Supplementary Materials of 'Estimating the Spatial Covariance Structure using the Geoadditive Model'

Yannick Vandendijck
Interuniversity Institute for Biostatistics and statistical Bioinformatics,
Hasselt University, B3590 Diepenbeek, Belgium

and

Christel Faes
Interuniversity Institute for Biostatistics and statistical Bioinformatics,
Hasselt University, B3590 Diepenbeek, Belgium

and

Niel Hens
Interuniversity Institute for Biostatistics and statistical Bioinformatics,
Hasselt University, B3590 Diepenbeek, Belgium
Centre for Health Economic Research and Modeling Infectious Diseases,
Vaccine and Infectious Disease Institute, University of Antwerp,
B2610 Wilrijk, Belgium

Journal: Environmental and Ecological Statistics
Corresponding author: `yannick.vandendijck@uhasselt.be`

In the Supplementary Materials more information is provided concerning inference from the geoadditive model, additional simulation results, additional results with respect to data applications, and computational details.

- Section 1: In this section we present the most important and often used functions g_τ and some more inference details for geoadditive models. More specific, we focus on the amount of smoothing of the spatial component and provide recommendations with respect to plotting of non-linear effects.
- Section 2: In this section we present additional simulation results of the simulation study presented in Section 4 of the manuscript. We also present the results of a simulation setting including clustering which is not presented in the manuscript.
- Section 3: In this section we give additional results of the two data applications presented in the paper. More specific, we provide tables of estimated parameters, investigate the influence of the number of knots and provide additional plots of the estimated effects. In addition, we present the results of an analysis of the zinc-levels in the Meuse dataset.
- Section 4: In this section we present computational details to implement the proposed methods in `R` software. We present some details on how all described methods in the paper can be implemented in `R` software.

1 Additional Concepts for the Geoadditive Model

Functions g_τ

In Table 1 below we present an overview of the most important (generalized) covariance functions that can be used as functions g_τ

Table 1: Some important and often used (generalized) covariance functions that can be used to model the spatial component S in the geoadditive model (3).

Exponential	$g_\tau(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|}{\tau}\right)$
Gaussian	$g_\tau(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|^2}{\tau^2}\right)$
Spherical	$g_\tau(\mathbf{x}) = \left(1 - \frac{3}{2}\frac{\|\mathbf{x}\|}{\tau} + \frac{1}{2}\frac{\|\mathbf{x}\|^3}{\tau^3}\right) I_{\|\mathbf{x}\|<\tau}$
Matérn ($\nu = \frac{3}{2}$)	$g_\tau(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}\|}{\tau}\right)\left(1 + \frac{\|\mathbf{x}\|}{\tau}\right)$
Circular	$g_\tau(\mathbf{x}) = 1 - \frac{2}{\pi}\left(\vartheta\sqrt{1-\vartheta^2} + \arcsin\vartheta\right)$, with $\vartheta = \min\left(\frac{\|\mathbf{x}\|}{\tau}, 1\right)$
Thin plate	$g(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|^k & , k = 1,3,5,\ldots \\ \|\mathbf{x}\|^k \ln(\|\mathbf{x}\|) & , k = 2,4,6,\ldots \end{cases}$

The parameter τ is positive for each function.

Amount of smoothing

The amount of smoothing is quintessential for penalized splines. For the geoadditive model with linear mixed model representation in (3), the variance ratios $\sigma_\varepsilon^2/\sigma_b^2$ and $\sigma_\varepsilon^2/\sigma_S^2$ quantify the amount of smoothness of the non-linear and spatial components f and S (Ruppert et al., 2003). Intuitively, large values of σ_S^2 lead to less penalization and an overfit of the spatial component, whereas a very small value of σ_S^2 leads to a constant spatial component S. Similar arguments hold for σ_b^2. It is natural to replace the variances σ_ε^2, σ_b^2 and σ_S^2 with their (restricted) maximum likelihood estimates to choose the amount of smoothness. This automatic selection of the amount of smoothing is an attractive feature of using the linear mixed model approach for fitting penalized splines. However, Ruppert et al. (2003) discourage blind acceptance of whatever answer this gives and recommend looking at other amounts of smoothing. The work of Chaudhuri and Marron (1999) is recommended to investigate other amounts of smoothing, but it does not fall within the scope of this paper.

An important and attractive tool to quantify the amount of smoothness is the computation of the degrees-of-freedom values. Linear terms have one degree of freedom, whereas the non-linear and spatial components could have positive non-integer degrees-of-freedom. The degrees-of-freedom can be calculated for the entire fit or for each component separately using standard matrix operations. Details are not provided here, we recommended Section 8.3 (p. 174-176) in Ruppert et al. (2003) for a full overview on degrees-of-freedom calculations for additive penalized spline models.

Plotting issues

To facilitate the interpretation of an estimated effect of a non-linear covariate, the curve estimate can be constructed by performing vertical alignment around zero. Using this plotting convention, a curve estimate of a non-linear covariate represents how the response changes relative to its mean with changes in the value of that covariate whilde holding the other covariates in the model constant (Ruppert et al., 2003). This can be done by partitioning $\mathbf{C}$ into $[\mathbf{1}|\mathbf{C}_r]$ (a partition into the intercept column and the remainder) and calculating $\tilde{\mathbf{C}} = \big(\mathbf{1}|(\mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^T)\mathbf{C}_r\big)$. The curve estimate is then constructed by matrix multiplication of the columns corresponding to the non-linear covariate of interest in $\tilde{\mathbf{C}}$ and the corresponding elements in $[\hat{\boldsymbol{\beta}}^T, \hat{\tilde{\mathbf{u}}}^T]^T$. This can also be applied to the surface estimate of the spatial component S.

2 Additional Simulation Results

In Section 4 of the paper an elaborate simulation study is described to evaluate the proposed methodology under controlled settings. The simulation study is divided into two main scenarios: (1) No covariates, and (2) with covariates in the mean effect function. In this section we provide additional tables and figures for each scenario. In addition, We present the results of a simulation setting including clustering which is not presented in the manuscript.

No Covariates

Boxplots of the estimated covariogram parameters (c_s, τ), the measurement error parameter σ_ε^2 and the ratio c_s/τ are presented in Figures 1-3. Estimated covariogram parameters using GM1$_{ML}$ and GM1$_{REML}$ are seriously biased, which also affects the estimation of the measurement error parameter σ_ε^2. GM2$_{ML}$ and GM2$_{REML}$ perform well and yield similar results as the direct likelihood approaches D$_{ML}$ and D$_{REML}$. Figure 4 presents the boxplots of the prediction bias averaged over the five spatial locations. The prediction bias is negligible for all considered methods. Estimated covariograms based on the proposed methodology of the 250 simulates are presented in Figure 5, together with the true covariogram and averaged covariogram over all 250 simulations. It can be observed that covariograms are estimated well.

Misspecification, in the sense that a wrong covariance is considered, is also investigated. We present two simulation results here. First, suppose data is simulated as explained in Section 4 using (14) with an exponential covariogram. Data are analysed using GM2$_{ML}$ with the exponential function g_τ (correct specification) and the Matérn function g_τ (misspecification). The estimated covariogram functions obtained from 250 simulations are presented in Figure 6. It can be observed that the misspecified covariance function still yields a good approximation of the true covariogram. The MSE of the predictive performance is 18.24 for the exponential function g_τ (see Table 2 in the manuscript). The MSE with the Matérn function g_τ is 18.49 and thus very similar. A second example is presented in Figure 7. Here, data is simulated as explained in Section 4 using (14) with a Gaussian covariogram. Data are analysed using GM2$_{ML}$ with the Gaussian function g_τ (correct specification) and the spherical function g_τ (misspecification). Again, the misspecified function g_τ yields an acceptable approximation of the true covariogram. The MSE of the predictive performance was 2.35 for the Gaussian function g_τ (see Table 2 in the manuscript). The MSE with the spherical function g_τ is 2.77 and thus very similar.

In Table 2 we also present simulation results on the predictive performance when thin plate splines are used as function g_τ in the spatial component in the geoadditive model. Data is simulated according to model (14) as discussed in Section 4 of the paper. Thin plate spline basis functions are often used to compare the performance of kriging and splines (Dubrule, 1984; Hutchinson and Gessler, 1994; Laslett, 1994; Altman, 2000). It is observed that MSE values are higher for thin plate splines as compared to the results in Table 2 of the manuscript.

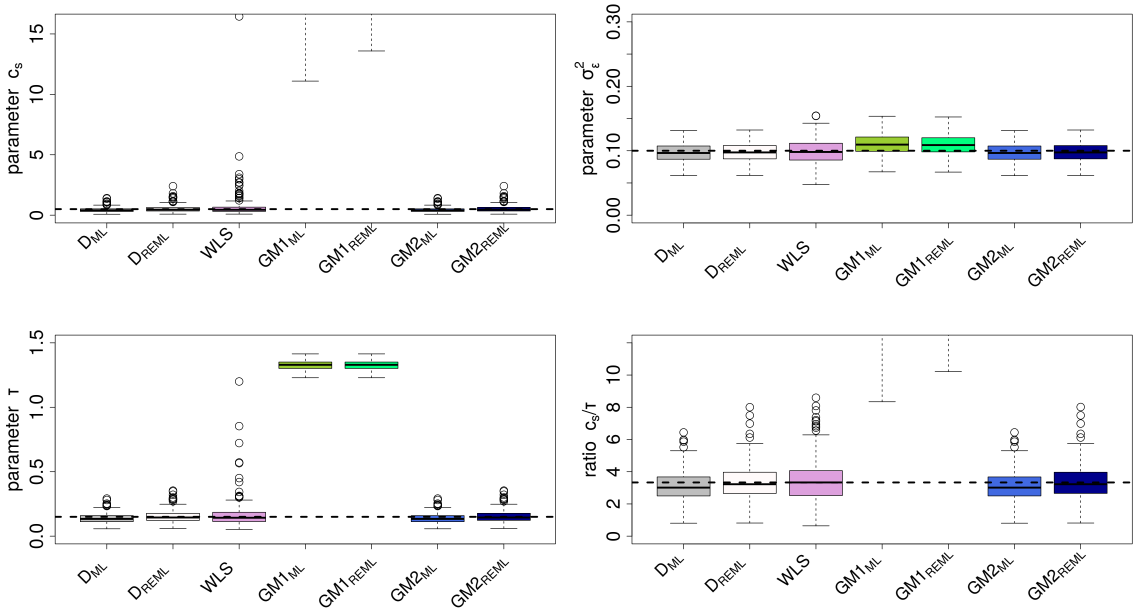

Figure 1: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ^2_ε over 250 simulations using seven estimation methods. Data is simulated according to model (14) in the paper with the Matérn covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

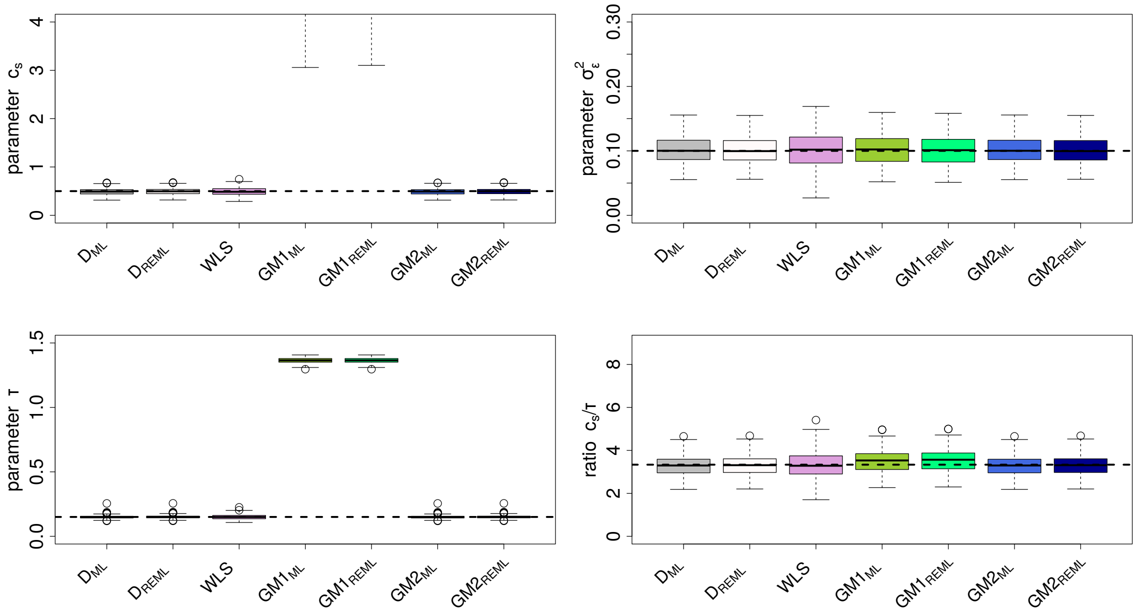

Figure 2: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ^2_ε over 250 simulations using seven estimation methods. Data is simulated according to model (14) in the paper with the spherical covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

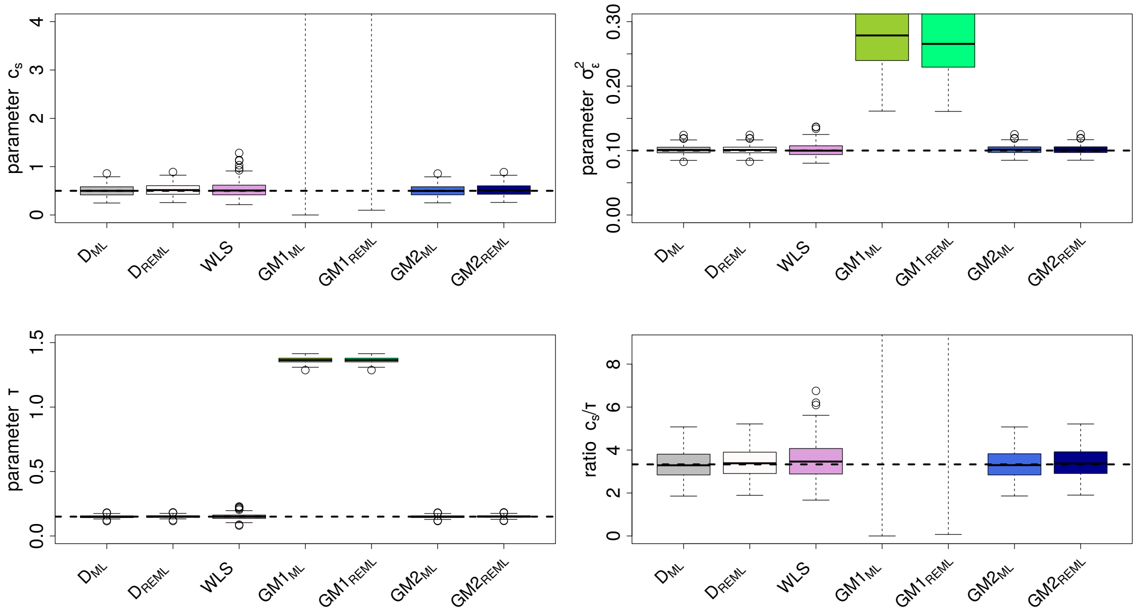

Figure 3: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ_ε^2 over 250 simulations using seven estimation methods. Data is simulated according to model (14) in the paper with the Gaussian covariogram model. The range of the boxplots of $\mathrm{GM1}_{ML}$ and $\mathrm{GM1}_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

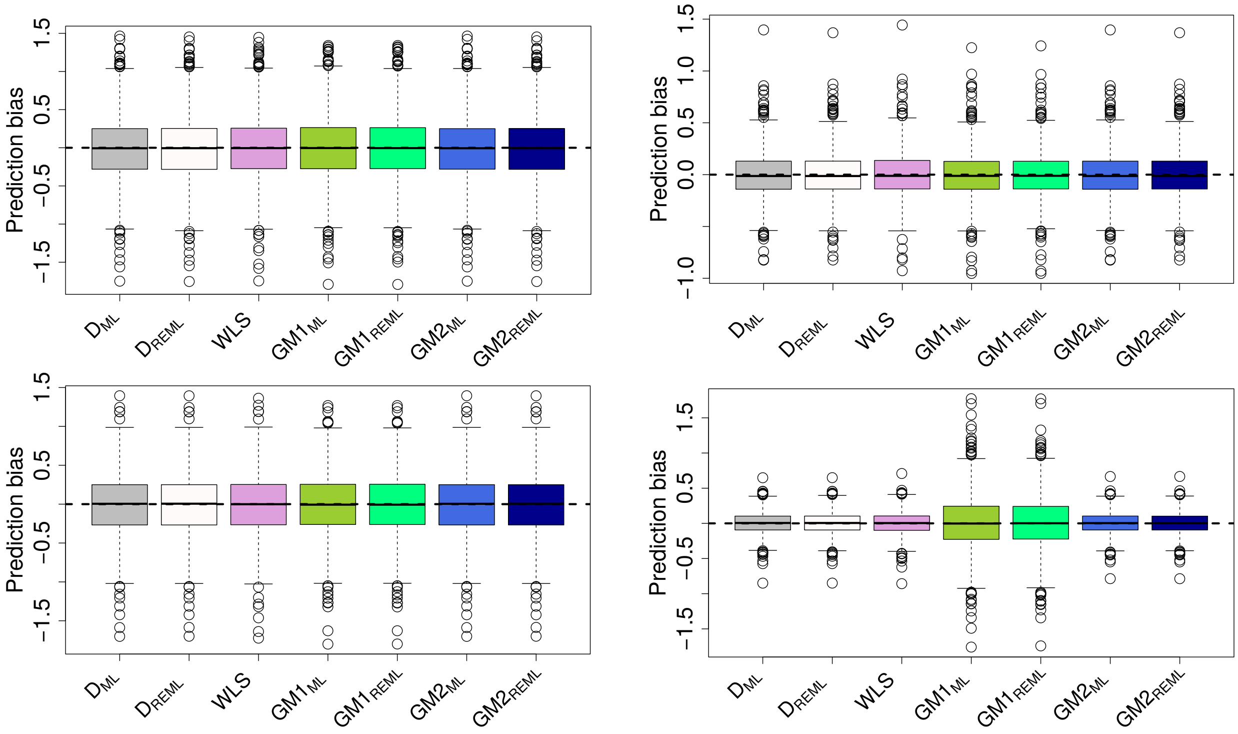

Figure 4: Prediction bias averaged over five spatial locations obtained from 250 simulations using seven estimation methods. Data is simulated according to model (14) in the paper with the exponential covariogram model (top left), the Matérn covariogram model (top right), the spherical covariogram model (bottom left) and the Gaussian covariogram model (bottom right).

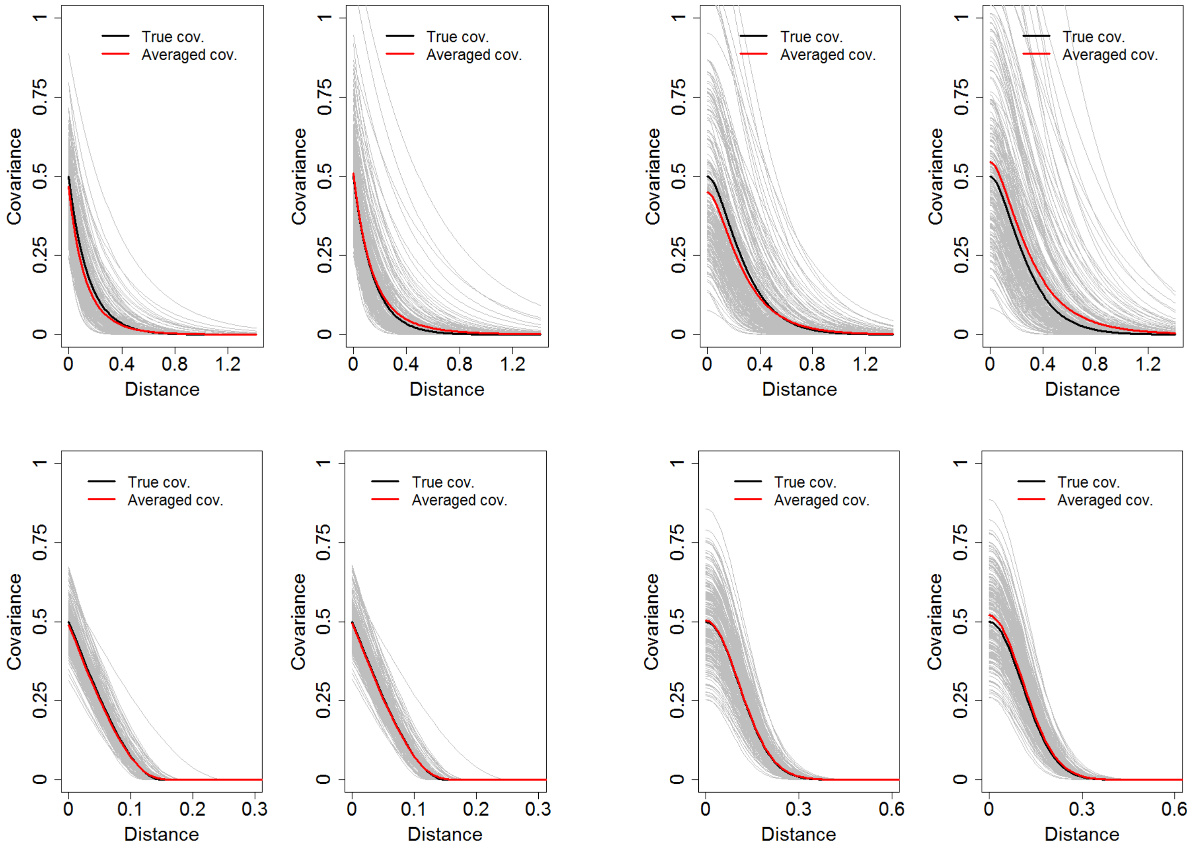

Figure 5: Estimated covariogram functions obtained from 250 simulations using GM2$_{ML}$ (left in each subplot) and GM2$_{REML}$ (right in each subplot). The thick black line represents the true covariogram function used in the simulation study and the thick red line represents the average over the estimated covariogram functions. Data is simulated according to model (14) in the paper with the exponential covariogram model (top left), the Matérn covariogram model (top right), the spherical covariogram model (bottom left), the Gaussian covariogram model (bottom right).

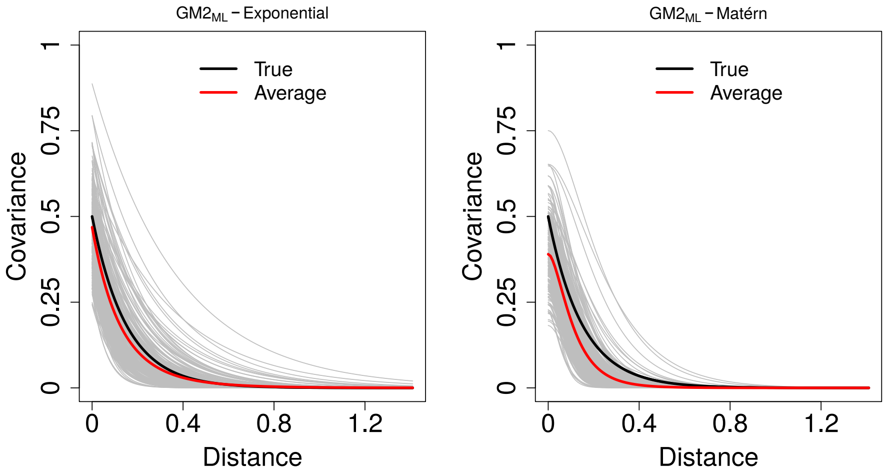

Figure 6: Estimated covariogram functions using GM2$_{ML}$ obtained from 250 simulations with the exponential function g_τ (left) and the Matérn function g_τ (right). The thick black line represents the true covariogram function and the thick red line represents the average over the estimated covariogram functions. Data is simulated according to (14) in the paper with the exponential covariogram model.

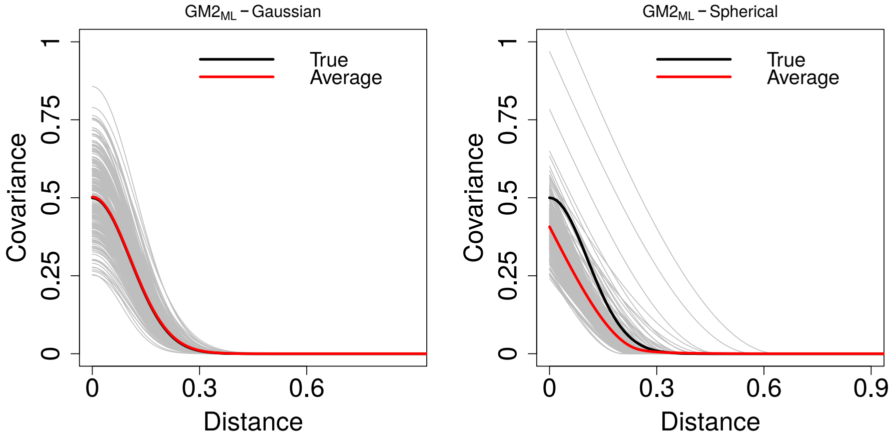

Figure 7: Estimated covariogram functions using GM2$_{ML}$ obtained from 250 simulations with the Gaussian function g_τ (left) and the spherical function g_τ (right). The thick black line represents the true covariogram function and the thick red line represents the average over the estimated covariogram functions. Data is simulated according to (14) in the paper with the Gaussian covariogram model.

Table 2: MSE (×100) of the predictive performance and the nominal coverage of the 95% confidence intervals over 250 simulation using thin plate splines as function g_τ in the geoadditive model. Data is simulated according to model (14) in the paper with four different covariogram models.

True underlying covariogram		ML	REML
Exponential			
	MSE (×100)	21.20	21.11
	95% coverage	74.2	74.4
Matérn			
	MSE (×100)	5.10	5.11
	95% coverage	91.1	91.1
Spherical			
	MSE (×100)	17.11	17.08
	95% coverage	75.2	75.3
Gaussian			
	MSE (×100)	2.81	2.81
	95% coverage	89.7	89.7

Covariates

Figure 8 presents the boxplots of the prediction bias averaged over the five spatial locations. The prediction bias is negligible for all considered methods. D$_{ML}$, D$_{REML}$, WLS, GM1$_{ML}$ and GM1$_{REML}$ are associated with higher variability of the obtained predictions. Boxplots of the estimated covariogram parameters (c_s, τ), the measurement error parameter σ_ε^2 and the ratio c_s/τ are presented in Figures 9 and 10. Estimated covariogram parameters using GM1$_{ML}$ and GM1$_{REML}$ are seriously biased. Entering the non-linear covariate x_2 linearly in the mean function for D$_{ML}$, D$_{REML}$ and WLS has an effect on the estimation of the covariogram parameters. It is observed that GM2$_{ML}$ and GM2$_{REML}$ perform the best.

The estimated non-linear covariate effect of x_2 over all 250 simulations are presented in Figures 11 and 12. It is observed that the geoadditive model estimated with the proposed methodology in Section 3 produce less variable estimates of the effect of x_2. In Figure 13 we show the estimates of the linear effect parameter of x_1. Again, it is observed that the geoadditive model estimated with the proposed methodology in Section 3 produce less variable estimates.

In Table 3 we also present simulation results on the predictive performance when thin plate splines are used as function g_τ in the spatial component in the geoadditive model. Data is simulated according to model (15) as discussed in Section 4 of the paper. Thin plate spline basis functions are often used to compare the performance of kriging and splines (Dubrule, 1984; Hutchinson and Gessler, 1994; Laslett, 1994; Altman, 2000). It is observed that MSE values are higher for thin plate splines as compared to the results in Table 3 of the manuscript.

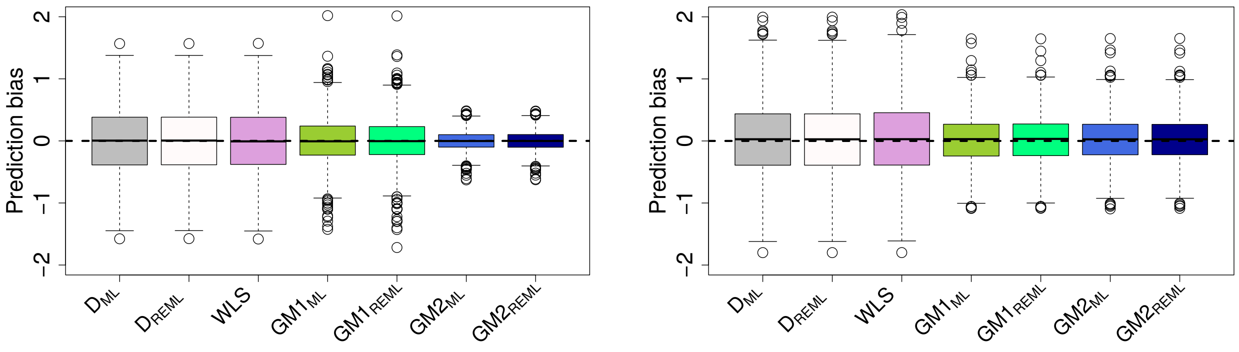

Figure 8: Prediction bias averaged over five spatial locations obtained from 250 simulations using seven estimation methods. Data is simulated according to model (15) in the paper with the Gaussian covariogram model (left) and the circular covariogram model (right).

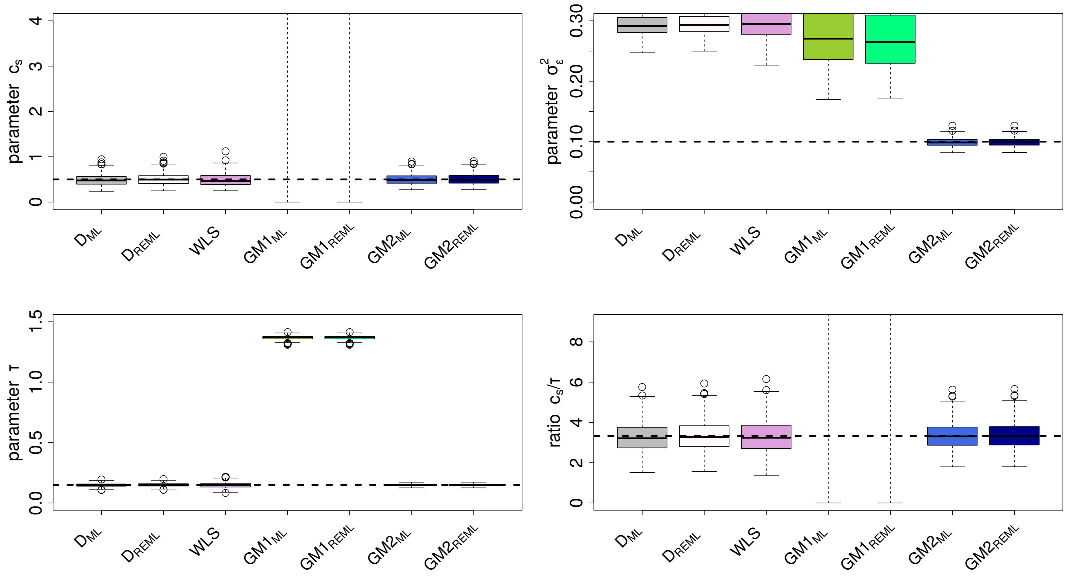

Figure 9: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ^2_ε over 250 simulations using seven estimation methods. Data is simulated according to model (15) in the paper with the Gaussian covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

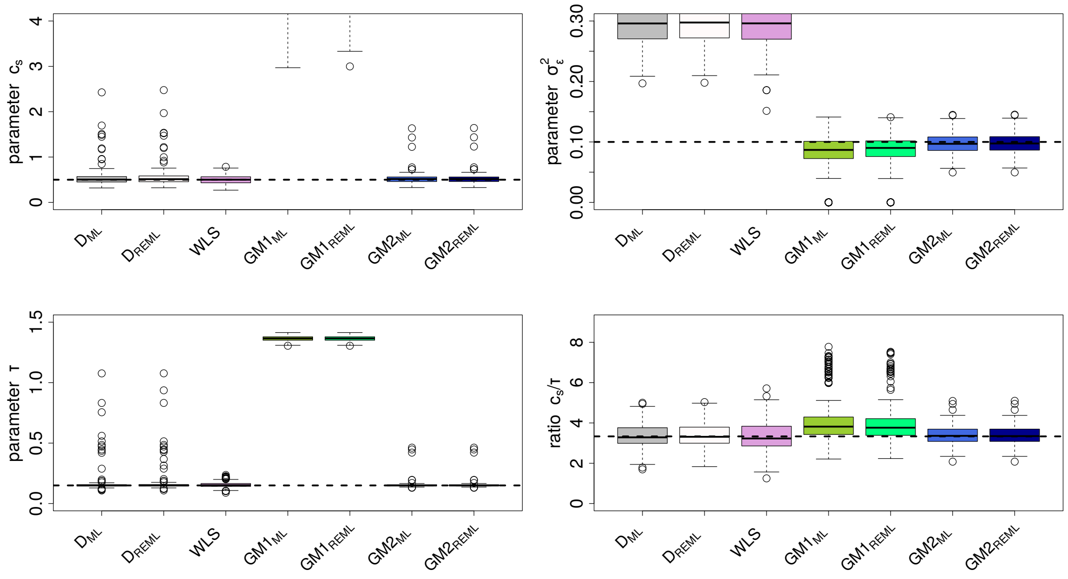

Figure 10: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ^2_ε over 250 simulations using seven estimation methods. Data is simulated according to model (15) in the paper with the circular covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

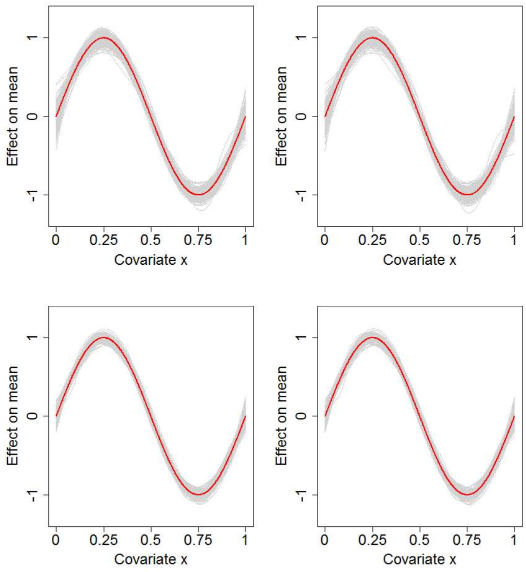

Figure 11: The estimated non-linear effects of the covariate x_2 from 250 simulations using GM1$_{ML}$ (top left), GM1$_{REML}$ (top right), GM2$_{ML}$ (bottom left) and GM2$_{REML}$ (bottom right) as estimation methods. The red line indicates the true non-linear effect of x_2. Data is simulated according to model (15) in the paper with the Gaussian covariogram model.

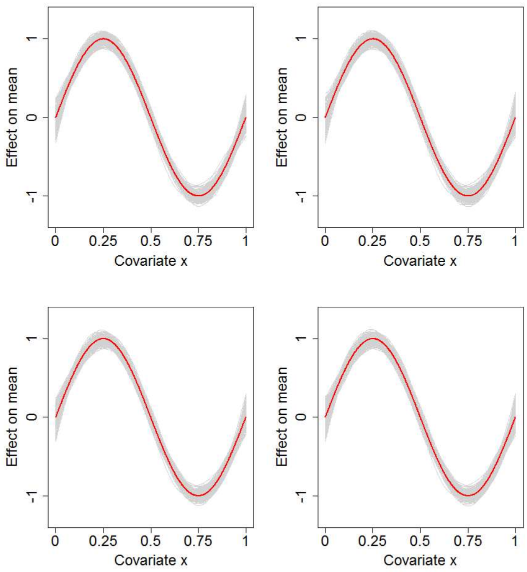

Figure 12: The estimated non-linear effects of the covariate x_2 from 250 simulations using GM1$_{ML}$ (top left), GM1$_{REML}$ (top right), GM2$_{ML}$ (bottom left) and GM2$_{REML}$ (bottom right) as estimation methods. The red line indicates the true non-linear effect of x_2. Data is simulated according to model (15) in the paper with the circular covariogram model.

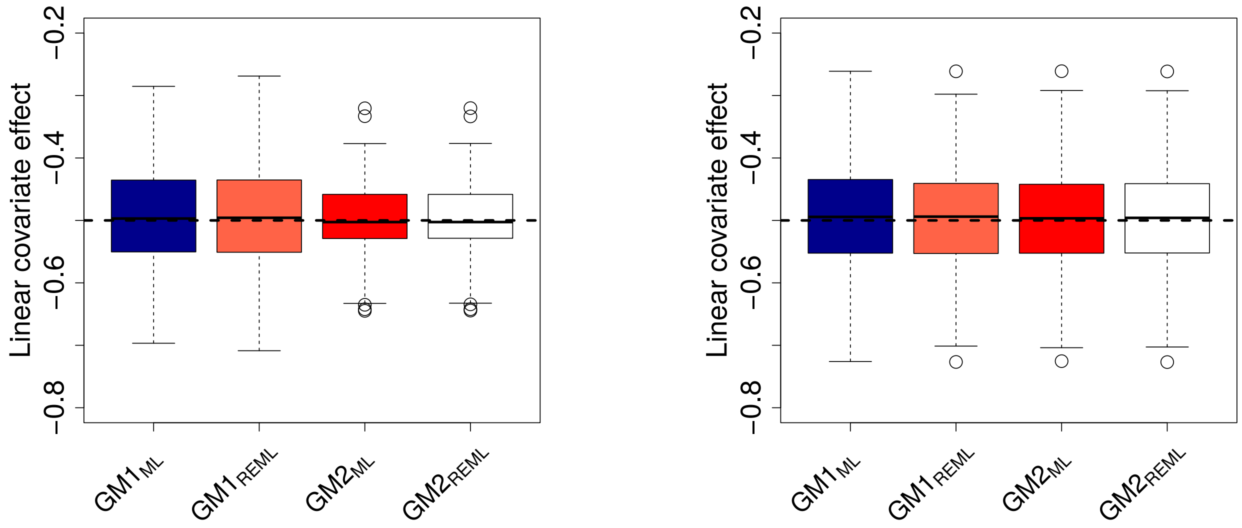

Figure 13: Boxplots of the estimated linear effect of the covariate x_1 from 250 simulations using GM1$_{ML}$, GM1$_{REML}$, GM2$_{ML}$ and GM2$_{REML}$ as estimation methods. Data is simulated according to model (15) in the paper with the Gaussian covariogram model (left) and the circular covariogram model (right).

Table 3: MSE (×100) of the predictive performance and the nominal coverage of the 95% confidence intervals over 250 simulation using thin plate splines as function g_τ in the geoadditive model. Data is simulated according to model (15) in the paper with two different covariogram models.

True underlying covariogram		ML	REML
Gaussian			
	MSE (×100)	2.73	2.73
	95% coverage	91.8	92.0
Circular			
	MSE (×100)	15.34	15.47
	95% coverage	76.6	76.8

Clustering

Additionally, we investigate the effect of clustering, meaning that measurements from the same cluster (*e.g.*, similar soil type, similar industry classification) are more similar to each other than measurements from other clusters, in thegeo additive model. Data on the unit square is simulated using model (14) in the manuscript, where again $\varepsilon_{\mathbf{s}} \sim \mathcal{N}(0, \sigma_\varepsilon^2 = 0.10)$ and $S(\mathbf{s})$ is a GRF with a Matérn or circular covariogram model. Similarly, we take $c_s = 0.50$ and $\tau = 0.15$. 250 realizations are simulated from which each time a sample of $n = 200$ is obtained. We assume that there are twenty clusters of size ten. Each sampled data location is randomly classified into one of these twenty clusters and a clustering effect is introduced by adding a random effect b_i to the simulated data value. The values of the random effects are simulated from $b_i \sim \mathcal{N}(0, \sigma_b^2 = 0.10)$, $i = 1, \ldots, 20$. The estimation methods D$_{ML}$, D$_{REML}$, WLS, GM1$_{ML}$, GM1$_{REML}$, GM2$_{ML}$ and GM2$_{REML}$ are used to obtain estimates of the covariogram parameters, the measurement error parameter σ_ε^2, the cluster variance σ_b^2 and predictions at five spatial locations. For D$_{ML}$, D$_{REML}$, WLS, the clustering is ignored (this is typically done in kriging), whereas the the geoadditive models take into account the clustering by a random intercept. For both covariogram models, all sampled locations are used as knots.

Figures and tables summarizing the results are presented below. In the main, the methods D$_{ML}$ and D$_{REML}$ have a slightly better performance than GM2$_{ML}$ and GM2$_{REML}$ in terms of MSE (see Table 4) for the parameter c_s because of a lower variance associated to the estimates of D$_{ML}$ and D$_{REML}$ (see Figures 14 and 15). However, the ratio c_s/τ is much better estimated for GM2$_{ML}$ and GM2$_{REML}$ in comparison with the other methods. D$_{ML}$, D$_{REML}$ and WLS all yield biased estimates of the ratio c_s/τ. The measurement error is only estimated without bias by the geoadditive model estimated using the proposed methodology (GM2$_{ML}$ and GM2$_{REML}$). The cluster variance σ_b^2 is well estimated by the geoadditive models (Figure 17).

All methods produce unbiased predictions (see Figure 16). However, the associated variance of the predictions is smallest for GM2$_{ML}$ and GM2$_{REML}$, which results in lower MSE values when compared with the other methods (see Table 5). The bootstrap prediction variance is again needed to obtain acceptable nominal coverages. Nevertheless, a slight undercoverage of the 95% confidence intervals is still observed for the Matérn covariogram model using the boostrap variance (around 92.8%).

In Table 6 we also present simulation results on the predictive performance when thin plate splines are used as function g_τ in the spatial component in the geoadditive model. Data is simulated according to the text above. Thin plate spline basis functions are often used to compare the performance of kriging and splines (Dubrule, 1984; Hutchinson and Gessler, 1994; Laslett, 1994; Altman, 2000). It is observed that MSE values are higher for thin plate splines as compared to the results in Table 5.

Table 4: MSE (×100) of the covariogram parameters (c_s, τ), the measurement error parameter σ_ε^2 and the ratio c_s/τ over 250 simulations using seven estimation methods. Data is simulated according to the text above with two different covariogram models.

	D_{ML}	D_{REML}	WLS	$\mathrm{GM1}_{ML}$	$\mathrm{GM1}_{REML}$	$\mathrm{GM2}_{ML}$	$\mathrm{GM1}_{REML}$
				Matérn covariogram			
c_s	3.70	6.30	35.90	$> 10^3$	$> 10^3$	4.71	8.76
τ	0.30	0.29	0.67	138.74	138.74	0.16	0.22
σ_ε^2	0.32	0.35	0.28	0.07	0.04	0.02	0.02
c_s/τ	247.88	287.85	419.02	$> 10^3$	$> 10^3$	90.50	114.29
				Circular covariogram			
c_s	1.94	3.93	1.89	$> 10^3$	$> 10^3$	4.60	5.02
τ	0.15	0.51	0.10	138.80	138.80	0.29	0.51
σ_ε^2	0.31	0.30	0.44	1.83	0.57	0.17	0.17
c_s/τ	110.85	116.70	226.13	181.83	100.28	56.28	48.69

Table 5: MSE (×100) of the predictive performance and the nominal coverage of the 95% confidence intervals over 250 simulation using seven estimation methods (see text). Data is simulated according to the text above with two different covariogram models.

	D_{ML}	D_{REML}	WLS	$\mathrm{GM1}_{ML}$	$\mathrm{GM1}_{REML}$	$\mathrm{GM2}_{ML}$	$\mathrm{GM1}_{REML}$
				Matérn covariogram			
MSE (×100)	15.07	14.99	15.14	10.49	9.77	9.64	9.49
95% coverage	80.3	80.2	80.4	91.9	92.9	91.2	91.5
95% coverage[a]						92.7	92.8
				Circular covariogram			
MSE (×100)	36.34	36.38	37.23	34.14	32.97	28.20	28.22
95% coverage	89.2	89.4	88.2	71.1	74.0	70.6	71.0
95% coverage[a]						94.2	94.4

a: Based on the boostrap procedure described in Section 3.

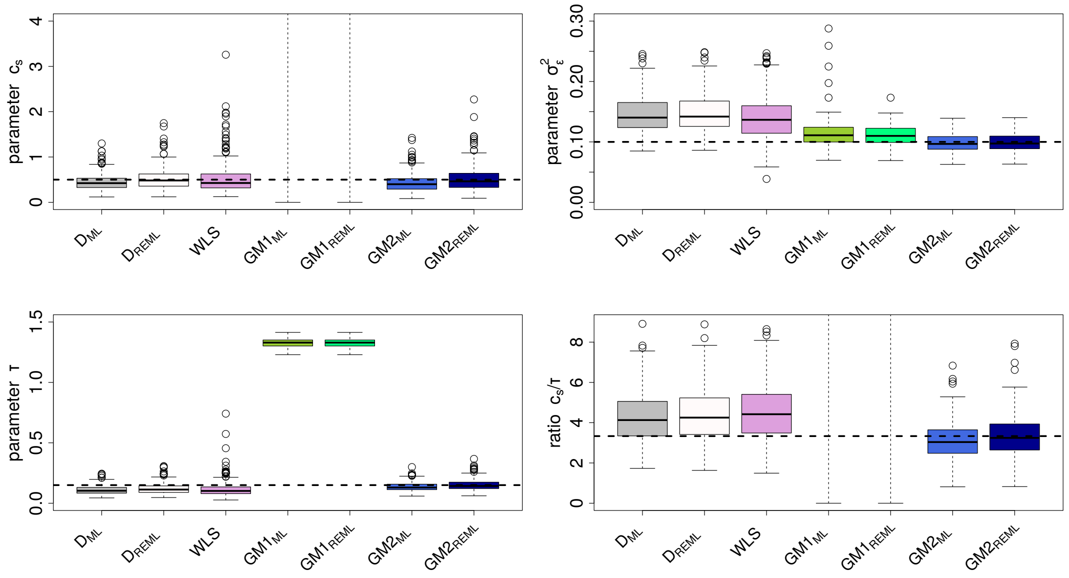

Figure 14: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ_ε^2 over 250 simulations using seven estimation methods. Data is simulated according to the text above with the Matérn covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

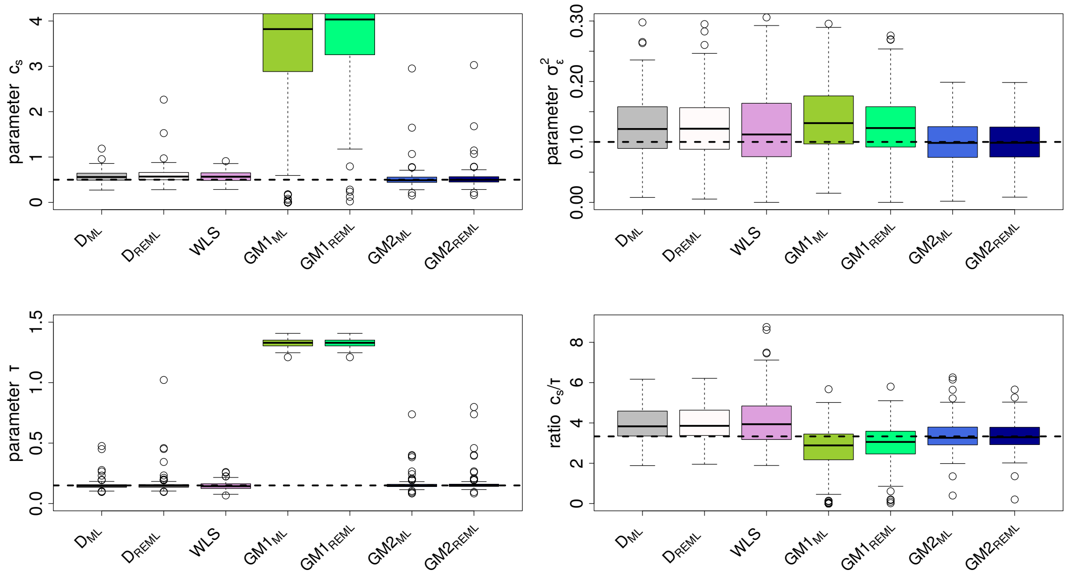

Figure 15: Boxplots of the estimated covariogram parameters (c_s, τ) and the measurement error parameter σ_ε^2 over 250 simulations using seven estimation methods. Data is simulated according to the text above with the circular covariogram model. The range of the boxplots of GM1$_{ML}$ and GM1$_{REML}$ exceed the range of the y-axis used and are, therefore, not fully depicted on the figure.

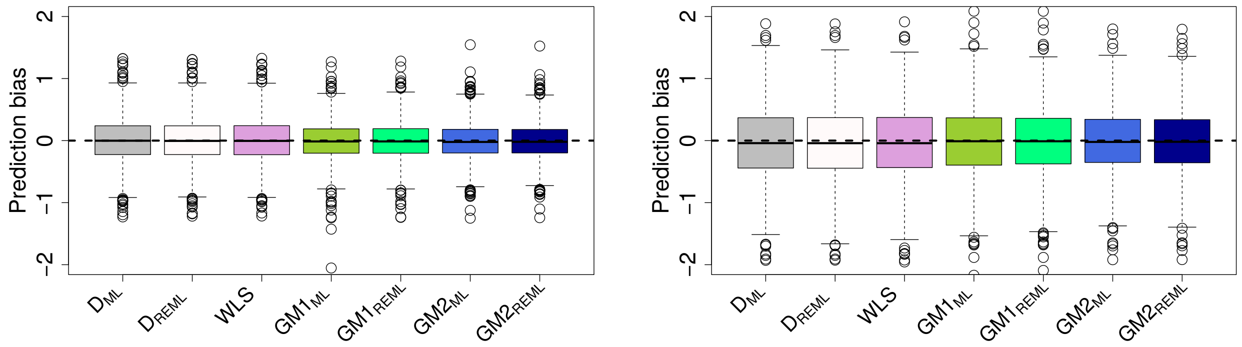

Figure 16: Prediction bias averaged over five spatial locations obtained from 250 simulations using seven estimation methods. Data is simulated according to the text above with the Matérn covariogram model (left) and the circular covariogram model (right).

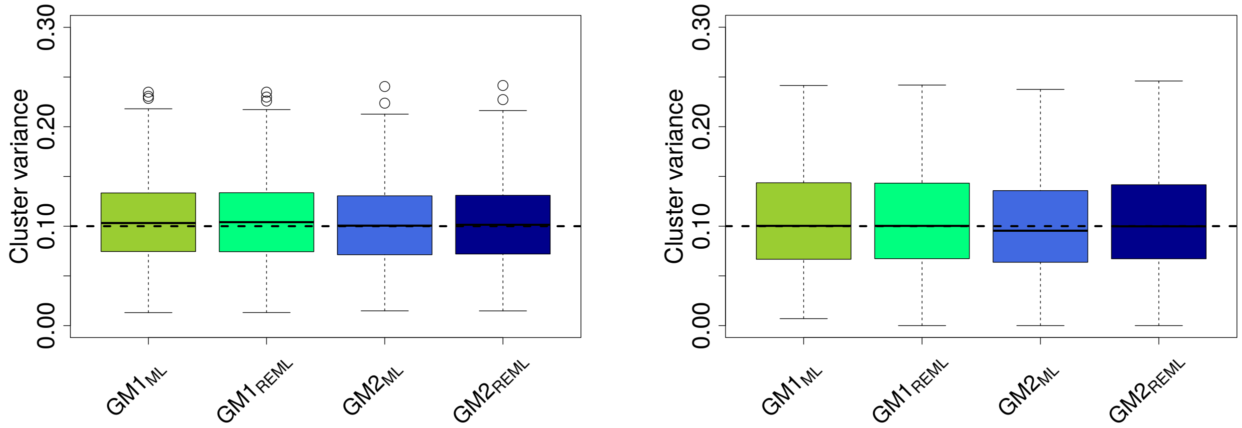

Figure 17: Boxplots of the estimated cluster variance from 250 simulations using GM1$_{ML}$, GM1$_{REML}$, GM2$_{ML}$ and GM2$_{REML}$ as estimation methods. Data is simulated according to the text above with the Matérn, covariogram model (left) and the circular covariogram model (right).

Table 6: MSE (×100) of the predictive performance and the nominal coverage of the 95% confidence intervals over 250 simulation using thin plate splines as function g_τ in the geoadditive model. Data is simulated according to the text above with two different covariogram models.

True underlying covariogram		ML	REML
Matérn			
	MSE (×100)	10.40	10.40
	95% coverage	93.2	93.2
Circular			
	MSE (×100)	38.92	38.55
	95% coverage	75.7	79.0

3 Data Application

Here, we give additional results of the two data applications presented in the paper. More specific, we provide tables of estimated parameters, investigate the influence of the number of knots and provide additional plots of the estimated effects. In addition, we present the results of an analysis of the zinc-levels in the Meuse dataset.

Paraná State, Brazil, Rainfall Data
To analyse this data model (16) was considered. For the spatial component S we used the exponential, Matérn, circular, spherical and Gaussian as g_τ function. The lowest AIC value for the geoadditive model fitted using the proposed methodology (GM2$_{ML}$) is observed for a circular g_τ function (see Table 7). We observe that for each choice of the function g_τ, all estimated covariogram parameters using the geoadditive model fitted using the proposed methodology (GM2$_{ML}$) were almost similar as results obtained by direct likelihood estimation (D$_{ML}$). This can be expected, since model (16) only includes linear covariate effects. It was also observed that geoadditive models fixing τ at (6), namely GM1$_{ML}$, have higher AIC values than the geoadditive models fitting τ using the proposed estimation procedure in Section 3 (GM2$_{ML}$). It is also observed that thin plate splines perform worse in terms of AIC.

All data locations are used as knots in the analysis. From Figure 18 it can be observed that the number of knots used can only be reduced slightly for this data example. From 120 knots onwards stable results are obtained for the AIC value and covariogram parameter estimates.

Table 7: Parameter estimates and AIC values of the Paraná rainfall data using model (16) with different estimation approaches.

Method	$g_\tau(\cdot)$	$\hat{c}_s$	$\hat{\tau}$	$\hat{\sigma}^2_\varepsilon$	AIC
D$_{ML}$	exponential	785.6	184.4	385.5	1339.7
D$_{ML}$	Gaussian	840.6	211.7	509.9	1338.2
D$_{ML}$	spherical	717.3	378.1	410.9	1336.0
D$_{ML}$	circular	706.4	330.4	415.0	1335.3
D$_{ML}$	Matérn	783.1	86.7	460.2	1337.9
GM1$_{ML}$	exponential	2157.7	619.5	399.0	1341.3
GM1$_{ML}$	Gaussian	39132.6	619.5	533.7	1352.1
GM1$_{ML}$	spherical	1602.2	619.5	402.8	1340.7
GM1$_{ML}$	circular	1602.2	619.5	402.8	1340.7
GM1$_{ML}$	Matérn	49013.4	619.5	491.3	1347.3
GM2$_{ML}$	exponential	785.7	184.4	385.5	1339.7
GM2$_{ML}$	Gaussian	840.5	211.7	509.9	1338.2
GM2$_{ML}$	spherical	717.3	378.1	410.9	1336.0
GM2$_{ML}$	circular	706.4	330.4	415.0	1335.3
GM2$_{ML}$	Matérn	783.1	86.7	460.2	1337.9
GM$_{ML}$	thin plate			469.5	1348.3

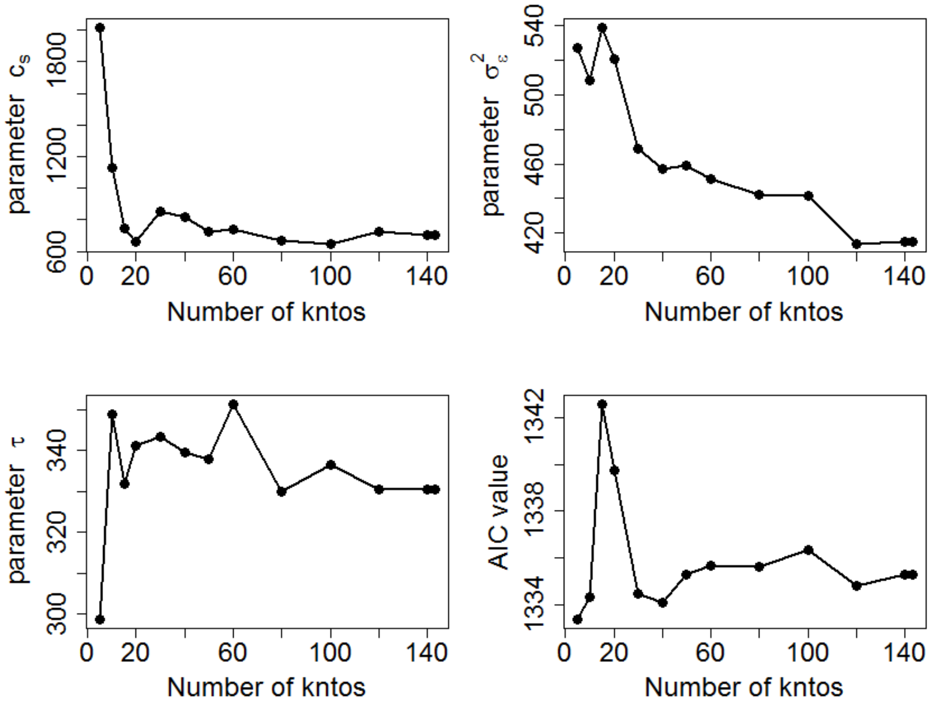

Figure 18: Parameter estimates and AIC value for several choices of the number of knots for the Paraná rainfall data using model (16). Parameters are estimated using GM2$_{ML}$ with a circular function g_τ.

Meuse Dataset: Lead

To analyse this data models (17) and (18) were considered which were estimated by the proposed methodology in Section 3 (GM2$_{ML}$). For the spatial component S we used the exponential, Matérn, circular, spherical and Gaussian as g_τ function. The circular function yielded the best fit in terms of AIC (see Table 8). We thus observe that the model accounting for the landuse by a random intercept has a slightly better fit.

In Figure 19 we present the estimated variance of the spatial component S of the best fitting model. In Figure 20 we present the estimated non-linear effect of the covariate distance to the river. The effects estimated by models (17) and (18) are very similar.

Table 8: Parameter estimates and AIC values estimated from geoadditive models (17) and (18) estimated by the proposed methodology in Section 3 (GM2$_{ML}$).

Model	$g_\tau(\cdot)$	$\hat{c_s}$	$\hat{\tau}$	$\hat{\sigma}^2_\varepsilon$	$\hat{\sigma}^2_\alpha$	AIC
(17)	exponential	0.153	0.224	0.060		178.31
(17)	Gaussian	0.115	0.217	0.092		178.57
(17)	spherical	0.132	0.805	0.093		175.66
(17)	circular	0.143	0.762	0.092		174.62
(17)	Matérn	0.119	0.127	0.093		177.67
(18)	exponential	0.144	0.279	0.065	0.024	179.11
(18)	Gaussian	0.115	0.217	0.092	0.000	180.57
(18)	spherical	0.125	0.816	0.085	0.028	175.44
(18)	circular	0.133	0.762	0.085	0.029	174.07
(18)	Matérn	0.114	0.179	0.098	0.028	178.15

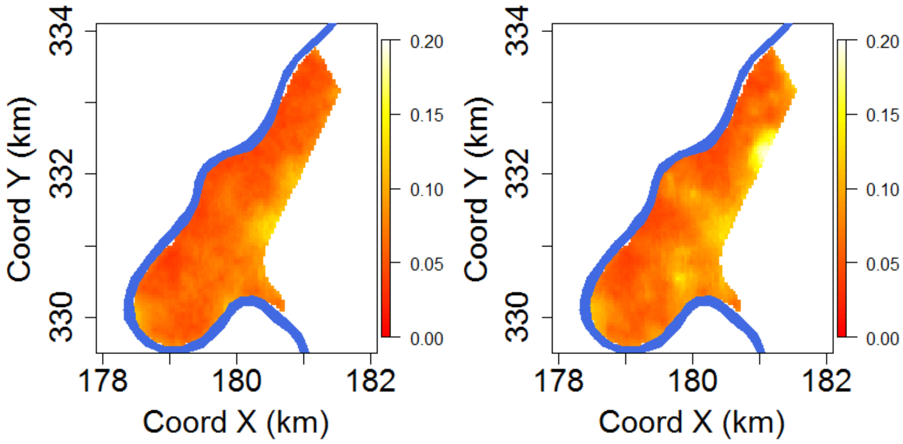

Figure 19: Estimated variance of the spatial component S obtained by model (17) (left) and model (18) (right) with a circular function g_τ which is estimated by the proposed methodology in Section 3 (GM2$_{ML}$).

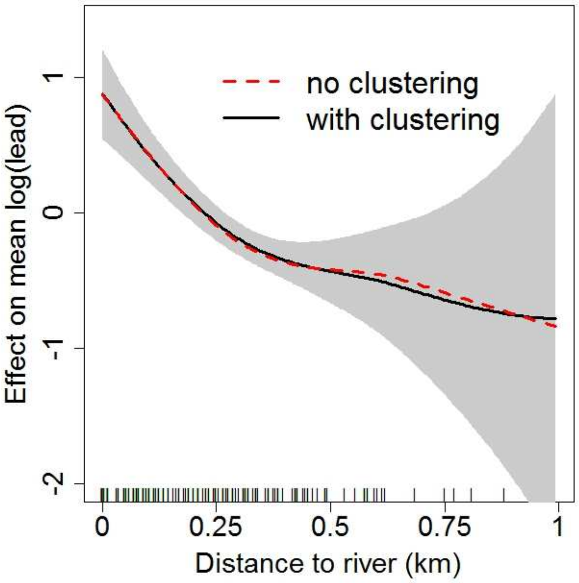

Figure 20: Estimated non-linear effect of the covariate distance to the river obtained by model (17) and (18) with a circular function g_τ which is estimated by the proposed methodology in Section 3 (GM2$_{ML}$). The shaded area represents the 95% point-wise confidence interval of the non-linear effect obtained by model (18).

Meuse Dataset: Zinc

The Meuse data is a classical geostatistical dataset used frequently to demonstrate various geostatistical analyses. The dataset comprises of four heavy metals measured in the top soil in a flood plain along the river Meuse. Polluted sediment is carried by the river, and mostly deposited close to the river bank. In total, 155 heavy metal concentration measurements (ppm) are available along with a number of soil and landscape variables (Figure 21 (a)). This dataset is available in the R-package `gstat`.

The goal is to create a prediction map of the zinc concentrations over the area of interest using the distance to the river as a covariate. The zinc concentrations are log-transformed to remove the skewness. Typical geostatistical analyses of the Meuse data consider a square root transformation of the distance to the river to obtain a linear effect with the log-zinc values. The results obtained from the geoadditive model estimated by the proposed methodology in the paper will therefore be compared with the results of kriging using direct maximum likelihood estimation using the square root distance to the river as covariate. The geoadditive model used is

$$\log(\text{zinc}_i) = \beta_0 + f(\text{dist}_i) + S(\mathbf{s}_i) + \varepsilon_i, \qquad 1 \leq i \leq 155, \tag{1}$$

where the distance to the river thus does not enter with a square root transformation in the model, but as a non-linear covariate effect. The Gaussian covariogram yielded the lowest AIC value for the direct maximum likelihood approach (see Table 9). Therefore, we also present the results with the Gaussian function g_τ in the geoadditive model for GM2_{ML}. We use 100 knots to model the geostatistical component S in model (1) (see Figure 22).

The covariogram parameters estimated by GM2_{ML} are comparable with those estimated by the direct likelihood approach, ($\hat{c}_s = 0.10, \hat{\tau} = 0.19$) and ($\hat{c}_s = 0.10, \hat{\tau} = 0.22$) respectively. Figure 21 (b)-(f) presents the prediction results and the effect of the covariate distance to the river on the mean log-zinc values. The GM2_{ML} approach yields 2.80 degrees of freedom for the non-linear effect of distance and 53.14 degrees of freedom for the spatial component S. The log-likelihood of the direct likelihood approach (-73.72) is somewhat greater than the log-likelihood of GM2_{ML} (-75.53). Nevertheless, this example nicely shows that geoadditive models estimated by the proposed methodology in Section 3 are useful to perform a geostatistical analysis using a non-linear covariate.

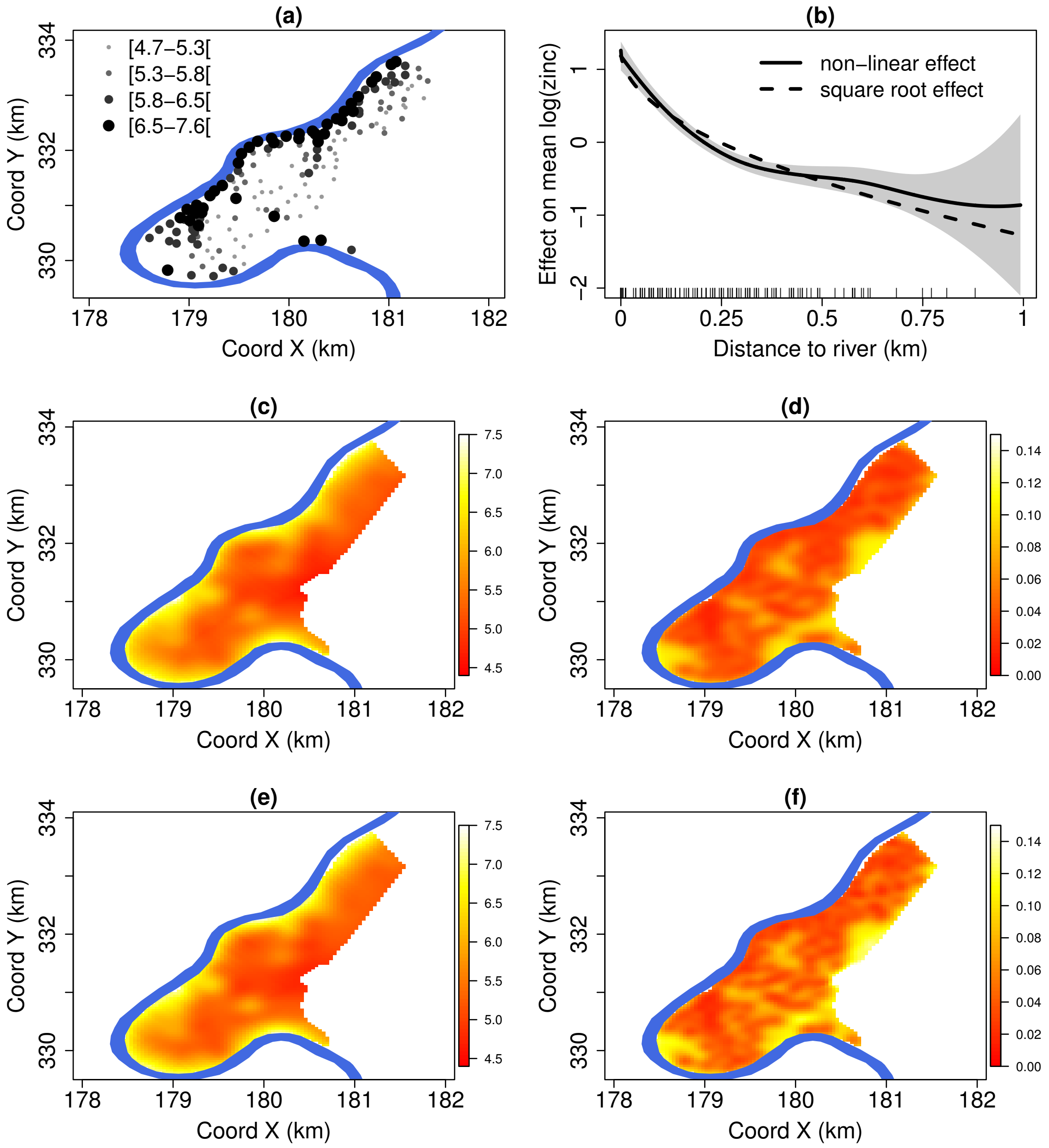

Figure 21: (a) Map showing the Meuse river and the measurement locations with corresponding recorded log-zinc values; (b) The estimated effect of distance to the river (the shaded area represents the 95% pointwise confidence interval of the non-linear effect); (c) Predicted surface using kriging with a Gaussian covariogram; (d) Prediction variance using kriging; (e) Predicted surface using the geoadditive model (1) with a Gaussian function g_τ estimated by the proposed methodology in Section 3; and (f) Prediction variance (based on the bootstrap approach) using the geoadditive model (1) with a Gaussian function g_τ estimated by the proposed methodology in Section 3.

Table 9: Parameter estimates and AIC values of the Meuse dataset using model (1) with different estimation approaches.

Method	$g_\tau(\cdot)$	$\hat{c}_s$	$\hat{\tau}$	$\hat{\sigma}_\varepsilon^2$	AIC
D_{ML}	exponential	0.14	0.17	0.05	159.84
D_{ML}	Gaussian	0.10	0.22	0.09	157.44
D_{ML}	spherical	0.12	0.42	0.06	158.21
D_{ML}	circular	0.12	0.35	0.06	158.12
D_{ML}	Matérn	0.11	0.10	0.08	158.44
$\mathrm{GM1}_{ML}$	exponential	1.18	4.44	0.09	173.04
$\mathrm{GM1}_{ML}$	Gaussian	0.00	4.44	0.18	198.96
$\mathrm{GM1}_{ML}$	spherical	0.00	4.44	0.18	198.96
$\mathrm{GM1}_{ML}$	circular	0.91	4.44	0.09	173.05
$\mathrm{GM1}_{ML}$	Matérn	0.78	4.44	0.09	172.79
$\mathrm{GM2}_{ML}$	exponential	0.13	0.20	0.08	161.71
$\mathrm{GM2}_{ML}$	Gaussian	0.10	0.19	0.08	162.96
$\mathrm{GM2}_{ML}$	spherical	0.12	0.43	0.08	161.44
$\mathrm{GM2}_{ML}$	circular	0.12	0.38	0.08	161.47
$\mathrm{GM2}_{ML}$	Matérn	0.11	0.09	0.08	161.76
GM_{ML}	thin plate			0.18	198.92

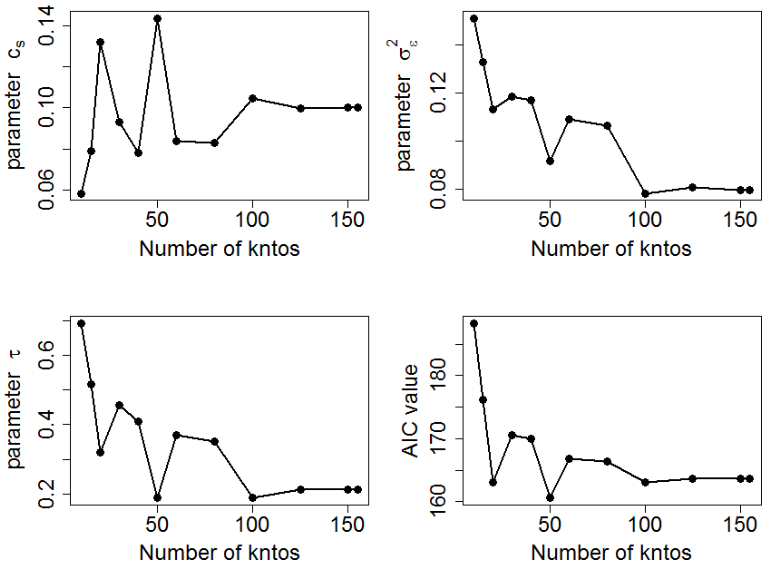

Figure 22: Parameter estimates and AIC value for several choices of the number of knots for the Meuse dataset using model (1) above. Parameters are estimated using $\mathrm{GM2}_{ML}$ with a Gaussian function g_τ.

4 R-module: Computational details

An R-module was made to fit geoadditive models using the proposed methodology in the paper. With this R-module all estimation methods used in the simulation study (Section 4) can be fit. The name of the R-module is `fit.spatial.process` and is available on request from the first author. This function has following inputs

```
fit.spatial.process = function(data, method, ini.parms=c(1,1,1),
    covariogram.model="exponential", number.of.knots=NULL,
    trend.input="cte", nonlinear.trend.input=NULL, cluster.input=NULL,
    tol1=1e-06, tol2=1e-06, max.iter=500, phi.upper.limit=1000)
```

The `data` is passed as a data frame with at least 3 columns. The first column contains the vector of responses. The second and third columns contain the 2-dimensional spatial coordinates. To fit a 1-dimensional coordinate system the third column should consist of zeros. It is optional to include other external covariates. These are given in columns 4, 5,. . . .

The `method` input takes a number between 1 and 7. These numbers correspond to the following options of fitting a spatial process:

- 1: Maximum likelihood (D_{ML}) estimation of the covariogram function. This option uses the `likfit` function from the `geoR` package. Default options are used for this function.

- 2: Restricted maximum likelihood (D_{REML}) estimation of the covariogram function. This option uses the `likfit` function from the `geoR` package. Default options are used for this function.

- 3: Weighted least squares (WLS) estimation of the variogram function. The functions `variogram` and `fit.variogram` from the `gstat` package are used. Weights of the empirical semivariogram are based on Cressie (1985). The corresponding weights are thus $|N(h(j))|/(\gamma(h(j);\theta)^2)$. The use of Cressie's weights corresponds with the option `fit.method=2` in the function `fit.variogram` in the R-package `gstat` that is used for the weighted least squares estimation.

- 4: Maximum likelihood estimation of the geoadditive model fixing τ at equation (6) ($\mathrm{GM1}_{ML}$). This method makes no use of the `ini.parms` statement. As mentioned in the paper, if the number of knots is less than the number of unique spatial locations, an efficient space filling algorithm (Johnson et al., 1990) is used to obtain a representative subset of knots. The linear mixed models obtained from the geoadditive model is fit using the function `lme()` with default options.

- 5: Restricted maximum likelihood of the geoadditive model fixing τ at equation (6) ($\mathrm{GM1}_{REML}$). This method makes no use of the `ini.parms` statement. As mentioned in the paper, if the number of knots is less than the number of unique spatial locations, an efficient space filling algorithm (Johnson et al., 1990) is used to obtain a representative subset of knots. The linear mixed models obtained from the geoadditive model is fit using the function `lme()` with default options.

- 6: Maximum likelihood of the geoadditive model using the estimation procedure proposed in Section 3 (GM2_{ML}). This method only uses the third value input of the ini.parms statement. As mentioned in the paper, if the number of knots is less than the number of unique spatial locations, an efficient space filling algorithm (Johnson et al., 1990) is used to obtain a representative subset of knots. In step (ii) of the proposed estimation procedure (see Section 3) the linear mixed models obtained from the geoadditive model is fit using the function lme() with default options. For the one parameter maximization of τ in step (iii) the function optimize() is used.

- 7: Restricted maximum likelihood of the geoadditive model using the estimation procedure proposed in Section 3 (GM2_{REML}). This method only uses the third value input of the ini.parms statement. As mentioned in the paper, if the number of knots is less than the number of unique spatial locations, an efficient space filling algorithm (Johnson et al., 1990) is used to obtain a representative subset of knots. In step (ii) of the proposed estimation procedure (see Section 3) the linear mixed models obtained from the geoadditive model is fit using the function lme() with default options. For the one parameter maximization of τ in step (iii) the function optimize() is used.

The function optimize() is a one-dimensional optimization function that uses a method which is a combination of golden section search and successive parabolic interpolation, and is designed for use with continuous functions. The lower endpoint of the interval to be searched is 0.001, and the upper endpoint of the interval to be searched is passed by the input phi.upper.limit.

If method equals 1 or 2, the return value is of the class likGRF. Method 3 returns a fitted variogram model (of the class variogramModel). Methods 4 to 7 returns an object of the class lme.

The ini.parms option takes a vector with 3 inputs specifying the initial parameter values, respectively the partial sill (c_s in kriging and σ_S^2 for splines), the nugget (c_0 in kriging and σ_ε^2 in splines) and the range (spatial-decay) parameter (τ).

The following covariogram models (covariogram.model) are supported by all methods: exponential, gaussian, spherical, circular, matern. The following options are only supported for methods 4 to 7: inverse.multiquadratic and thin.plate.

The input number.of.knots specifies the number of knots used for the spatial component in methods 4 to 7.

The trend.input specifies the mean trend of the spatial process. This input is always given as "...". By default the option "cte" is used which assumes a constant mean trend. The option "1st" assumes a first order polynomial on the coordinates and "2nd" assumes a second order polynomial on the coordinates. If external covariates are implemented this should be specified as "~formula". For example, a first order polynomial plus an age covariate effect is given as: "~x.coord + y.coord + age", where x.coord and y.coord are the data frame names of the spatial coordinate variables and age is the variable name of the external covariate. The nonlinear.trend.input argument is used when a covariate should be entered non-linearly into the model (only for methods 4 to 7). It should also be specified as "~formula". If one specifies a covariate in nonlinear.trend.input, it should also be specified in the `trend.input` input. The input `cluster.input` requires a vector of the same length as the `data` in which the cluster covariate is specified.

The tolerance statements `tol1` and `tol1` are only used in methods 6 and 7. The first tolerance `tol1` specifies the tolerance used in the optimization of a new τ parameter (thus passed to the `optimize()` function). The second tolerance `tol2` is used to check if two successive τ parameters are close enough to stop the doubly iterative procedure described in Section 3. The maximum number of iterations `max.iter` specifies the maximum number of times the τ parameter may be updated until the iterative procedure stops without reaching convergence.

In the R-module a function is available to perform predictions at (new) spatial locations:

```
predict.results = function(fit, data, covariogram.model, trend.input,
    nonlinear.trend.input=NULL, locations, cov.locations=NULL,
    cluster.input=NULL, cluster.locations=NULL)
```

In the R-module a function is available to calculation the prediction variance using the bootstrap approach described in Section 3:

```
bootstrap.variance = function(fit, data, covariogram.model, method,
    trend.input, nonlinear.trend.input=NULL, locations,
    cov.locations=NULL, cluster.input=NULL, cluster.locations=NULL,
    boot.samples=100, approximate=TRUE, seeding.bootstrap=12345){
```

For the simulation of $\mathbf{S}^*$ in step [iii] of the bootstrap approach (see Section 3) we use the function `RFsimulate()` of the R-package `RandomFields` (Schlather et al., 2014). This `RFsimulate()` function is also used in the simulation study to simulate the Gaussian Random Fields S.

To finalize, we present a small discussion on computational time. All computations of the data applications were performed on a laptop personal computer with Intel Core i5-4210M @ 2.60HGz processor. We only discuss computational time here of geoadditive models fit using the proposed methodology in Section 3. In the first application (Paraná State, Brazil, Rainfall data) model fitting for the best model took approximately 302 seconds and convergence was attained after 10 iteration steps. In the second example (Meuse Dataset: Lead) model fitting took 235 seconds (7 iteration steps) for the model without clustering, and 262 seconds (8 iteration steps) for the model with clustering. For the example discussed in this Supplementary Materials (Meuse Dataset: Zinc), the best fitting model took approximately 11 seconds to be estimated and convergence was attained after 8 iteration steps.

References

Altman, N. (2000). Krige, smooth, both or neither. *Australian and New Zealand Journal of Statistics 42*(4), 441–461.

Chaudhuri, P. and J. S. Marron (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association 94*, 807–823.

Cressie, N. A. C. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology 5*, 563–586.

Dubrule, O. (1984). Comparing splines and kriging. *Computers & Geosciences 10*, 327–338.

Hutchinson, M. F. and P. E. Gessler (1994). Splines – more than just a smooth interpolator. *Geoderma 62*, 45–67.

Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. *Journal of Statistical Planning and Inference 26*, 131–148.

Laslett, G. M. (1994). Kriging and splines: An empirical comparison of their predictive performance in some applications (with discussion). *Journal of the American Statistical Association 89*, 392–409.

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers & Geosciences 30*, 683–691.

R Core Team (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro Jr, P. J. and P. J. Diggle (2001). geoR: A package for geostatistical analysis. *R-NEWS 1*(2), 14–18. ISSN 1609-3631.

Ruppert, D., M. P. Wand, and R. J. Carroll (2003). *Semiparametric Regression.* Cambridge: University Press.

Schlather, M., A. Malinowski, M. Oesting, D. Boecker, K. Strokorb, S. Engelke, J. Martini, F. Ballani, P. J. Menck, S. Gross, U. Ober, K. Burmeister, J. Manitz, P. Ribeiro, R. Singleton, B. Pfaff, and R Core Team (2014). *RandomFields: Simulation and Analysis of Random Fields.* R package version 3.0.44.