \in N_d} (e_i - e_j)^2,$$

where N_d is the collection of pairs in the bin with mid-point d, and $|N_d|$ are the number of these pairs.

- Figure 5 shows the binned histogram, note the different scale in the vertical direction, as compared to Figure 4.

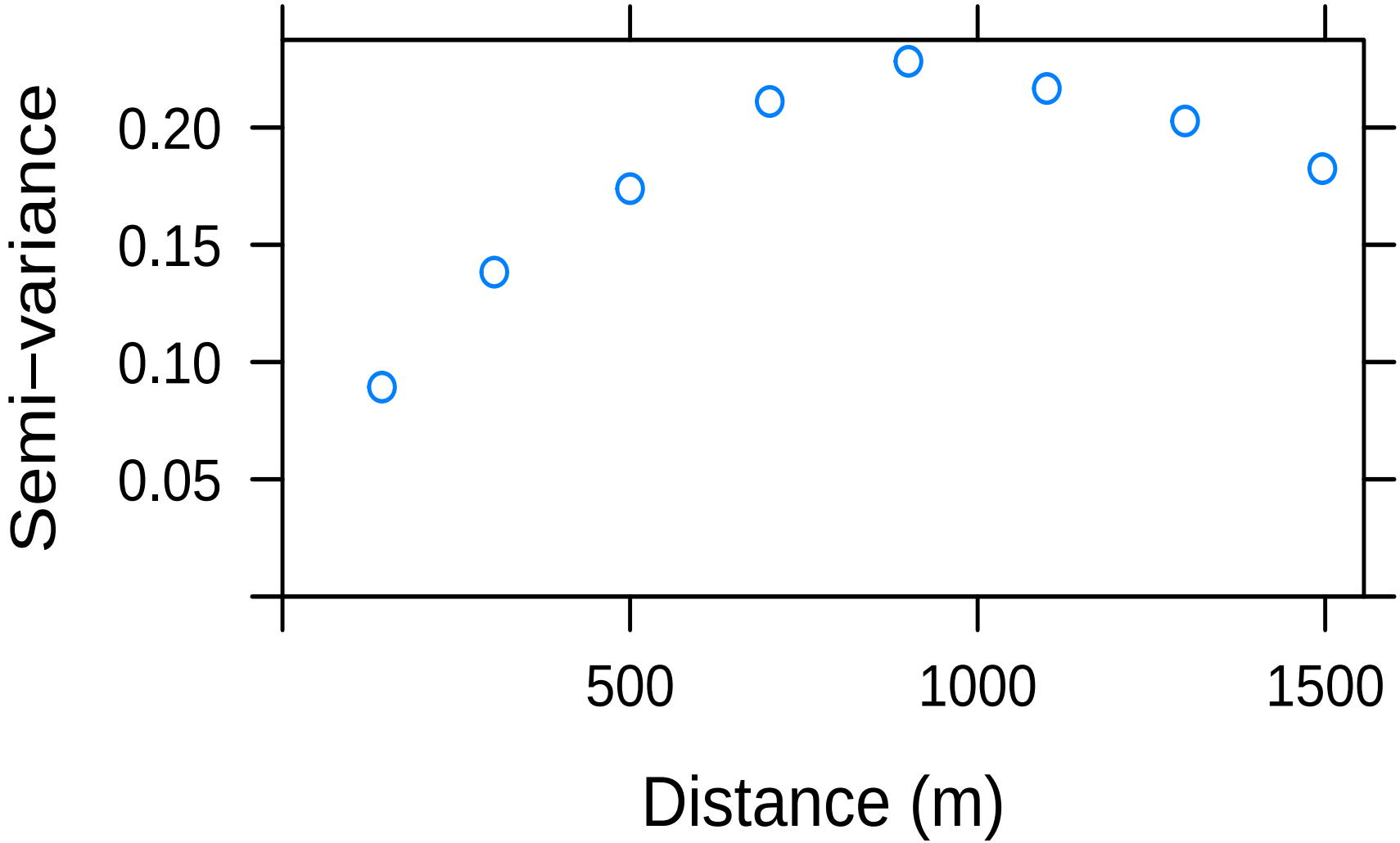

Figure 5: Binned variogram for the log zinc data.

Modeling possibilities

As usual in regression modeling, we have parameters in the regression model and in the variance-covariance model.

The variogram is an explanatory tool: how do we make more formal inference?

Various possibilities are available for fitting parametric models to spatial exposure data:

- Apply least squares (ordinary or weighted) to the empirical variogram – useful for initial estimates.
- Maximum likelihood estimation (MLE) and Restricted MLE (REML).
- Bayesian estimation.

For the second and third possibilities we must assume a probability model for the data (possibly after transformation).

For continuous data the usual choice is a normal distribution.

Gaussian Process Models

Assume we have data Y_i, $i = 1, \ldots, n$ with:

$$\boldsymbol{Y} = [Y_1, \ldots, Y_n]^\top \sim \mathrm{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

- $\boldsymbol{\mu} = \boldsymbol{\mu}(\beta) = [\mu_1, \ldots, \mu_n]$ with regression parameters β and
- $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\theta)$ is an $n \times n$ variance-covariance matrix, with parameters θ.

We have log-likelihood

$$l(\beta, \theta) = -\frac{1}{2} \log |\boldsymbol{\Sigma}(\theta)| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}(\theta)^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$$

which we need to maximize as a function of β and θ.

The determinant and inverse can be problematic to evaluate if n is large.

As an example, in the exponential correlation model we have $\boldsymbol{\theta} = (\sigma^2_\epsilon, \sigma^2_s, \phi)$.

In general, closed from estimators do not exist.

Covariance Models

The response Y_i is measured at location $\boldsymbol{s}_i$.

The model is

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + \epsilon_i + S_i$$

with $\epsilon_i \sim_{iid} \mathrm{N}(0, \sigma^2_\epsilon)$ and $\boldsymbol{S} \sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2_s\boldsymbol{R})$ where $\boldsymbol{R}$ is the $n \times n$ correlation matrix.

With the nugget effect σ^2_ϵ we have

$$\boldsymbol{\Sigma} = \sigma^2_\epsilon\boldsymbol{I} + \sigma^2_s\boldsymbol{R}$$

where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

We have already seen the common spatial exponential model with residual correlations between responses at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$:

$$R_{ij} = \exp(-d_{ij}\phi)$$

and $d_{ij} = ||\boldsymbol{s}_i - \boldsymbol{s}_j||$ is the distance between spatial locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

GeoR Example: Modeling of zinc concentration

We take as predictor variables here the distance from river and elevation variables.

Figure 6 is a variogram that was produced with a model that removes the linear effects of distance and elevation from the log zinc measurements:

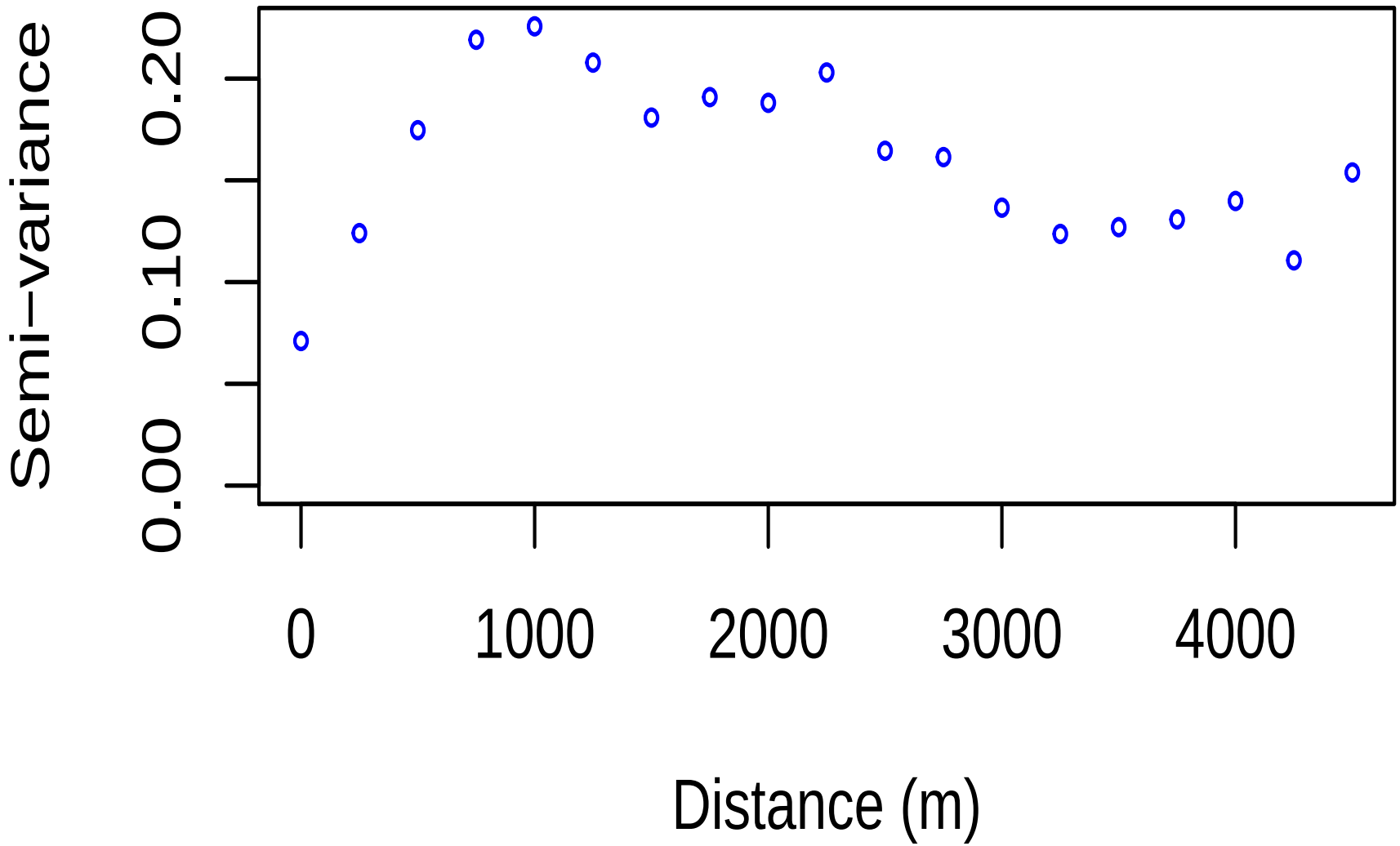

Figure 6: Binned variogram for the log zinc data, after removal of distance from river and elevation covariates effect.

Monte Carlo Test of No Spatial Correlation in the Variogram

We describe how a Monte Carlo test of

$$H_0 : \text{ No spatial correlation}$$

can be performed.

Under the null hypothesis the data can be randomly permuted (recall that we require stationarity, so trends should be removed), and we can examine the resultant variograms.

The steps are:

1. Randomly permute the data locations a set number of times.
2. Compute the semi-variogram for each permutation.
3. Calculate envelopes for each bin.
4. Plot the variogram along with the envelopes.

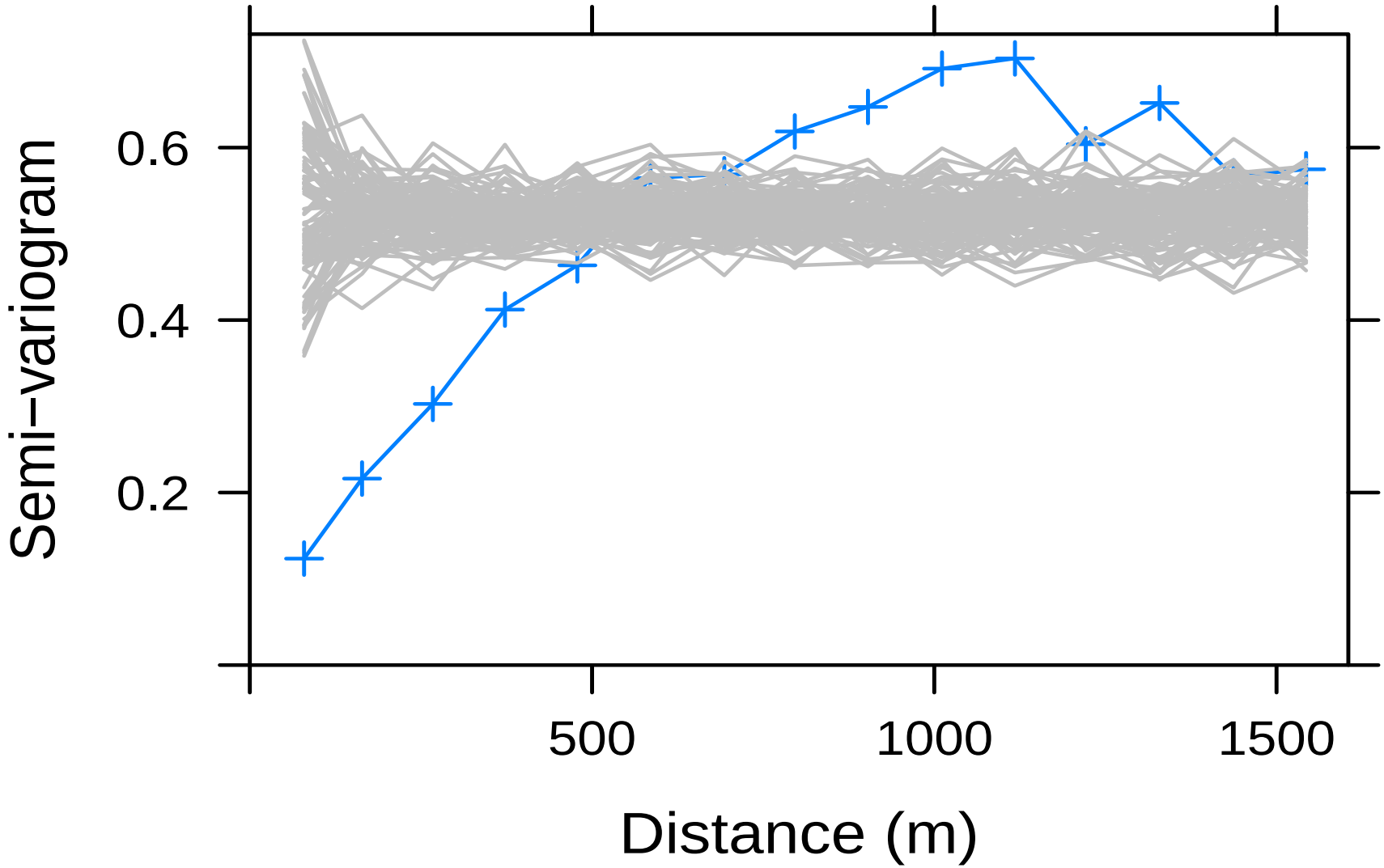

Figure 7: Binned variogram along with Monte Carlo simulations.

Fitted Variograms

We now compare the empirical binned semi-variograms with various fits.

Many models can be fitted, we illustrate with the exponential covariance model which is parameterized as $\sigma_s^2 \exp(-d/\phi)$ in `geoR`.

The nugget variance is σ_ϵ^2.

Fitted Variograms

Least squares can be used treating the data as the binned variogram values, as a function of distance; this is a nonlinear model since the semi-variogram model is not linear in σ_ϵ, σ_s^2, ϕ.

Although this approach does not rely on a distribution for the data, there are a number of problems with this approach:

- Loss of information in moving from the full dataset to the binned variogram values.
- Arbitrary binning – a different binning would produce different 'data'.
- The points are based on different numbers of observations, and are dependent, since each original datapoint contributes to multiple bins.
- WLS can address the former issue, but not the latter.

Fitted Variograms

MLE and Bayes provide more reliable estimation procedures, though they require more computation.

In random effects models REML provides, in general, more accurate estimates of variance-covariance parameters by adjusting for the degrees of freedom lost in estimation of the fixed effects (the β's).

Figure 8 compares the various fits.

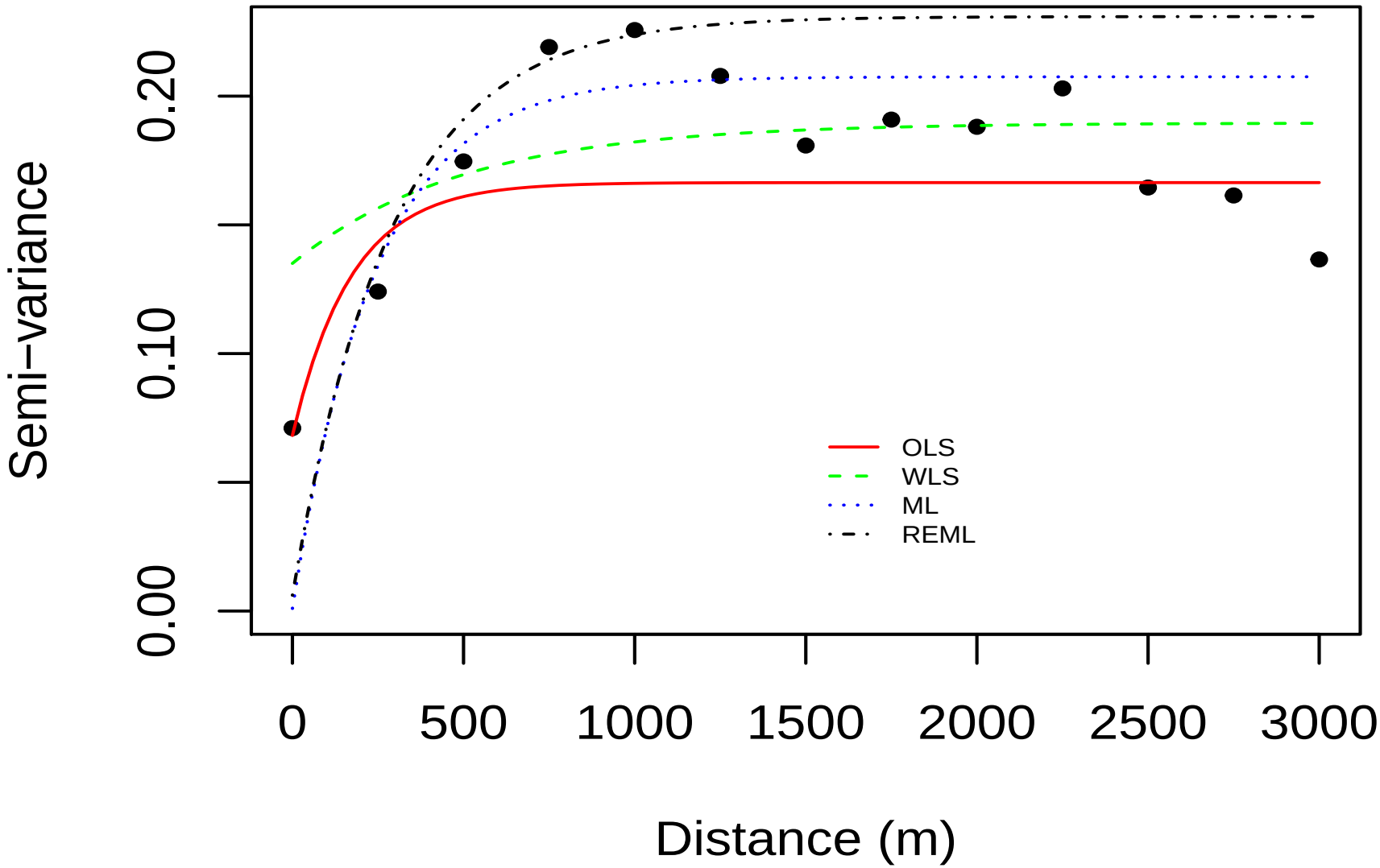

Figure 8: Binned variogram with fitted models.

Prediction of a Spatial Surface

Prediction of a Spatial Surface

We now describe how to construct a predicted spatial surface for a continuous variable. We concentrate on a geostatistical technique known as kriging, which has many flavors.

The model for a general location $\boldsymbol{s}$ is

$$Y(\boldsymbol{s}) = \beta_0 + \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}_1 + S(\boldsymbol{s}) + \epsilon$$

with ϵ are independent error terms and $S(\boldsymbol{s})$ are spatial terms.

For the observed data

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + S_i + \epsilon_i$$

for the observed data. We have a noisy version of a spatial process.

We suppose that we would like to predict the outcome at a location $\boldsymbol{s}_0$; we would like point and interval predictions with the latter not including the 'measurement error' (nugget contribution).

To do this we need to include the contribution from the underlying spatial surface, S_0

Prediction of a Spatial Surface

S_0, the spatial random effect associated with location $\boldsymbol{s}_0$.

One justification is to minimize the mean squared error (MSE) of a predictor $\widetilde{S}_0$:

$$\text{MSE} = \text{E}[(\widetilde{S}_0 - S_0)^\top \boldsymbol{A} (\widetilde{S}_0 - S_0)]$$

Then it can be shown that the form that minimizes the MSE is

$$\widetilde{S}_0 = \text{E}[S_0 | \boldsymbol{Y}]$$

To get specific forms we need to have a more detailed model.

Prediction of a Spatial Surface

Consider the model

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + S_i + \epsilon_i$$

where S_i are spatial effects with $\boldsymbol{S} = (S_1, \ldots, S_n) \sim \mathrm{N}_n(\boldsymbol{0}, \sigma_s^2\boldsymbol{R})$ and $\epsilon_i \sim_{iid} \mathrm{N}(0, \sigma_\epsilon^2)$.

Then,

$$\begin{bmatrix} S_0 \\ \boldsymbol{Y} \end{bmatrix} \sim \mathrm{N}_{n+1}\left(\begin{bmatrix} 0 \\ \beta_0 + \boldsymbol{x}\boldsymbol{\beta}_1 \end{bmatrix}, \begin{bmatrix} \sigma_s^2 & \sigma_s^2\boldsymbol{R}_0^{\mathsf{T}} \\ \sigma_s^2\boldsymbol{R}_0 & \sigma_\epsilon^2\boldsymbol{I} + \sigma_s^2\boldsymbol{R} \end{bmatrix}\right)$$

where $\boldsymbol{R}_0$ is an $n \times 1$ column vector with i-th entry describing the correlation between locations $\boldsymbol{s}_0$ and $\boldsymbol{s}_i$, $i = 1, \ldots, n$.

Prediction of a Spatial Surface

Then, using properties of the multivariate normal distribution

$$\widetilde{Y}_0 = \beta_0 + \boldsymbol{x}_0\boldsymbol{\beta}_1 + \widetilde{S}_0,$$

where

$$\begin{aligned} \widetilde{S}_0 = \mathrm{E}[S_0|\boldsymbol{Y}] &= \sigma_s^2 \boldsymbol{R}_0^{\mathsf{T}}(\sigma_\epsilon^2 \boldsymbol{I} + \sigma_s^2 \boldsymbol{R})^{-1}(\boldsymbol{Y} - \beta_0 - \boldsymbol{x}\boldsymbol{\beta}_1) \\ &= \sum_{i=1}^{n} w_i r_i \end{aligned} \tag{2}$$

and where w_i are a set of weights and $r_i = Y_i - \beta_0 - \boldsymbol{x}_i\boldsymbol{\beta}_1$ are the set of residuals.

Writing as

$$\widetilde{S}_0 - \beta_0 - \boldsymbol{x}_0\boldsymbol{\beta}_1 = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \boldsymbol{x}_i\boldsymbol{\beta}_1)$$

makes it clearer that the prediction for the residual at the new point is a weighted combination of the residuals from the n observed data points.

Prediction of a Spatial Surface

Note that if there is no spatial dependence ($\sigma_s^2 = 0$) the weights are zero and we only have contributions from $\boldsymbol{x}_0$.

The sizes of the weights depend on the proximity of the new location to the n data points, and on the magnitude of the spatial dependence, in comparison with the non-spatial.

The variance of the prediction is given by

$$\text{var}(\widetilde{S}_0|\boldsymbol{y}) = \sigma_s^2 - \sigma_s^2 \boldsymbol{R}_0^{\mathsf{T}} (\sigma_\epsilon^2 \boldsymbol{I} + \sigma_s^2 \boldsymbol{R})^{-1} \boldsymbol{R}_0 \sigma_s^2$$

again using properties of the multivariate normal distribution; this is never less that the spatial variance σ_s^2 which is good news.

The above estimator may also be justified as the best linear unbiased estimator (if the data are not Gaussian then the optimal estimator will not necessarily be linear in the data).

In practice estimates of $\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}$, where $\boldsymbol{\theta} = (\sigma_\epsilon^2, \sigma_s^2, \phi)$ and ϕ are the parameters of $\boldsymbol{R}$, are substituted into the above formulas.

Derivation of Kriging estimates

The Kriging estimator and variance may be derived in various ways, and does not require normality of the data.

One approach is to find the best linear unbiased predictor (BLUP), see Robinson (1991) for discussion.

A Small Simulated Example

$n = 6$ points were randomly simulated in the pretend study area $[0, 10] \times [0, 10]$ and then we suppose we wish to krige a prediction at the point (5,5); the left hand panel of Figure 9 shows the 6 generated sampling points in blue and the prediction point in red.

Recall that the prediction is of the form $\sum_{i=1}^{n} w_i r_i$ where r_i are the residuals from the sampling points, and

$$\boldsymbol{w}^{\mathsf{T}} = [w_1, \ldots, w_n] = \sigma_s^2 \boldsymbol{R}_0^{\mathsf{T}} (\sigma_\epsilon^2 \boldsymbol{I} + \sigma_s^2 \boldsymbol{R})^{-1},$$

for $i = 1, \ldots, n$.

Values: $\sigma_s^2 = 0.5$, $\rho = 0.9$, $\sigma_\epsilon^2 = 0.1$.

The weights associated with the 6 points are plotted against distance in the right hand panel of Figure 9; they are not monotonic in distance (point 4 has a greater weight than 6, even though at a greater distance, since it is more isolated and therefore more of an independent predictor).

Also shown is the correlation function ρ^d for $\rho = 0.9$.

A Small Simulated Example

The matrix Λ contains diagonal elements $\Lambda_{ii} = \sigma_s^2$ and off-diagonal elements $\Lambda_{ij} = \sigma_s^2 \rho^{d_{ij}}$:

$$\begin{bmatrix} 0.50 & 0.24 & 0.30 & 0.25 & 0.41 & 0.42 \\ 0.24 & 0.50 & 0.38 & 0.28 & 0.20 & 0.22 \\ 0.30 & 0.38 & 0.50 & 0.29 & 0.25 & 0.27 \\ 0.25 & 0.28 & 0.29 & 0.50 & 0.23 & 0.26 \\ 0.41 & 0.20 & 0.25 & 0.23 & 0.50 & 0.44 \\ 0.42 & 0.22 & 0.27 & 0.26 & 0.44 & 0.50 \end{bmatrix}.$$

The precision Λ^{-1} is

$$\begin{bmatrix} 8.39 & 0.09 & -1.73 & -0.01 & -3.32 & -3.14 \\ 0.09 & 5.21 & -3.53 & -1.07 & 0.03 & 0.13 \\ -1.73 & -3.53 & 6.14 & -0.62 & 0.26 & -0.29 \\ -0.01 & -1.07 & -0.62 & 3.51 & 0.14 & -1.15 \\ -3.32 & 0.03 & 0.26 & 0.14 & 10.55 & -6.79 \\ -3.14 & 0.13 & -0.29 & -1.15 & -6.79 & 11.34 \end{bmatrix}$$

Note there are no zeroes so there are no conditional independencies.

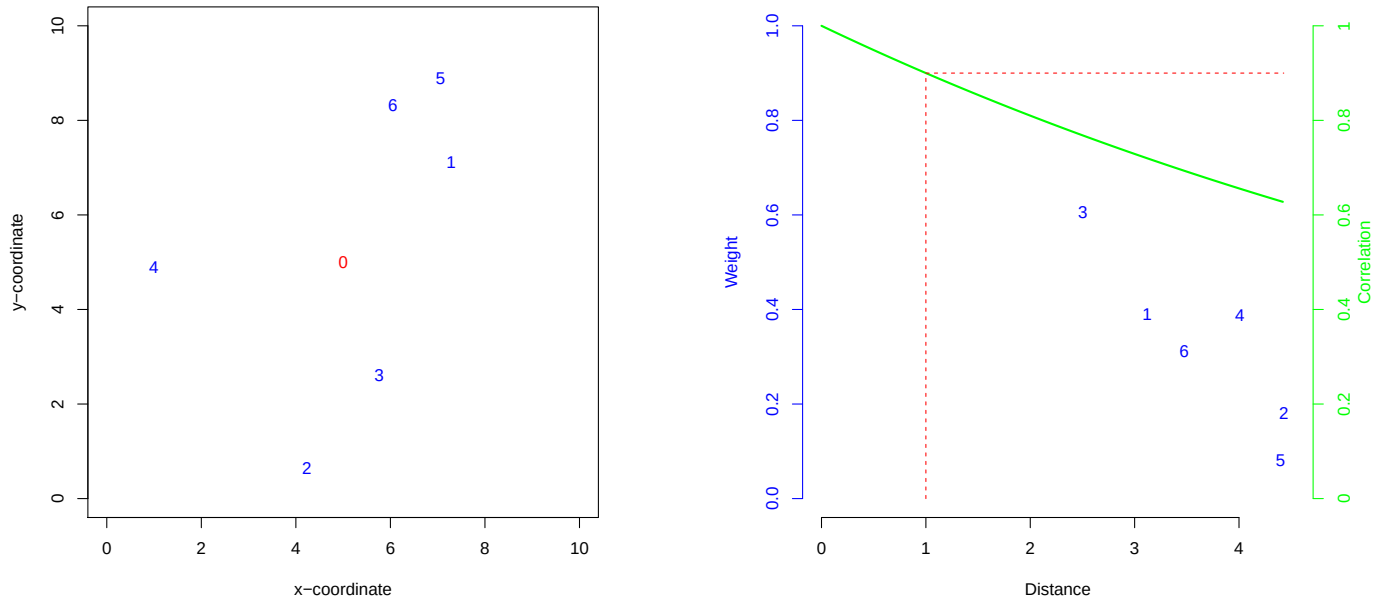

Figure 9: Left: Locations of $n = 6$ sampling points (1–6) and prediction point (0). Right: Weights and correlation function versus distance; the value 0.9 is marked since this corresponds to the correlation at 1 unit of distance.

Flavors of Kriging

Simple kriging: linear prediction assuming a known mean.

Ordinary kriging: linear prediction with a constant unknown mean.

Universal kriging: linear prediction with a non-constant mean.

Trans Gaussian kriging transforms the response to approximate normality.

For more details, see Chapter 10 of Elliott *et al.* (2000) or Chapter 8 of Waller and Gotway (2004). A more recent treatment is Zimmerman and Stein (2010).

We have already described universal kriging.

Example: Predictions of log zinc

We trend the data and then look at the residuals in Figure 10 (note the reference box which we carry out prediction over later).

Now we carry out Kriging on the residual surface.

Figures 11 and 12 show the predicted surface and the standard error surface (note the low standard errors at the sampling points).

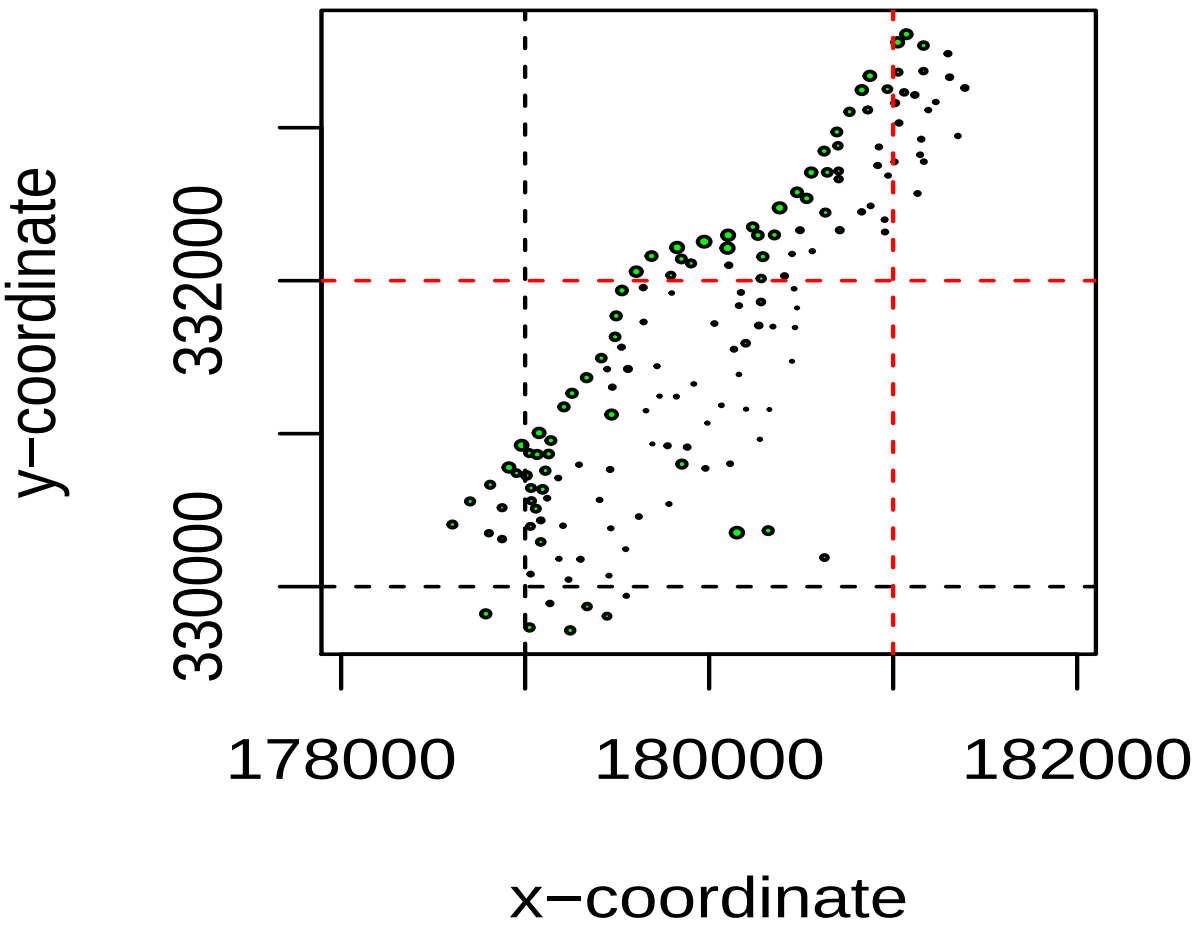

Figure 10: Values of residuals obtained after removing effect of distance from river and elevation.

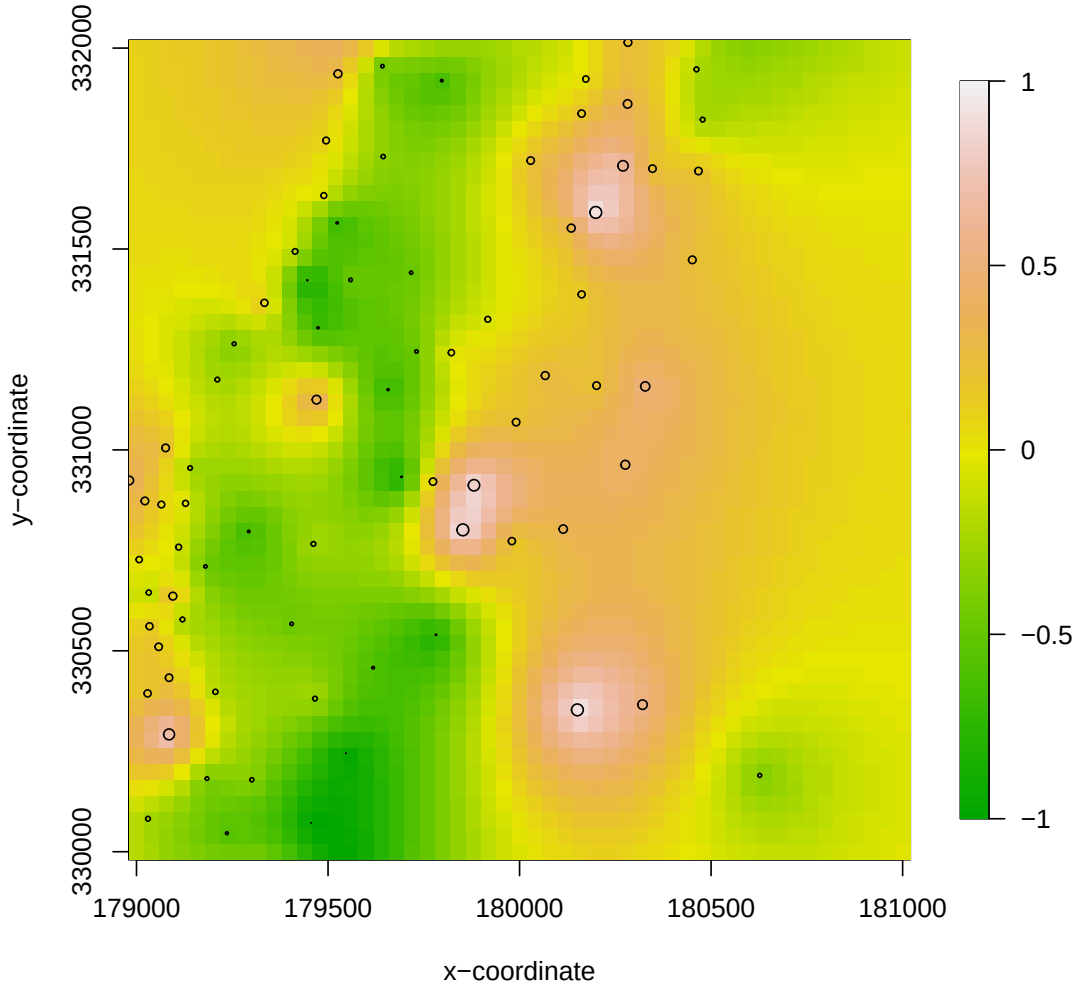

Figure 11: Predicted spatial residual surface.

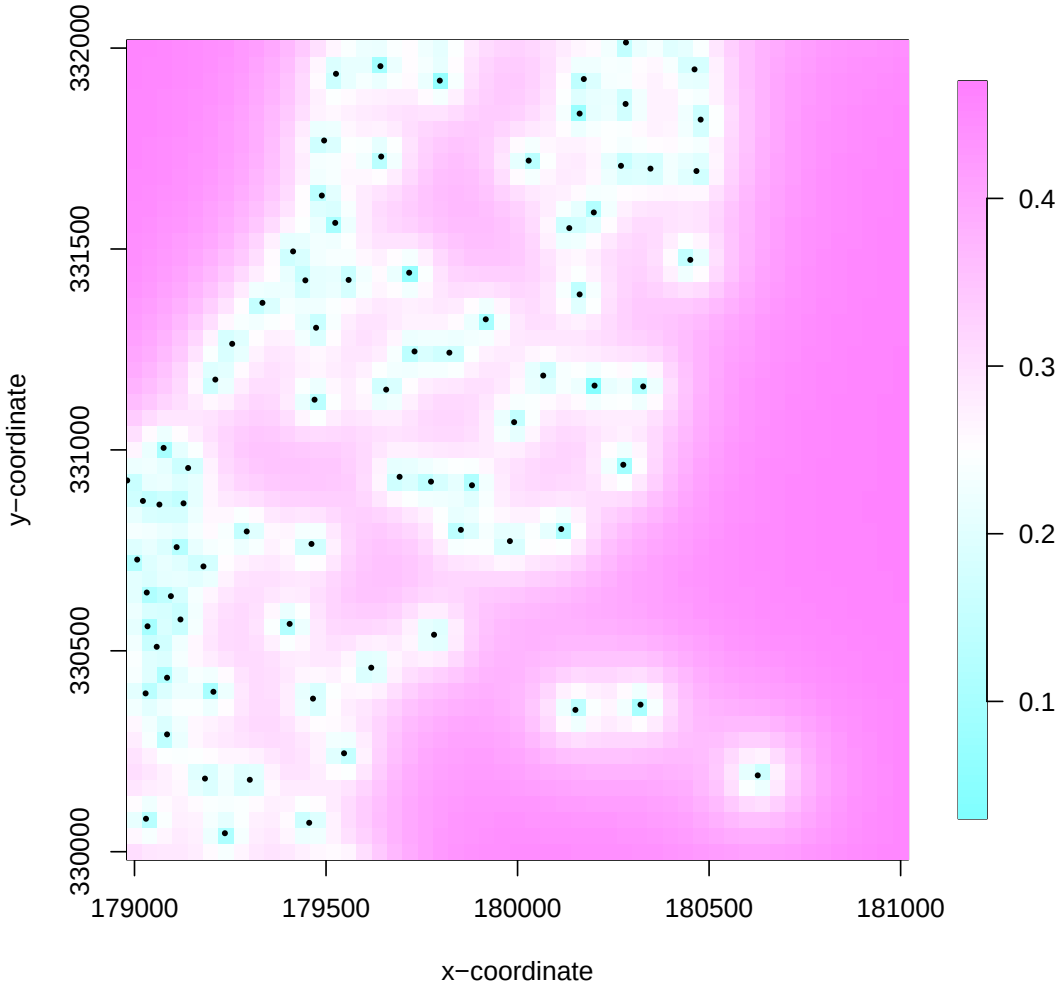

Figure 12: Standard error of prediction of residual.

Bayesian Approach

The above procedures do not account for the uncertainty in parameter estimation (like empirical Bayes) – we have used $E[S_0|\boldsymbol{y}, \widehat{\beta}_0, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\theta}}]$.

This may be acknowledged using a Bayesian approach in which we consider the posterior distribution,

$$p(S_0|\boldsymbol{y}) = \int_{\beta_0} \int_{\boldsymbol{\beta}_1} \int_{\boldsymbol{\theta}} p(S_0|\boldsymbol{y}, \beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}) \times p(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}|\boldsymbol{y})\, d\beta_0 d\boldsymbol{\beta}_1 d\boldsymbol{\theta}$$

where we average over the posterior distribution for $\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}$, and

$$p(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}|\boldsymbol{y}) \propto L(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta}) p(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta})$$

so that a prior $p(\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\theta})$ is required.

Overview of Approaches

See Bradley *et al.* (2016) and Heaton *et al.* (2017) for reviews of approaches, with an emphasis on big data.

Fixed rank kriging (Cressie and Johannesson, 2008); `R` implementation via the `FRK` package:

`https://cran.r-project.org/web/packages/FRK/index.html`

Predictive processes (Banerjee *et al.*, 2008). `R` implementation available in the `spBayes` package.

Lattice kriging (Nychka *et al.*, 2015); `R` implementation via the `LatticeKrig` package:

`https://cran.r-project.org/web/packages/LatticeKrig/index.html`

Approximation based on a stochastic partial differential equation (SPDE) and utilizing INLA (Lindgren *et al.*, 2011; Simpson *et al.*, 2012).

Bayesian Approach: Computation

As we have seen, Bayesian models may be fitted in `inla` (using analytical approximations and numerical integration).

- The INLA framework fits Markov random field (MRF) models, and the Gaussian process model does not fit into this framework, but **?**) show that the GP model may be represented as an MRF, thus allowing the GP model to be fitted within `inla`.
- The package `geostatsp` package makes the fitting of these models simpler (Brown, 2014).

Bayesian models may also be fitted via MCMC:

- in `WinBUGS`, with `spatial.exp` providing the exponential correlation model.
- The `krige.bayes()` function within `geoR` carries out Bayesian inference via Markov chain Monte Carlo (MCMC).
- `Stan` is improving its capabilities for spatial models, and there is an efficient implementation for the BYM model (Morris *et al.*, 2019).
- For continuous spatial models, efficient MCMC implementations are more challenging.

Generalized Additive Models

Generalized Additive Models (GAMs)

A different approach to modeling the spatial structure is via a deterministic model – this may work well if there is not too much small-scale spatial variation

A GAM extends the usual generalized linear model (GLM) by adding a non-parametric component.

For continuous (nominally normal) data the model is of the form

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + S(\boldsymbol{s}_i) + \epsilon_i$$

where $\epsilon_i \sim \text{N}(0, \sigma^2_\epsilon)$ and $S(\boldsymbol{s}_i)$ is a smooth function of the spatial location $\boldsymbol{s}_i$.

This smooth function can be specified in a number of ways.

A popular choice is a spline model, though kernels may also be used.

An excellent text on GAMs is Wood (2006).

Generalized Additive Models (GAMs)

The following leans on Chapters 11 and 12 of Wakefield (2013).

Suppose wish to minimize the penalized sum of squares,

$$\sum_{i=1}^{n} [y_i - f(s_{i1}, s_{i2})]^2 + \lambda P(f) \tag{3}$$

where $f(\boldsymbol{s})$ is the unknown latent spatial function that we would like to estimate.

The penalization term is,

$$P(f) = \int \int \left[\left(\frac{\partial^2 f}{\partial s_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial s_1 \partial s_2} \right)^2 + \left(\frac{\partial^2 f}{\partial s_2^2} \right)^2 \right] ds_1 ds_2. \tag{4}$$

This term penalizes wiggliness.

Generalized Additive Models (GAMs)

As shown by Green and Silverman (1994, Chapter 7) the unique minimizer is provided by the natural thin plate spline with knots at the observed data (to give a so-called smoothing spline), which is defined as

$$f(\boldsymbol{s}) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \sum_{i=1}^{n} b_i \eta(||\boldsymbol{s} - \boldsymbol{s}_i||), \tag{5}$$

where

$$\eta(r) = \begin{cases} \frac{1}{8\pi} r^2 \log(r) & \text{for } r > 0 \\ 0 & \text{for } r = 0 \end{cases}$$

and the unknown b_i are constrained via

$$\sum_{i=1}^{n} b_i = \sum_{i=1}^{n} b_i s_{i1} = \sum_{i=1}^{n} b_i s_{i2} = 0.$$

Generalized Additive Models (GAMs)

Such a spline provides the unique minimizer of $P(f)$ amongst interpolating functions. Interested readers are referred to Theorems 7.2 and 7.3 of Green and Silverman (1994) and to Duchon (1977), who proved optimality and uniqueness properties for natural thin plate splines.

Natural thin plate splines are very appealing since they remove the need to decide upon knot locations or basis functions; each is contained in (5).

In practice, however, thin plate splines have too many parameters. A thin plate regression spline (TPRS) truncates the space of the "wiggly" basis (the b_i's in (5)), while leaving β unchanged.

Various approaches are available for selecting the key smoothing parameter λ in (3) including cross-validation and REML (under a mixed model representation).

Spline Detail

Wahba (1978) showed that this smoothing spline solution could be reached using a Bayesian approach.

For simplicity, we describe in the one-dimensional case, with the variable over which smoothing is being carried out is t on the range $[0, 1]$.

A frequentist approach chooses $f(\cdot)$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}[\, y_i - f(t_i)\,]^2 + \lambda \int_0^1 \frac{d^2 f}{du} du, \tag{6}$$

with $\lambda > 0$.

The polynomial pieces connect and the resulting function has two continuous derivatives.

Spline Detail

The prior on $f(t)$ is of the form

$$f(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + Z(t),$$

where $Z(t)$ is a zero-mean Gaussian process with

$$\frac{d^2 Z(t)}{dt^2} = \delta^{1/2} \frac{dW(t)}{dt}, \tag{7}$$

where $\delta > 0$ is a scale parameter and $dW(t)/dt$ is scaled white noise, i.e., $W(t)$ is a zero-mean Wiener process with variance t.

Spline Detail

Wahba (1978) characterizes $Z(t)$ as the integrated Wiener process:

$$Z(t) = \int_0^t (t-u)^{-1} dW(u).$$

The prior on $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2]$ is an improper flat prior, which is taken as the limit of a two-dimensional normal, $\mathrm{N}(\mathbf{0}, \xi \boldsymbol{I})$ as the precision $\xi \to \infty$.

In this situation, the posterior mean of f corresponds to the usual frequentist estimator with $\lambda = \sigma^2/\delta$.

Example of GAM fitting: Zinc Data

We fit a GAM to the log zinc data using thin plate regression splines.

We display the fitted surface $S(\boldsymbol{s})$ over a grid of $\boldsymbol{s}$ values, in Figure 14.

A comparison between the two predictions is given in Figure 16 and we see that the dame medium- to large-scale features are being picked up.

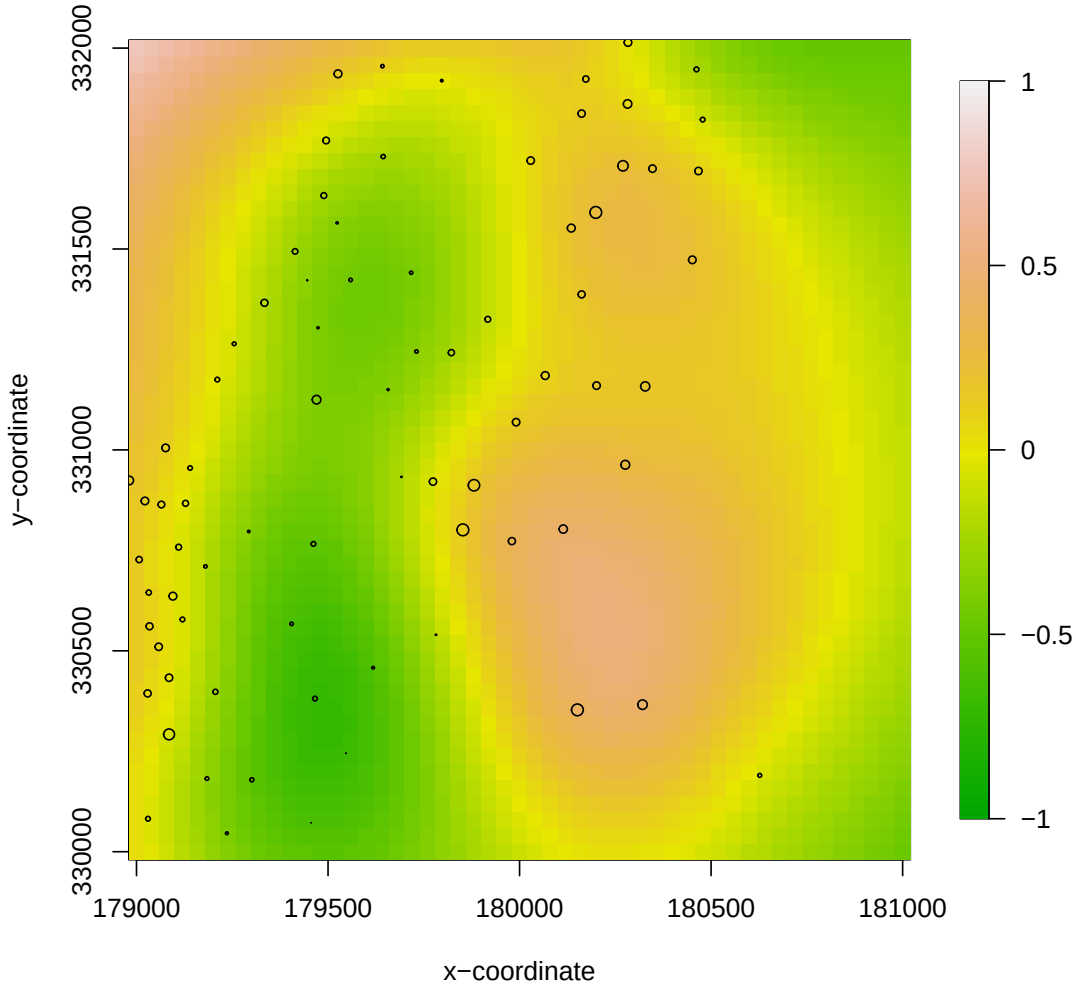

Figure 13: Predicted (residual) surface from the GAM model.

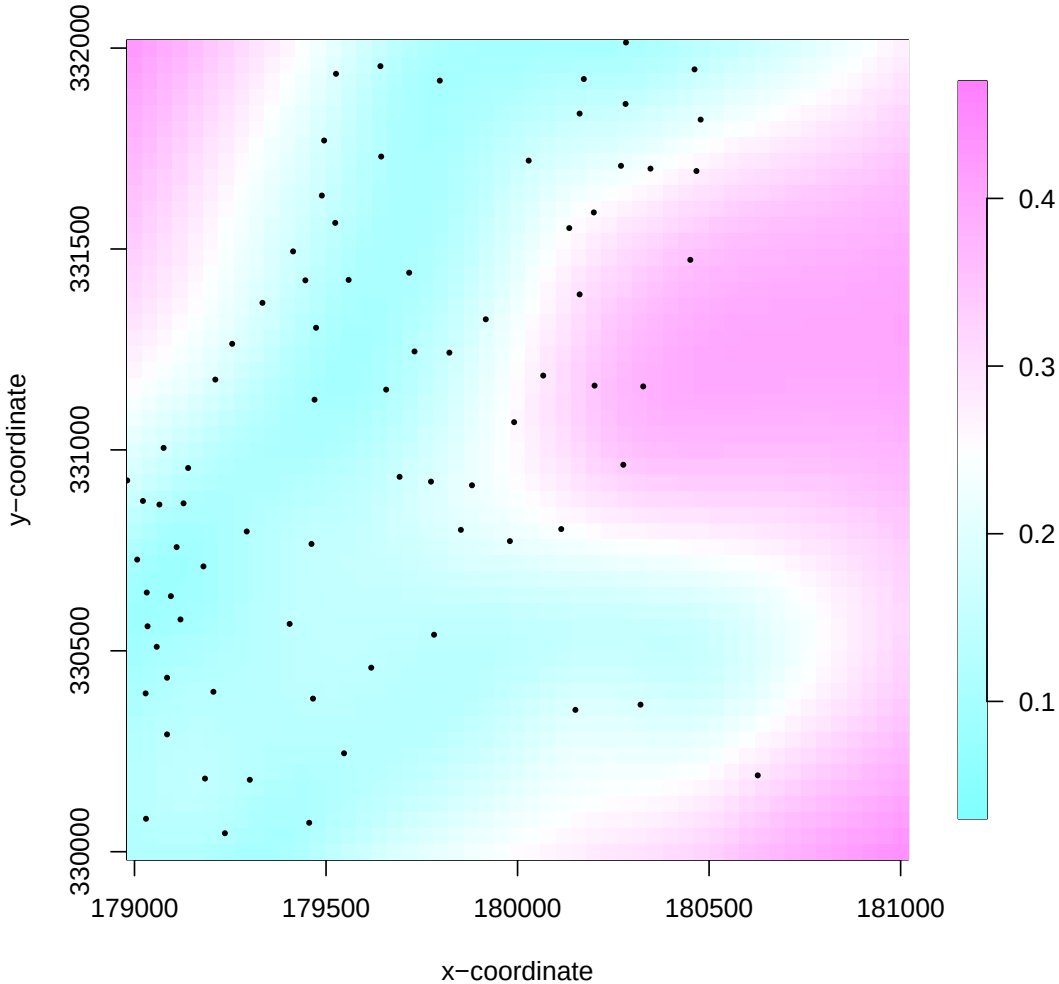

Figure 14: Uncertainty from the GAM model.

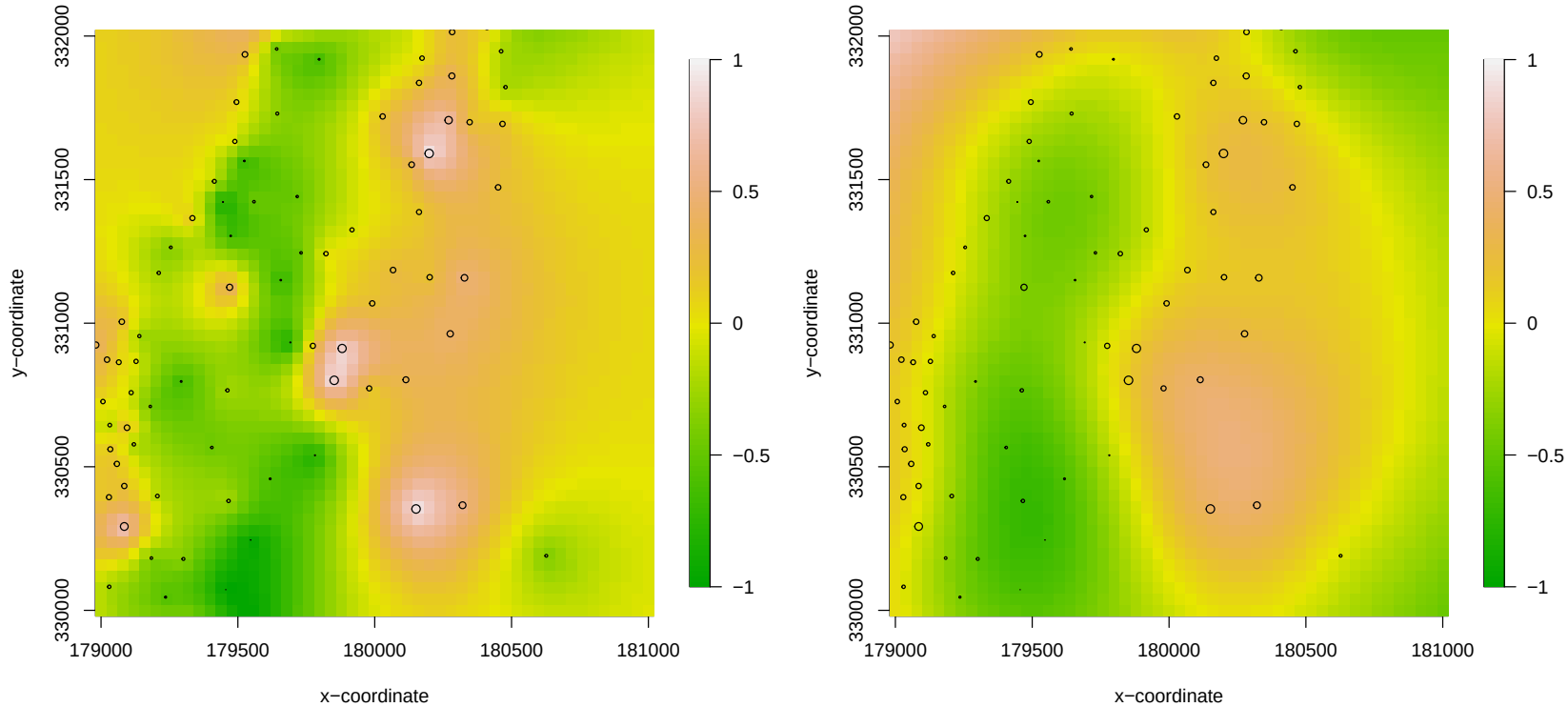

Figure 15: Predicted (residuals) surfaces from the geostatistical and GAM models.

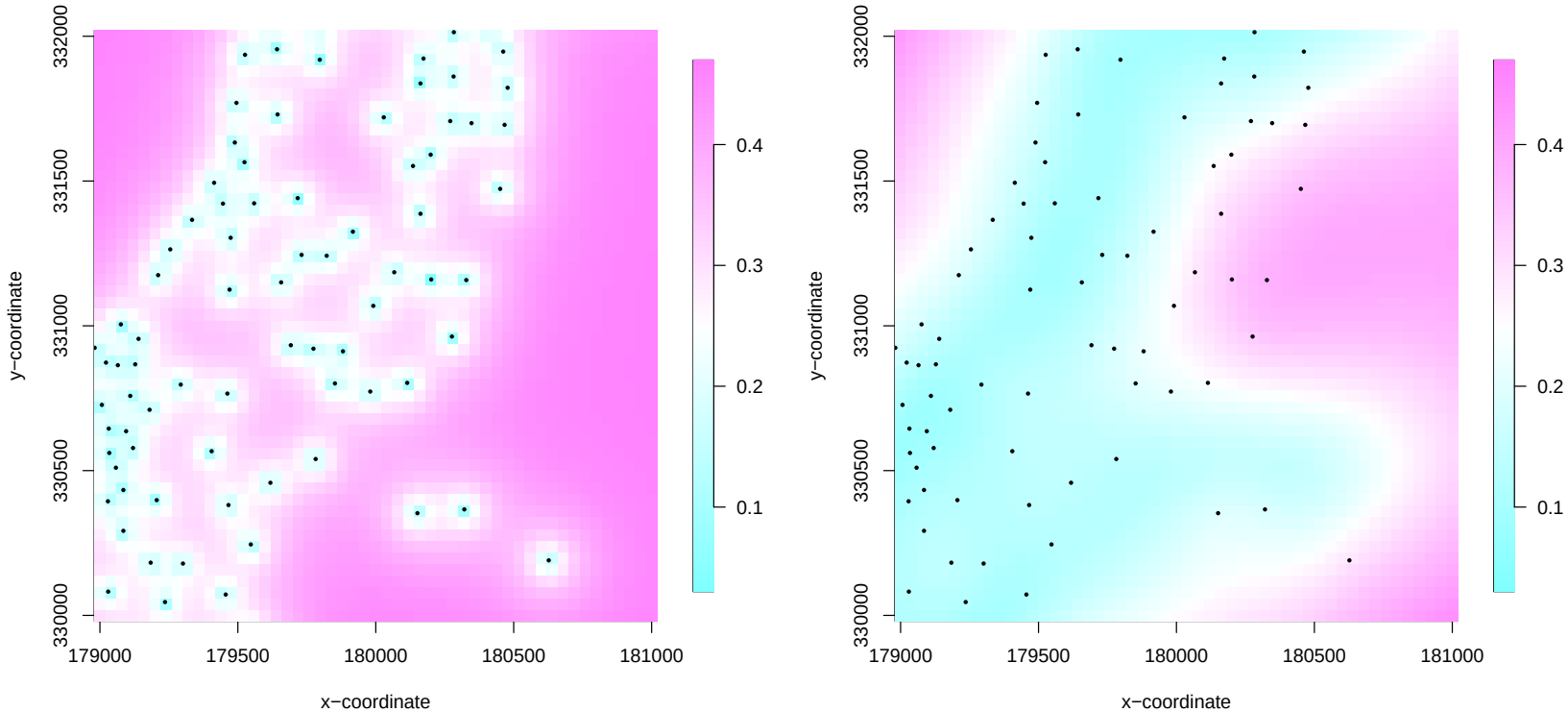

Figure 16: Uncertainty surfaces from the geostatistical and GAM models.

Summary for Continuous Responses

The semi-variogram can be a useful exploratory tool, but they can be very noisy.

Other explanatory tools for examining spatial dependence/clustering include:

- Moran's I for count data.
- F, G, K, L functions for point process data.

Summary for Continuous Responses

We have described two models for modeling continuous point data.

In the geostatistical (kriging) model we have

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + S_i + \epsilon_i$$

where $S_i = S(\boldsymbol{s}_i)$ is the spatial random effect at location $\boldsymbol{s}_i$.

This model can be good at picking up small scale behavior but assumes the same form of dependence across the study region.

Tricky aspects include:

- knowing when we have the data to obtain reliable estimates – this includes both the number of data points, and the configuration (what's the distribution of distances between points?).
- Checking the model:
  - Stationarity over study region?
  - Deterministic component?
  - Covariance Function?
  - Distribution of errors?

Summary for Continuous Responses

In the GAM we have

$$Y_i = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + g(\boldsymbol{s}_i) + \epsilon_i$$

with $g(\boldsymbol{s}_i)$ a smooth term that may be estimated using splines (for example).

This model is sometimes known as a geoadditive model (Kammann and Wand, 2003).

The GAM is good at picking up medium- and large-scale spatial trends but not so good at small scale.

Being nonparametric it does allow flexibility across the map.

I view as more of an exploratory tool.

Overview of Approaches

At a generic location $\boldsymbol{s}$ we assume,

$$Y(\boldsymbol{s}) = \eta(\boldsymbol{s}) + \epsilon, \tag{8}$$

where ϵ represents zero-mean measurement error (which may have known variance) and $\eta(\boldsymbol{s})$ consists of the regression contribution and the latent field.

The latent field may be decomposed as

$$\eta(\boldsymbol{s}) = \beta_0 + \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}_1 + S(\boldsymbol{s}) + e, \tag{9}$$

with $S(\boldsymbol{s})$ and e representing small-scale and micro-scale spatial variation, respectively.

The component $\beta_0 + \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}_1$ (which may be replaced by a non-linear function) is sometimes known as the drift or spatial trend.

Overview of Kriging Type Approaches

In the Earth sciences, (8) is known as the data model and (9) the process model.

In the following, all terms ϵ, e and $S(\boldsymbol{s})$ (if treated as stochastic) are assumed independent.

Data $\{y(\boldsymbol{s}_i), \boldsymbol{x}(\boldsymbol{s}_i), i = 1, \ldots, n\}$, are observed at locations $\boldsymbol{s}_i$ and the aim is often prediction at a set of unobserved locations.

Overview of Approaches

There are many different options for $S(\boldsymbol{s})$ — we have already discussed the kriging approach to prediction in which a covariance model is assumed (for example, the Matérn) and the prediction (2) is used.

Unfortunately, the computational complexity is $O(n^3)$ so prohibitive for large datasets.

A crucial observation is that the kriging estimator is a sum of basis functions with coefficients that are linear in the observed $\boldsymbol{y}$ and where the bases functions are defined in terms of the covariance function (Nychka, 2000).

Overview of Approaches

A non-stochastic (deterministic) approach for modeling $S(\boldsymbol{s})$ is provided by splines, an example of which we have already seen.

If a smoothing spline is used, then the computational complexity is again $O(n^3)$.

In a thin-plate regression spline with K knots, the complexity reduces to $O(nK^2 + K^3)$.

Mapping of Binary Data

Overview of Point Data Mapping

We now consider binary data, and the visualization of relative risk/odds ratio surfaces.

In an epidemiology context such data may arise from a case-control study (since we will rarely know the locations of all non-cases).

Alternatively, again in epidemiology (or in the social sciences) cross-sectional surveys are carried out.

Also in epidemiology, particularly in a developing world context, prevalence mapping is an important endeavor, with the data available from surveys carried out in particular villages (say).

Binary point data are difficult to visualize, since they are binary!

This observation suggests we need to bin (to create denominators greater than 1) or smooth geographically.

Overview of Disease Mapping for Binary Data

We will describe three approaches:

1. Kernel Density Estimation (KDE) Approach:
   - We can view the non-case locations, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_0}$ as a sample from a probability density $f_0(\boldsymbol{s})$.
   - We can view the case locations, $\boldsymbol{s}_{n_0+1}, \ldots, \boldsymbol{s}_{n_0+n_1}$ as a sample from a probability density $f_1(\boldsymbol{s})$.
   - These densities can these be estimated using KDE.
   - This approach can be used in exploratory analyses, but alone is less good for inference.
2. Generalized Additive Models (GAMs)
   - Supplement a generalized linear model (GLM) with a smoother component in space, such as a kernel or a spline, known as a GAM.
   - In this way one can carry out inference and adjust for confounders such as age and gender, and other predictors such as exposures.

Overview of Disease Mapping for Binary Data

3 Geostatistical Approach:

- Let $p(\boldsymbol{s})$ be the probability of being a case at location $\boldsymbol{s}$.
- We can then consider models of the form

$$\log\left(\frac{p(\boldsymbol{s})}{1-p(\boldsymbol{s})}\right) = \beta_0 + \beta_1 x(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) + S(\boldsymbol{s}),$$

  where $x(\boldsymbol{s})$ is a spatially references covariate and $\epsilon(\boldsymbol{s})$ and $S(\boldsymbol{s})$ are non-spatial and spatial residual error terms.
- The likelihood is

$$Y_i | p(\boldsymbol{s}_i) \sim \text{Bernoulli}[p(\boldsymbol{s}_i)],$$

$i = 1, \ldots, n_0, n_0 + 1, \ldots, n_0 + n_1$.

Analogy with regression: If we obtain a simple random sample of (x, y) pairs, we can reconstruct the x distribution, but if the x's are fixed by design we cannot learn about the x distribution, but we can model $y|x$.

Binary Data: GAMs

If we define

$$p_i = p(\boldsymbol{s}_i) = \Pr(Y_i = 1 | \boldsymbol{x}_i, \boldsymbol{s}_i)$$

as the probability of a case at location $\boldsymbol{s}_i$ with covariates $\boldsymbol{x}_i$ then

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + g(\boldsymbol{s}_i).$$

Kelsall and Diggle (1998) suggest using a generalized additive model (GAM) to simultaneously estimate the effect of covariates and use KDE on the residuals, i.e. to estimate $g(\boldsymbol{s})$.

In general, different smoothers may be used; for example, penalized regression splines are a common choice.

See Bivand et al. (2013, Section 7.5.2).

Binary Data: GAMs

Recall

$$p_i = p(\boldsymbol{s}_i) = \Pr(Y_i = 1 | \boldsymbol{x}_i, \boldsymbol{s}_i)$$

as the probability of a case at location $\boldsymbol{s}_i$ with covariates $\boldsymbol{x}_i$ then

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_1 + \boldsymbol{x}_i\boldsymbol{\beta}_1 + g(\boldsymbol{s}_i).$$

In general, there are many possibilities for modeling $g(\boldsymbol{s})$ including kernels, splines and local polynomials.

We can replace $g(\boldsymbol{s}_i)$ with $\epsilon(\boldsymbol{s}_i) + S(\boldsymbol{s}_i)$ where $\epsilon(\boldsymbol{s}_i)$ and $S(\boldsymbol{s}_i)$ are random effects without and with spatial structure – known as the geostatistical approach.

Geostatistical Approach

- Let $p(\boldsymbol{s})$ be the probability of being a case at location $\boldsymbol{s}$.
- We can then consider models of the form

$$\log\left(\frac{p(\boldsymbol{s})}{1-p(\boldsymbol{s})}\right) = \beta_0 + \beta_1 x(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) + S(\boldsymbol{s}),$$

  where $x(\boldsymbol{s})$ is a spatially references covariate and $\epsilon(\boldsymbol{s})$ and $S(\boldsymbol{s})$ are non-spatial and spatial residual error terms.
- The likelihood is

$$Y_i | p(\boldsymbol{s}_i) \sim \text{Bernoulli}[p(\boldsymbol{s}_i)],$$

  $i = 1, \ldots, n_0, n_0 + 1, \ldots, n_0 + n_1$.
- Inference using INLA via wrappers is available (Brown, 2014) or using a full-on SPDE approach (Lindgren *et al.*, 2011).

Summary for Binary Data

The KDE approach with points can be useful for exploratory purposes, but I wouldn't trust it for inference, because it's so sensitive to the smoothing parameter.

The GAM approach is better in this respect, but picking up small-scale variation using splines (for example) is not straightforward.

The model-based geostatistics approach is flexible and can provide reliable inference, but many crucial choices in the implementation, not least of which is the prior on the variances (spatial and nugget).

References

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, **70**, 825–848.

Bartlett, M. (1964). The spectral analysis of two-dimensional point processes. *Biometrika*, **51**, 299–311.

Bivand, R., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R, 2nd Edition*. Springer, New York.

Bradley, J. R., Cressie, N., Shi, T., *et al.* (2016). A comparison of spatial predictors when datasets could be very large. *Statistics Surveys*, **10**, 100–131.

Brown, P. (2014). Model-based geostatistics the easy way. *Journal of Statistical Software*. In Press.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B*, **70**, 209–226.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–350.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Solobev spaces. In W. Schemp and K. Zeller, editors,

Construction Theory of Functions of Several Variables, pages 85–100. Springer, New York.

Elliott, P., Wakefield, J. C., Best, N. G., and Briggs, D. J. (2000). *Spatial Epidemiology: Methods and Applications*. Oxford University Press, Oxford.

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall/CRC Press, Boca Raton.

Heaton, M. J., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., *et al.* (2017). Methods for analyzing large spatial data: A review and comparison. *arXiv preprint arXiv:1710.05013*.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics*, **52**, 1–18.

Kelsall, J. and Diggle, P. (1998). Spatial variation in risk of disease: a non-parametric binary regression approach. *Applied Statistics*, **47**, 00–00.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, **73**, 423–498.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. *Spatial and Spatio-Temporal Epidemiology*, **31**, 100301.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, **24**, 579–599.

Nychka, D. W. (2000). Spatial-process estimates as smoothers. *Smoothing and regression: approaches, computation, and application*, pages 393–424.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**, 15–32.

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, **1**, 16–29.

Tarr, G. A., Shringi, S., Phipps, A. I., Besser, T. E., Mayer, J., Oltean, H. N., Wakefield, J., Tarr, P. I., and Rabinowitz, P. (2018). Geogenomic segregation and temporal trends of human pathogenic escherichia coli o157: H7, Washington, USA, 2005–2014. *Emerging infectious diseases*, **24**, 32.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40**, 364–372.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer, New York.

Waller, L. and Gotway, C. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley and Sons.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC Press, Boca Raton.

Zimmerman, D. and Stein, M. (2010). Classical geostatistical methods. In A. Gelfand, P. Diggle, M. Fuentes, and P. Guttorp, editors, *Handbook of spatial statistics*, pages 29–44. CRC Press.