# TMA4250 Spatial Statistics
# Assignment 1: Continuous Spatial Variables

IMF/NTNU/HO

January 2020

## Introduction

This assignment contains problems related to continuous spatial variables and Gaussian random fields (RF). We recommend using R for solving the problems, and relevant functions can be found in the R libraries geoR, akima and fields.

## Problem 1: Gaussian RF - model characteristics

Consider the continuous spatial variable $\{r(x) : x \in \mathtt{D} : [1, 50] \subset \mathbb{R}^1\}$, and assume that it is modeled as a stationary 1D Gaussian RF with the following model parameters:

$$\begin{aligned} E\{r(x)\} &= \mu_r = 0 \\ Var\{r(x)\} &= \sigma_r^2 \\ Corr\{r(x), r(x')\} &= \rho_r(\tau) \end{aligned}$$

where $\rho_r(\tau); \tau = |x - x'|/10$ is the spatial correlation function. Let $\mathtt{D} : [1, 50]$ be discretized in $\mathtt{L} \in \{1, 2, \ldots, 50\}$ and define the discretized Gaussian RF $\{r(x); x \in \mathtt{L}\}$.

Let the spatial correlation function $\rho_r(\tau)$, be either Powered exponential with parameter $\nu_r \in \{1, 1.9\}$ or Matern with parameter $\nu_r \in \{1, 3\}$. Let the variance take the values $\sigma_r^2 \in \{1, 5\}$.

**a)** The spatial correlation function must be a positive definite function, specify this requirement mathematically. Explain why this requirement is necessary ?

Display the two spatial correlation functions for $\tau \in \mathbb{R}_{\oplus}$ for the model parameters specified above. Discuss the features of the spatial correlation function which are crucial for the associated Gaussian RF, and the relations between

the variance and correlation function with the variogram function. Display the associated variogram functions $\gamma_r(\tau)$.

Use the functions cov.spatial(.) and/or Matern(.).

**b)** Specify the pdf for the corresponding Gaussian model and let it be the prior model.

Simulate four realizations of the Gaussian RF on L for each of the eight different sets of model parameters defined above and present them in eight displays with four realizations in each.

Discuss the relation between the realizations and the model parameters.

Let the spatial variable be observed as $\{d(x); x \in \{10, 25, 30\} \subset \mathtt{L}\}$ according to the acquisition model,

$$d(x) = r(x) + \epsilon(x) \qquad\qquad x \in \{10, 25, 30\}$$

with measurement errors $\epsilon(\cdot)$ being centered, independent and identically Gaussian distributed with variance $\sigma_\epsilon^2$. Further, assume that $r(x)$ and $\epsilon(x')$ are independent for all $x, x'$.

**c)** Specify the expression for the corresponding likelihood model and explain why this is not a pdf, and the consequences thereof.

Consider the simulated realizations in **b)** with $\sigma_r^2 = 5$ and select one realization, and use the values at $x \in \{10, 25, 30\}$ in this realization as the observed values $\{d(x); x \in \{10, 25, 30\} \subset \mathtt{L}\}$. Let the observation error variance take the values $\sigma_\epsilon^2 \in \{0, 0.25\}$.

**d)** Specify the pdf for the discretized posterior Gaussian RF given the observations.

Use the prior model, the likelihood model with the two error variances listed above, and the actual observed values. Compute the corresponding two predictions for the spatial variable $\{\hat{r}(x); x \in \mathtt{L}\}$ with associated prediction 0.9-intervals, and present the results in two displays.

Discuss these two displays and the relation between the model parameters and the predictions with prediction intervals. Inspect carefully the appearance of the predictions at the observation locations.

**e)** Simulate 100 realizations from each of the two discretized posterior Gaussian RF models, and estimate empirically the prediction with associated prediction 0.9-intervals based on these realizations for each model.

Present the simulated realizations in two displays, one for each model, and over-print the corresponding empirically estimated predictions and prediction intervals in each display.

Discuss the relation between the model parameters and the realizations, and discuss the relation between the analytically and empirically obtained predictions with prediction intervals.

**f)** Consider the non-linear function on $\{r(x); x \in \mathtt{D}\}$,

$$A_r = \sum_{x \in \mathtt{L}} I(r(x) > 2)(r(x) - 2)$$

which approximates the area under the spatial variable and above level 2. Use the 100 realizations from the posterior model with $\sigma_\epsilon^2 = 0$ to provide a prediction $\hat{A}_r$ with associated prediction variance.

An alternative predictor for this area is based on the predicted spatial variable with $\sigma_\epsilon^2 = 0$ being $\{\hat{r}(x); x \in \mathtt{L}\}$,

$$\tilde{A}_r = \sum_{x \in \mathtt{L}} I(\hat{r}(x) > 2)(\hat{r}(x) - 2).$$

Calculate this prediction.

Consider the two predictions and the prediction variance of the former. Compare the predictions and use Jensen's inequality to explain why one expects $\hat{A}_r \geq \tilde{A}_r$.

**g)** Present a short summary of the experiences you have made on evaluating the model characteristics.

## Problem 2: Gaussian RF - real data

Consider observations of terrain elevation, available in the file topo.dat. The 52 observations are located in the domain $\mathtt{D} = [(0, 315) \times (0, 315)] \subset \mathbb{R}^2$ . Let the 52-vector of exact observations be $\boldsymbol{d} = (r(\boldsymbol{x}_1^o), \ldots, r(\boldsymbol{x}_{52}^o))^T$.

**a)** Display the observations in various ways. Is a stationary Gaussian RF a suitable model for the terrain elevation in domain D?
The functions interp(.), contour(.) and image.plot(.) in the R libraries akima and fields may be useful.

Let the terrain elevation in domain D be modeled by the Gaussian RF $\{r(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{D} \subset \mathbb{R}^2\}$ with

$$\begin{aligned} E\{r(\boldsymbol{x})\} &= \boldsymbol{g}(\boldsymbol{x})^T \boldsymbol{\beta}_r \\ Var\{r(\boldsymbol{x})\} &= \sigma_r^2 \\ Corr\{r(\boldsymbol{x}), r(\boldsymbol{x}')\} &= \rho_r(\tau/\xi) \end{aligned}$$

where $\boldsymbol{g(x)} = (1, g_2(\boldsymbol{x}), ..., g_{n_g}(\boldsymbol{x}))^T$ is a $n_g$-vector of known explanatory spatial variables on $\boldsymbol{x} \in \mathtt{D}$, and $\boldsymbol{\beta}_r = (\beta_1, \ldots, \beta_{n_g})^T$ is a $n_g$-vector of unknown parameters. Moreover, let the variance be $\sigma_r^2 = 2500$ and the spatial correlation function be $\rho_r(\tau) = \exp\{-(0.01\tau)^{1.5}\}$ with $\tau = |\boldsymbol{x} - \boldsymbol{x}'|$.

**b)** Develop the expression for the minimization problem to be solved for the universal kriging predictor and the associated prediction variance at an arbitrary location $\boldsymbol{x_0} \in \mathtt{D}$. The actual optimization need not be solved. Is it natural to change the value of the variance if the parameterization of expectation function is changed?

**c)** Consider the case with $E\{r(\boldsymbol{x})\} = \beta_1$, the so-called ordinary Kriging model.
Discretize the Gaussian RF to $\{r(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$ with the grid $\mathtt{L} : [(1, 315) \times (1, 315)] \in \mathtt{D}$. Calculate the universal Kriging predictor with associated prediction variance, $\{\hat{r}(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$ and $\{\sigma^2_{\hat{r}}(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$. Display the results and comment on them.
Use the function krige.conv(.) and the arguments trend.d and trend.l in krige.control(.) to specify the form of the expectation function - the function expand.grid(.) may also be useful.

**d)** Let the reference variable $\boldsymbol{x} \in \mathtt{D} \subset \mathbb{R}^2$ be denoted $\boldsymbol{x} = (x_v, x_h)$, set $n_g = 6$ and define the set of known polynomial functions $\boldsymbol{g}(\boldsymbol{x})$ to be all polynomials $x_v^k x_h^l$ for $(k, l) \in \{(0, 0), (1, 0), (0, 1), (1, 1), (2, 0), (0, 2)\}$.
Specify the resulting $n_g$-vector $\boldsymbol{g}(\boldsymbol{x})$ and the expected value of $r(\boldsymbol{x})$.
Discretize the Gaussian RF to $\{r(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$ with the grid $\mathtt{L} : [(1, 315) \times (1, 315)] \in \mathtt{D}$. Calculate the universal Kriging predictor with associated prediction variance, $\{\hat{r}(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$ and $\{\sigma^2_{\hat{r}}(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$. Display the results and comment on them.
Use the function krige.conv(.) and the arguments trend.d and trend.l in krige.control(.) to specify the form of the expectation function - the function expand.grid(.) may also be useful.

**e)** Use the ordinary Kriging predictor with associated prediction variance and consider grid node $\boldsymbol{x}_0 = (100, 100)$. Calculate the probability for the elevation to be higher than 850 m at this location. Further, calculate the elevation for which it is 0.90 probability that the true elevation is below it.

**f)** Present a short summary of the experiences you have made on evaluating the real data.

## Problem 3: Parameter estimation

Consider the stationary Gaussian RF $\{r(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{D} \subset \mathbb{R}^2\}$ with $\mathtt{D} : [(1, 30) \times (1, 30)]$, with

$$\begin{aligned} E\{r(\boldsymbol{x})\} &= \mu_r = 0 \\ Var\{r(\boldsymbol{x})\} &= \sigma_r^2 \\ Corr\{r(\boldsymbol{x}), r(\boldsymbol{x}')\} &= \exp\{-\tau/\xi_r\} \end{aligned}$$

with $\tau = |\boldsymbol{x} - \boldsymbol{x}'|$.

**a)** Consider the discretized Gaussian RF $\{r(\boldsymbol{x}); \boldsymbol{x} \in \mathtt{L}\}$ on grid $\mathtt{L} : [(1, 30) \times (1, 30)] \in \mathtt{D}$. Set the model parameters $\sigma_r^2 = 2$ and $\xi_r = 3$ and generate one realization of the discretized Gaussian RF and display it.

**b)** Compute the empirical variogram based on exact observations of the full realization, and display the estimate jointly with the correct variogram function. Comment on the result, particularly the precision of the estimates due to finite domain `D`.

Use the function variog(.) in GeoR .

**c)** Repeat point **a)** and **b)** three times. Comment on the results.

**d)** Generate 36 locations uniformly randomly in the grid `L`. Compute the empirical variogram estimate based on the corresponding 36 exact observations. Display the estimate jointly with the correct variogram function, and comment on the results.

Consider the model parameters variance $\sigma_r^2$ and $\xi_r$ to be unknown. Estimate the parameters by a maximum likelihood criterion based on exact observation of the full realization and based on the 36 exact observations. Display the corresponding estimated variogram functions jointly with the correct variogram function, and comment on the result.

Use the function likfit(.) in GeoR .

Discuss and compare all the results above.

**e)** Repeat the procedure in **d)** with 9, 64 and 100 uniformly randomly generated exact observations from the realization. Present the estimates jointly with the correct variogram function in separate displays, and comment on the results.

**f)** Present a short summary of the experiences you have made on parameter estimation.