Model-Based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Data Sets

Luc Steinbuch^{1,2,3,4} · Thomas G. Orton⁴ ·
Dick J. Brus⁵

Received: 30 November 2018 / Accepted: 4 November 2019 / Published online: 27 November 2019
© The Author(s) 2019

Abstract Area-to-point kriging (ATPK) is a geostatistical method for creating high-resolution raster maps using data of the variable of interest with a much lower resolution. The data set of areal means is often considerably smaller (< 50 observations) than data sets conventionally dealt with in geostatistical analyses. In contemporary ATPK methods, uncertainty in the variogram parameters is not accounted for in the prediction; this issue can be overcome by applying ATPK in a Bayesian framework. Commonly in Bayesian statistics, posterior distributions of model parameters and posterior predictive distributions are approximated by Markov chain Monte Carlo sampling from the posterior, which can be computationally expensive. Therefore, a partly analytical solution is implemented in this paper, in order to (i) explore the impact of the prior distribution on predictions and prediction variances, (ii) investigate whether certain aspects of uncertainty can be disregarded, simplifying the necessary computations, and (iii) test the impact of various model misspecifications. Several approaches using simulated data, aggregated real-world point data, and a case study on aggregated crop yields in Burkina Faso are compared. The prior distribution is

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11004-019-09840-6>) contains supplementary material, which is available to authorized users.

✉ Luc Steinbuch
luc.steinbuch@wur.nl

¹ Soil Geography and Landscape, Wageningen University, Wageningen, The Netherlands

² Team Soil, Water and Land Use, Wageningen Environmental Research, Wageningen, The Netherlands

³ ISRIC - World Soil Information, Wageningen, The Netherlands

⁴ School of Agriculture and Food Sciences, The University of Queensland, Brisbane, Australia

⁵ Biometris, Wageningen University and Research, Wageningen, The Netherlands

found to have minimal impact on the disaggregated predictions. In most cases with known short-range behaviour, an approach that disregards uncertainty in the variogram distance parameter gives a reasonable assessment of prediction uncertainty. However, some severe effects of model misspecification in terms of overly conservative or optimistic prediction uncertainties are found, highlighting the importance of model choice or integration into ATPK.

Keywords Area-to-point kriging · Spatial disaggregation · Bayesian statistics · Closed-form solutions

1 Introduction

An important challenge often encountered in scientific research is spatial prediction using areal-support data, that is, data about the variable of interest that is available as areal means only. Data may be aggregated for privacy protection, administrative, technical or other reasons. Examples include data on the failures in semiconductor chip production (White et al. 2017), cancer mortality (Goovaerts 2006), precipitation (Park 2013), soil properties (Kerry et al. 2012; Orton et al. 2012; Malone et al. 2009; Brus et al. 2014), soil hydraulic properties (Horta et al. 2014) and, the motivating example in this research, crop yields. You et al. (2009) state that, because of limited resources, crop yield information in many sub-Saharan countries is available only sparsely and in aggregated format; however, such information is needed at a finer spatial resolution to support efforts to increase crop productivity [see for example Orton et al. (2018)] and thereby improve both human welfare and ecological sustainability.

Model-based spatial prediction is usually done using algorithms based on point support (Diggle and Ribeiro 2007). In the case of areal-supported input, area-to-point kriging (ATPK) is a popular approach for creating fine-scale raster maps of the variable of interest and the corresponding prediction uncertainty (called spatial disaggregation or downscaling). In ATPK, regression coefficients (in the presence of covariates) and variogram parameters (describing spatial relationships) have to be estimated, for example by a least square estimator for the regression coefficients combined with an iterative variogram fitting deconvolution algorithm ('method of moments') on the regression residuals (Goovaerts 2008). More recent methods such as restricted maximum likelihood (REML) in combination with universal kriging¹ (UK) (both, and from hereon, referring to their application in the ATPK setting) consider the uncertainty in the regression coefficients (Webster and Oliver 2007). However, uncertainty in the variogram model parameters might also be a relevant source of uncertainty (Jansen 1998; Kitaniadis 1986; Minasny et al. 2011). Truong et al. (2014) showed that variogram uncertainty can have a substantial impact on ATPK variances. Brus et al. (2018) summarised earlier work by Pardo-Igúzquiza and Dowd (2001) showing that uncertainty in the variogram parameters can be quantified by the inverse Fisher matrix of the variogram parameters, but did not integrate this uncertainty in the kriging prediction uncertainty itself.

¹ Universal kriging in this work is defined as geostatistical prediction with the trend uncertainty included and where this trend is based on one or more covariates, which may or may not include the spatial coordinates.

In the Bayesian statistics paradigm, parameters can be considered stochastic rather than fixed but unknown (Gotway 2005). From a Bayesian perspective, REML in combination with UK considers the regression coefficients as stochastic and subsequently integrates them out—from the likelihood function for estimation of the variogram parameters as well as from the prediction. In this paper, this Bayesian direction is continued by successively integrating out the spatial variance parameter (analytically) and the spatial correlation distance parameter (numerically). By applying analytical solutions whenever possible, Markov chain Monte Carlo (MCMC) sampling from the Bayesian posterior distribution as proposed for example by Minasny et al. (2011) is avoided. MCMC can be computationally expensive and may, when used in a spatial context, be difficult to converge to a posterior distribution due to correlated parameters (Christensen et al. 2006). The extra effort of taking variogram parameter uncertainty into account could be most beneficial in the case of ATPK or area-to-area kriging, because the provided data set of areal means, from which regression coefficients and variogram parameters at high-resolution support must be inferred, can often be limited in size. However, taking variogram parameter uncertainty into account can also be beneficial in the case of point-to-point kriging (Le and Zidek 1992; Berger et al. 2001) and for sampling design (Marchant and Lark 2004, 2007).

It is not uncommon in ATPK studies to have only a small data set of areal means and no available relevant expert knowledge to inform model parameters (for example Brus et al. 2018). Therefore, this research aims to provide some insight into the applicability and behaviour of ATPK methods under these circumstances. To provide this insight, the following questions are answered: (i) What is the impact of different prior distributions—selected to represent a lack of prior knowledge about model parameters—on the quality of the ATPK predictions and prediction uncertainties? (ii) Can some aspects of uncertainty be disregarded, which might allow for computational benefits? (iii) How sensitive are results to misspecifications of the underlying statistical model?

In the following sections, the theoretical framework of model-based geostatistics for areal data, the Bayesian paradigm and the combination of these are briefly introduced. In the simulation part of this paper, REML is compared with more Bayesian approaches to perform and assess ATPK using a simulated spatial signal, including data sets as small as nine observations. Using real-world data, different approaches on self-aggregated remote sensing data are tested, referred to as the synthetic case study. Finally, as the motivating example, millet and sorghum yields, known as areal means only, are downscaled for each of the 45 provinces of Burkina Faso to a fine-scale grid of predicted yields.

2 Theory

2.1 Geostatistics Basics: Gaussian Random Field

According to the general theoretical framework of model-based geostatistics (Diggle and Ribeiro 2007), the spatial variable of interest is represented by a Gaussian random field

$$\mathbf{Z} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{C}(\phi)), \quad (1)$$

with MVN indicating a multivariate normal distribution; $\mathbf{X}$ the design matrix containing location-specific covariate values, including a column with ones to represent the regression intercept; $\boldsymbol{\beta}$ the vector of k regression coefficients (also called trend parameters); σ^2 in the terminology of geostatistics the partial sill variance; and $\mathbf{C}(\phi)$ the spatial correlation matrix as a parametric function of distance parameter ϕ .

Among several other possibilities, the exponential covariance function

$$(\sigma^2 \mathbf{C}(\phi))_{i1,i2} = \sigma^2 \exp\left(-\frac{h(s_{i1}, s_{i2})}{\phi}\right) \quad (2)$$

is assumed, where $i1$ and $i2$ index two discrete point locations s_{i1} and s_{i2} within the Gaussian random field, and $h(s_{i1}, s_{i2})$ represents the Euclidean distance between those locations, while $(\dots)_{i1,i2}$ means the element in row $i1$, column $i2$ of the indicated covariance matrix. Spatial directional dependence is not considered. For notational convenience, $\mathbf{C}(\phi)$ is indicated by $\mathbf{C}$ from now on.

2.2 Area-to-Point Kriging

In the case of ATPK, the observations are m areal means

$$\bar{z}_j = \frac{1}{|A_j|} \int_{s \in A_j} z(s) \, ds, \quad j \in 1, 2, \dots, m, \quad (3)$$

together $\bar{\mathbf{z}}$, with $z(s)$ an unobserved realisation of an infinite number of point values $\mathbf{Z}$ in area A_j and $|A_j|$ the surface area of area A_j , turning Eq. (1) into

$$\bar{\mathbf{Z}} \sim \text{MVN}(\bar{\mathbf{X}}\boldsymbol{\beta}, \sigma^2 \bar{\mathbf{C}}), \quad (4)$$

with $\bar{\mathbf{Z}}$ the stochastic representation of observations $\bar{\mathbf{z}}$, $\bar{\mathbf{X}}$ containing covariate values averaged over the corresponding areas, and $\sigma^2 \bar{\mathbf{C}}$ the matrix with average covariances between and within the areas. Note that $\boldsymbol{\beta}$, σ^2 and ϕ are equivalent in Eqs. (4) and (1): the parameters on point support are estimated from the areal data. Note the absence of a nugget effect (often indicated by τ^2), which might represent measurement errors and micro-scale variation, in the covariance model. Such a nugget effect cannot be identified based purely on areal-support data. Although Truong et al. (2014) demonstrated the potential of expert prior knowledge to help define a nugget parameter, in the situation investigated here, no such knowledge is available. This issue will return in the discussion.

To be able to predict values $\mathbf{z}^*$ at n^* point locations $\mathbf{s}^*$, together with the corresponding prediction variances $\mathbf{v}^*$, it is necessary to define (1) matrix $\bar{\mathbf{C}}^*$, the mean correlation between points in the observation areas and the prediction points and also implicitly a function of ϕ ; (2) matrix $\mathbf{C}^{**}$, the correlation matrix between the prediction points, again a function of ϕ ; (3) the design matrix for the prediction locations $\mathbf{X}^*$; and finally (4) the generalised least squares (GLS) estimator for $\boldsymbol{\beta}$ as given by

$$\hat{\beta} = (\bar{X}^T \bar{C}^{-1} \bar{X})^{-1} \bar{X}^T \bar{C}^{-1} \bar{z}. \quad (5)$$

The $n^* + m$ -dimensional distribution of the predictions and observations together can be written as

$$\begin{bmatrix} z^* \\ \bar{z} \end{bmatrix} | \beta, \sigma^2, \phi \sim \text{MVN}_{n^*+m} \left(\begin{bmatrix} X^* \beta \\ \bar{X} \beta \end{bmatrix}, \sigma^2 \begin{bmatrix} C^{**} & \bar{C}^* \\ (\bar{C}^*)^T & \bar{C} \end{bmatrix} \right). \quad (6)$$

According to UK theory, the resulting prediction is given by

$$\hat{z}_{\text{UK}}^* = \bar{C}^* \bar{C}^{-1} (\bar{z} - \bar{X} \hat{\beta}) + X^* \hat{\beta}, \quad (7)$$

an implicit function of ϕ , with prediction variance–covariance matrix

$$\begin{aligned} \text{var}[\hat{z}_{\text{UK}}^*] &= \sigma^2 C^{**} - \sigma^2 \bar{C}^* \bar{C}^{-1} (\bar{C}^*)^T + \sigma^2 (X^* - \bar{C}^* \bar{C}^{-1} \bar{X}) \\ &\quad (\bar{X}^T \bar{C}^{-1} \bar{X})^{-1} (X^* - \bar{C}^* \bar{C}^{-1} \bar{X})^T, \end{aligned} \quad (8)$$

an implicit function of ϕ and σ^2 . The diagonal of $\text{var}[\hat{z}_{\text{UK}}^*]$ is the vector of prediction error variances, better known as the universal kriging variance $\hat{v}_{\text{UK}}^*$, for every prediction point.

For a mathematical elaboration of the above equations, refer to Wackernagel (2014). The term regression kriging (RK) is used later to refer to simple kriging of the trend residuals, an approach which disregards any uncertainty in the estimated plug-in values of the trend parameters, which results in a different kriging variance. In the following sections, the framework presented above will be extended to a Bayesian one.

2.3 Bayesian Statistics

In the Bayesian framework, parameters are considered random variables (McElreath 2016). Formulated in general probability notation where Θ stands for ‘parameters of interest’ and E for evidence (observations), Bayes’ rule states that

$$p(\Theta|E) = \frac{p(\Theta) \times p(E|\Theta)}{p(E)}, \quad (9)$$

where the probability distribution $p(\Theta|E)$ indicates the posterior belief in the parameters given the evidence, $p(\Theta)$ indicates the prior belief in the parameters, independent of the evidence, $p(E|\Theta)$ is the probability of the evidence as a function of the parameters—called the likelihood—and $p(E)$ is the probability of the evidence. Note that to assess the likelihood, a correctly defined probability distribution of the modelled process is assumed. Note also that $p(E)$ is often left out when a proportional value for $p(\Theta|E)$ is sufficient. Mathematically, $p(E)$ equals $p(E, \Theta)$ integrated over its parameters

$$p(E) = \int p(\Theta, E) d\Theta. \quad (10)$$

In a Bayesian context, prediction entails formulating a distribution conditional on the evidence E . Inserting the parameters in the equation shows the derivation of the posterior predictive integral

$$\begin{aligned} p(P|E) &= \int p(P, \Theta|E) d\Theta \\ &= \int p(P|\Theta, E)p(\Theta|E) d\Theta, \end{aligned} \quad (11)$$

where P stands for the predicted values.

In the Bayesian framework, the prior $p(\Theta)$ needs to be defined, stating the current state of knowledge—or, inversely formulated, the current state of ignorance—about the parameters. In many cases, a low-informative prior is desired; however, formulating a prior that ‘lets the data speak for itself’ is not always straightforward (Seaman et al. 2012; Lindley 2004). Also, a ‘conjugate’ prior is often chosen so that its distribution function matches the likelihood, resulting in a closed-form description of the posterior (Albert 2009).

For various reasons, prior distributions that do not integrate to a finite value (i.e., cannot be normalised to integrate to one) might be considered, termed in the Bayesian framework as ‘improper’ priors. Improper priors can result in improper (poorly defined) posterior probability densities.

In the remainder of this paper, a posterior probability distribution—proper or not—is indicated by $f_p(\dots|\dots)$, the likelihood by $f_l(\dots|\dots)$, and a prior distribution—again having propriety or not—by $f_0(\dots)$. When the function type is ambiguous, its interpretation depends on the context and its arguments.

3 Implementation

3.1 Partly Analytical Bayesian Area-to-Point Algorithm

In this section, a partly analytical and partly numerical algorithm to execute Bayesian ATPK is described, based on integrating out the trend and variance parameters and systematically exploring gridded values in the correlation distance parameter space. This algorithm is developed as an alternative to methods based on sampling from the posterior distribution. Starting from Eq. (4), the likelihood of data generated by the geostatistical model is based on the multivariate normal distribution

$$f_l(\bar{z}|\beta, \sigma^2, \phi) = \frac{1}{(2\pi)^{\frac{m}{2}} (\sigma^2)^{\frac{m}{2}} |\bar{C}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} (\bar{z} - \bar{X}\beta)^T \bar{C}^{-1} (\bar{z} - \bar{X}\beta) \right\}, \quad (12)$$

where $|\dots|$ indicates the determinant.

Throughout this work, the priors for the trend, variance and correlation distance parameters are given by

$$f_0(\boldsymbol{\beta}, \sigma^2, \phi) \propto \frac{1}{\sigma^2} f_0(\phi). \quad (13)$$

This prior represents a priori independence between the parameters with an unlimited uniform (and thus improper) prior for the regression coefficient vector; a prior for the variance that is equivalent to an unlimited uniform prior for $\ln(\sigma^2)$, again an improper prior; and $f_0(\phi)$, for which different options are considered. It falls under the more general formulation of Berger et al. (2001), who considered appropriate objective (uninformative) priors for the analysis of spatial point-support data.

Given the above prior and likelihood function, the joint posterior distribution for the parameters is (up to a constant of proportionality)

$$f_p(\boldsymbol{\beta}, \sigma^2, \phi | \bar{\mathbf{z}}) \propto \frac{1}{(2\pi)^{\frac{m}{2}} (\sigma^2)^{\frac{m}{2}} |\bar{\mathbf{C}}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} (\bar{\mathbf{z}} - \bar{\mathbf{X}}\boldsymbol{\beta})^T \bar{\mathbf{C}}^{-1} (\bar{\mathbf{z}} - \bar{\mathbf{X}}\boldsymbol{\beta}) \right\} \frac{1}{\sigma^2} f_0(\phi). \quad (14)$$

Based on the above assumptions, Fig. 1 illustrates our partly analytical Bayesian algorithm, Bayesian areal kriging (BAK), to infer the marginal posterior distributions of all parameters and to calculate and summarise predictive distributions. The relevant equations and their derivation are given in Online Resource A; the summary stating the main equations follows in the coming sections.

3.1.1 Marginal Posterior Distance Parameter

Given the joint posterior (Eq. 14), $\boldsymbol{\beta}$ and σ^2 are analytically integrated out to arrive at the analytical solution for the marginal posterior for ϕ given by

$$f_p(\phi | \bar{\mathbf{z}}) \propto f_0(\phi) \frac{1}{|\bar{\mathbf{C}}|^{\frac{1}{2}} |\bar{\mathbf{X}}^T \bar{\mathbf{C}}^{-1} \bar{\mathbf{X}}|^{\frac{1}{2}} \left[(\bar{\mathbf{z}} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}})^T \bar{\mathbf{C}}^{-1} (\bar{\mathbf{z}} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}) \right]^{\frac{m-k}{2}}}, \quad (15)$$

where $\hat{\boldsymbol{\beta}}$ is defined according to Eq. (5).

Numerically, BAK creates a one-dimensional grid covering the parameter space of ϕ , calculates the marginal posterior for each ϕ , and normalises the marginal posterior to a distribution that integrates to one within the bounds of the ϕ grid.

3.1.2 Marginal Posterior Sill

For the marginal posterior of σ^2 , $\boldsymbol{\beta}$ is analytically integrated out from the joint posterior (Eq. 14) to arrive at

Fig. 1 Bayesian areal kriging (BAK) algorithm workflow. The posterior density can be represented by one 1-dimensional and several 2-dimensional parameter grids, each of which can be numerically integrated and normalised. The shown densities are meant for illustration

$$f_p(\sigma^2 | \bar{z}) \propto \int f_0(\phi) \frac{\exp \left\{ -\frac{1}{2\sigma^2} \left[(\bar{z} - \bar{X}\hat{\beta})^T \bar{C}^{-1} (\bar{z} - \bar{X}\hat{\beta}) \right] \right\}}{\left| \bar{C} \right|^{\frac{1}{2}} \left| \bar{X}^T \bar{C}^{-1} \bar{X} \right|^{\frac{1}{2}} (\sigma^2)^{\frac{m-k+2}{2}}} d\phi. \quad (16)$$

As, to the authors' knowledge, there is no analytical way of integrating out ϕ , BAK creates a two-dimensional grid over the parameter space of ϕ and σ^2 and calculates the joint posterior for σ^2 and ϕ (i.e., the integrand) for every grid point; then it performs a trapezoidal integration over ϕ and normalises to arrive at the marginal distribution for σ^2 .

3.1.3 Marginal Posterior Regression Coefficients

The marginal posteriors of the individual regression coefficients β_q , $q = 1 \dots k$ are based on the joint posterior for the vector $\boldsymbol{\beta}$

$$f_p(\boldsymbol{\beta}|\bar{\mathbf{z}}) \propto \int f_0(\phi) \frac{1}{|\bar{\mathbf{C}}|^{\frac{1}{2}} \left[(\bar{\mathbf{z}} - \bar{\mathbf{X}}\boldsymbol{\beta})^T \bar{\mathbf{C}}^{-1} (\bar{\mathbf{z}} - \bar{\mathbf{X}}\boldsymbol{\beta}) \right]^{\frac{m}{2}}} d\phi. \quad (17)$$

The integrand here can be shown to be proportional to a multivariate t distribution (Roth 2013) for $\boldsymbol{\beta}$ with $m - k$ degrees of freedom, location (vector) parameter $\hat{\boldsymbol{\beta}}$ and scale (matrix) parameter

$$\boldsymbol{\Sigma}_\beta = \frac{(\bar{\mathbf{z}} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}})^T \bar{\mathbf{C}}^{-1} (\bar{\mathbf{z}} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}})}{m - k} (\bar{\mathbf{X}}^T \bar{\mathbf{C}}^{-1} \bar{\mathbf{X}})^{-1}. \quad (18)$$

This integrand can be marginalised to a scaled t -distribution for the individual regression coefficients, as an implicit function of ϕ , and rearranged to give

$$f_p(\beta_q|\bar{\mathbf{z}}) \propto \int f_p(\phi|\bar{\mathbf{z}}) t_v(\beta_q; \hat{\beta}_q, \Sigma_q) d\phi \quad (19)$$

with $f_p(\phi|\bar{\mathbf{z}})$ as indicated in Eq. (15) and where

$$\begin{aligned} & t_v(\beta_q; \hat{\beta}_q, \Sigma_q) \\ &= \frac{\Gamma[(v+1)/2]}{\Gamma(v/2)(v\pi)^{1/2} |\Sigma_q|^{1/2}} \left[1 + \frac{1}{v} (\beta_q - \hat{\beta}_q)^T \Sigma_q^{-1} (\beta_q - \hat{\beta}_q) \right]^{-(v+1)/2} \end{aligned} \quad (20)$$

defines a t -distribution for β_q with degrees of freedom $v = m - k$, location parameter $\hat{\beta}_q$ and scale parameter Σ_q the q th element on the diagonal of $\boldsymbol{\Sigma}_\beta$. Note that the variance of this t -distribution is $\Sigma_q v / (v - 2)$.

Similarly to Sect. 3.1.2, BAK creates two-dimensional grids covering the parameter spaces of ϕ and β_q (for all q) and applies the trapezoidal rule to calculate the integral over ϕ in Eq. (19); finally, it normalises to get the marginal distributions for each individual β_q .

3.1.4 Posterior Predictive Distribution

The conditional distribution for the variable of interest, given the data and any particular value of the distance parameter, $f(z^*|\bar{\mathbf{z}}, \phi)$, is a t -distribution with degrees of freedom $v = m - k$, with location parameter $\hat{z}^*|\phi$ —an implicit function of ϕ —already given in Eq. (7), and with scale parameter as provided in Eq. A85 in Online Resource A. The variance of this conditional distribution, also a function of ϕ , is given by

$$\begin{aligned} \text{var}[\hat{z}^*|\phi] &= \frac{m-k}{m-k-2} \frac{(\bar{z} - \bar{X}\hat{\beta})^T \bar{C}^{-1} (\bar{z} - \bar{X}\hat{\beta})}{m-k} \\ &\times \left\{ C^{**} - \bar{C}^* \bar{C}^{-1} (\bar{C}^*)^T + (X^* - \bar{C}^* \bar{C}^{-1} \bar{X})(\bar{X}^T \bar{C}^{-1} \bar{X})^{-1} (X^* - \bar{C}^* \bar{C}^{-1} \bar{X})^T \right\}. \end{aligned} \quad (21)$$

Note that Eq. (21) is an increased universal kriging variance [see for comparison Eq. (8)] because the uncertainty in σ^2 is also considered—hence the increment expressed in the first fraction. The second fraction equals the REML estimate for σ^2 given ϕ .

The posterior predictive distribution is defined as an integral of the above conditional distribution with respect to the posterior distribution of the distance parameter,

$$f_p(z^*|\bar{z}) = \int_{\phi} f(z^*|\bar{z}, \phi) f_p(\phi|\bar{z}) d\phi, \quad (22)$$

which is numerically approximated. BAK first creates for each prediction point s^* a vector of predictions and a vector of corresponding prediction variances, both as a function of ϕ . Finally, the algorithm calculates the mean and variance of the posterior predictive distribution (or, more formally, of a finite mixture distribution that approximates this distribution, with weights defined based on $f_p(\phi|\bar{z})$ and the spacings of the ϕ parameter grid).

3.2 Methodological Details

In this work, a number of increasingly Bayesian approaches to ATPK are applied and compared (Table 1). The first three rows of the table represent plug-in approaches for some of the parameters (i.e., the stated parameters are first estimated, by maximising a likelihood or marginal likelihood function, before being plugged into the relevant predictive distribution equation for prediction), while the final row represents the fully Bayesian approach. In the case of maximum likelihood estimation (ML, and not implemented in this work), all parameters (in the geostatistical context: regression coefficients and spatial covariance parameters) are estimated by analytically or numerically maximising the likelihood. This general approach was consolidated by Fisher almost a century ago (Stigler 2007) and applied in geostatistics for example by Kitanidis (1983) and Lark (2000). REML, which has been advocated for several decades in geostatistics, is based on a likelihood function for a set of projected data rather than the original data, and gives conditionally unbiased estimates for the spatial parameters (Webster and Oliver 2007; Lark and Cullis 2004); see also Sect. A2 in Online Resource A. REML represents a form of marginal likelihood (a likelihood function in which some parameters have been marginalised), and has been presented in a Bayesian framework as such (the integral of the likelihood function with respect to the trend parameters, assuming a flat improper prior for these parameters) (Harville 1974). Note that $f_0(\beta)$ can be considered an uninformative prior when neglected—this is often valid for centrality parameters but not for other parameters. Underpinning the same approach, UK takes the uncertainty in the trend coefficients into account, making it a logical combination with REML. Within this research, the combined application

Table 1 Geostatistical approaches from maximum likelihood to full Bayesian. Corresponding to their universal kriging counterparts, $\hat{z}_{\text{RK}}^*$ and $\hat{v}_{\text{RK}}^*$ indicate the regression kriging and regression kriging variance, respectively

	Basis for estimation, function to be maximised/integrated posterior distribution	Estimated plug-ins	Basis for prediction	Predictive distribution	Prediction	Prediction variance
ML—maximum likelihood	$f_I := f_I(z \beta, \sigma^2, \phi)$	$\hat{\beta}, \hat{\sigma}^2, \hat{\phi}$	$f(z^* z, \hat{\beta}, \hat{\sigma}^2, \hat{\phi})$	N	$\hat{z}_{\text{RK}}^*$	$\hat{v}_{\text{RK}}^*$
REML—restricted maximum likelihood	$\int f_I f_O(\beta) d\beta$	$\hat{\sigma}^2, \hat{\phi}$	$\int f(z^* z, \beta, \hat{\sigma}^2, \hat{\phi}) \times f(\beta z, \hat{\sigma}^2, \hat{\phi}) d\beta$	N	$\hat{z}_{\text{UK}}^*$	$\hat{v}_{\text{UK}}^*$
MML—maximum marginal likelihood	$\int f_I f_O(\beta, \sigma^2) \times d(\beta, \sigma^2)$	$\hat{\phi}$	$\int f(z^* z, \beta, \sigma^2, \hat{\phi}) \times f(\beta, \sigma^2 z, \hat{\phi}) d(\beta, \sigma^2)$	t	$\hat{z}_{\text{UK}}^*$	$\hat{v}_{\text{UK}}^* \times \frac{m-k}{m-k-2}$
Full Bayesian	$\int f_I f_O(\beta, \sigma^2 \phi) \times d(\beta, \sigma^2, \phi)$	—	$\int f(z^* z, \beta, \sigma^2, \phi) \times f_P(\beta, \sigma^2, \phi z) d(\beta, \sigma^2, \phi)$	a	b	c

^aMixture of t
^bWeighted average of $\hat{z}_{\text{UK}}^*$, see Online Resource A
^cFunction of $\hat{z}_{\text{UK}}^*$, $\hat{v}_{\text{UK}}^* \times \frac{m-k}{m-k-2}$ and $f_P(\phi|\bar{z})$, see Online Resource A

of REML and UK is indicated by ‘REML approach’. The next gradation towards the fully Bayesian approach is maximum likelihood with both trend and variance integrated out, in the context of this paper indicated by the generic term ‘maximum marginal likelihood’ (MML). Finally, the full Bayesian approach (also referred to as ‘Bayesian approach’) provides a posterior distribution of all parameters, while in the prediction all parameters are integrated out and the uncertainty of all parameters is taken into account.

In the following sections, REML, MML and the Bayesian approach are compared, and for the Bayesian approach different priors for ϕ (as defined in the following section) are applied. All algorithms (including the central BAK algorithm as presented in Fig. 1) are written in the statistical programming language R, and are available at Steinbuch et al. (2019).

3.2.1 Prior Distributions for ϕ

For $f_0(\phi)$, three potential forms of prior distribution are compared, intended to represent limited prior knowledge. These are (1) a uniform prior with limited bounds; (2) the reference prior as suggested by Berger et al. (2001) for analysis of point-support data, applied in the context of areal-support data, and explained in Online Resource B; and—in the simulation ensemble—(3) an inverse-gamma distribution. The bounded uniform and the inverse-gamma distributions are proper; the assumed propriety of the reference prior will be discussed later.

3.2.2 Estimation and Prediction with REML and MML

For REML, the approach as described by Brus et al. (2018) was applied. For MML, the posterior mode of ϕ was calculated using the Bayesian approach with a uniform prior for ϕ . The predictive distribution was then defined conditionally on this value of ϕ . Mathematically, this equals integrating out β and σ^2 to arrive at an estimated $\hat{\phi}$, which is successively used as a single plug-in value for MML prediction; the mean and variance of the predictive distribution (representing the prediction and prediction variance) are shown in Table 1 and Eqs. (7) and (21).

3.2.3 Estimating Average Covariances

The average correlation matrices, $\bar{C}$ and $\bar{C}^*$, can be approximated in different ways. In this research, many discretisation points within each area are defined and the relevant Euclidean distances between those points are calculated, followed by construction of the corresponding correlation matrix, based on the correlation function—such as given in Eq. (2)—and distance parameter ϕ . Then, all correlations per area-area combination are averaged to arrive at $\bar{C}$, and per area-prediction point combination to arrive at $\bar{C}^*$. The discretisation points were on a regular grid in the simulation study, and selected by simple random sampling in the two case studies.

3.2.4 Validation

To quantify the performance of each approach—for the simulation study and synthetic case study, where the original point data were available—the predictions z^* and prediction uncertainties v^* were assessed in relation to the original signal z . As an indication of the quality of the prediction, the root mean squared error (RMSE) defined as

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \{z(s_i) - z^*(s_i)\}^2} \quad (23)$$

was calculated, where a smaller number indicates more accurate predictions (Oliver and Webster 2014). For comparison, a baseline approach was also included, for which point predictions were defined simply by the areal mean data for the corresponding area. Unbiasedness of predictions was tested using the mean error (ME), $\frac{1}{n} \sum_{i=1}^n \{z(s_i) - z^*(s_i)\}$. The mass preserving property (MPP) of the predictions was checked, which states that, in the case of ATPK, the mean of all predictions in any observed area should equal the observed areal mean (Kyriakidis 2004). This check was summarised by showing the maximum observed difference between areal-average data and the mean of the corresponding predictions.

As an indication of the quality of the prediction uncertainty, a motivating factor for this work, the standardised squared error (StSE) defined as

$$\text{StSE}(s_i) = \frac{\{z(s_i) - z^*(s_i)\}^2}{v^*(s_i)} \quad (24)$$

was calculated. This StSE should ideally have a mean of one (Lark 2000). Higher values indicate an underestimation of uncertainty, which is labelled ‘optimistic’, and lower values indicate an overestimation of uncertainty, labelled ‘conservative’.

4 Simulation Study

The following shows a single one-dimensional simulation where REML and full Bayesian (defined with $f_0(\phi) \sim \text{uniform}$) are compared for illustration purposes. Following the illustration, an ensemble of many simulations is applied to assess several settings. Online Resource D contains similar results for two-dimensional simulations.

4.1 Single Simulation

4.1.1 Simulated Data Set

A line of length 300 abstract units (au) was created, and filled with $n = 600$ equally spaced nodes. Using the exponential covariance function, a spatially correlated signal (with a zero-nugget exponential model; $\sigma_{\text{sim}}^2 = 5$, $\phi_{\text{sim}} = 60$) was generated and added to the trend of a linear function of the coordinate ($\beta_{1\text{sim}} = 0$ for the intercept, $\beta_{2\text{sim}} =$

Fig. 2 One simulation and resulting predictions, selected for illustration purposes. The predictions of REML and Bayesian with uniform prior for ϕ coincide largely, but Bayesian shows a larger prediction uncertainty. Distance is measured in abstract units (au)

0.02 for the slope on coordinate). For the Bayesian approach, $f_0(\phi) \sim \text{uniform}$ was used, bounded by $\phi_{l,u} = \{10, 300\}$; these bounds also defined bounds for the REML parameter search. The priors for β and σ^2 are provided in Eq. (13). The above settings are referred to as the standard settings. The line was split into $m = 10$ equal one-dimensional ‘areas’ or line sections. Finally, both z and the covariate over the areas were averaged to arrive at the observed means $\bar{z}$ and the averaged design matrix $\bar{X}$.

4.1.2 Results

The original signal (assumed to be unobservable, and represented as point values), the areal means (the ‘observations’) and the predictions are shown in Fig. 2. The difference between the REML and the full Bayesian approach is mainly in the prediction uncertainty: the Bayesian approach gives a slightly larger prediction interval.

The marginal densities in Fig. 3 show that, based on this—small—simulated data set, it is rather difficult to identify the distance parameter ϕ , which has a very flat mode. The REML estimates for ϕ and σ^2 (point values) are close to the modes of the respective marginal posteriors from the Bayesian approach. For the trend parameters, the marginal Bayesian posterior distributions are slightly skewed and wider than the corresponding distributions based on REML (Gaussian distributions parameterised by the GLS estimate and estimation variance), indicating that more uncertainty is included.

The pairwise joint posteriors (ϕ with each other parameter) are shown in Fig. 4: σ^2 and ϕ seem to have a positive correlation (subfigure a), while β_1 has a slightly negative correlation with ϕ (b) and β_2 a slightly positive correlation (c). Figure 5 shows the variogram models obtained with REML and the Bayesian-uniform approach. Table

Fig. 3 Marginal posterior distributions of spatial parameters σ^2 (a) and ϕ (b), and of the two trend parameters $\beta_{1,2}$ (c, d), resulting from the single run with Bayesian-uniform approach. REML parameter estimations and the data simulation settings (i.e., the true values) are also shown

Fig. 4 The joint posterior distributions of the spatial parameters σ^2 and ϕ (a), and of the two β elements with ϕ (b, c). REML parameter estimation and the data simulation settings are also shown. The grey shade is an indication of the posterior probability density (the darker, the larger)

Fig. 5 Variogram model as estimated by REML and Bayesian-uniform models (for several value combinations of σ^2 and ϕ) shaded according to their probability densities. The empirical residual variogram is added for reference

2 shows the validation results of this single simulation. The quality of the prediction (RMSE) is almost equal for the REML and Bayesian approaches, and better than for the baseline approach. The ME and max(MPP) are close to zero, indicating the absence of bias, and a discretisation grid (which is different from the prediction grid) of sufficient density to provide good approximations of areal-average covariances and

Table 2 Results of single simulation run with validation on original data points

	mean(StSE)	RMSE	ME	max(MPP)
REML	1.169	0.585	0.000	0.009
Bayesian-uniform	0.827	0.588	0.000	0.009
Baseline	–	0.666	0.000	0.000

REML restricted maximum likelihood, *StSE* standardised squared error, *ME* mean error, *RMSE* root mean squared error, *max(MPP)* maximal found deviation from the mass preserving property

areal-average values of covariates. The uncertainty validation value mean(StSE) shows that REML is on average optimistic, while the Bayesian approach is conservative. Note that the baseline approach does not provide a quantification of prediction uncertainty.

4.2 Simulation Ensemble

In this section, results are generalised by generating many simulations, varying only the random number seed, while comparing validation statistics on the outcomes of several approaches: REML, MML and full Bayesian with the three different priors for ϕ indicated earlier. The applied inverse-gamma prior for ϕ is set as somewhat informative with shape = 11 and rate = 600. This results in a mean of 60 au, emulating a situation where decent prior knowledge about the range is available. Also, the number of observations m is varied by dividing the line into m sections of equal length; this together is one ‘session’. Furthermore, to investigate how the approaches behave for differently simulated data sets and different inference settings, both are varied into an ‘ensemble’ of many sessions. The settings as used in standard session 1 are given in Sect. 4.1.1. Sessions 2 and 3 vary the upper bound for the uniform prior and for the REML and MML searches for estimation of ϕ in comparison with standard session 1 (where $\phi = 300$ au). In session 4, the trend is removed (so that the inferential model has to infer the mean only); in session 5, the trend is based on a separate Gaussian random field (GRF) rather than on the coordinates. Sessions 6, 7 and 8 introduce a misfit between the simulation model and the inferential model, where the correlation function in the simulation model is changed or a nugget component is added—the inferential model stays unchanged. Finally, sessions 9 and 10 show the effects of a misfit between the actual signal and the support of the available data (i.e., short or very long distance parameter used in simulation compared with the area sizes and total extent), which might make it difficult to identify parameters.

Table 3 presents the results expressed as the average (and, in small font, the corresponding standard deviation) over 250 means of the standardised squared error (mean(StSE)). Table D1 in Online Resource D shows the results of the analogous two-dimensional simulations. Additional validation statistics and assessments regarding ϕ and σ^2 for the simulation ensemble can be found in Online Resource C, which also includes $m = 15$ and $m = 30$ for the one-dimensional simulations, and in Online Resource E for the corresponding figures of the two-dimensional simulations.

Table 3 Mean standardised squared error (mean(StSE)) for the one-dimensional simulation ensemble, comparing restricted maximum likelihood (REML), maximum marginal likelihood (MML) and three full Bayesian approaches with different priors for ϕ

Session	m	REML	MML	InvGam	Uniform	Reference
1: Standard settings	10	1.25 ^{1.75}	0.94 ^{1.32}	0.99 ^{1.06}	1.07 ^{1.23}	0.40 ^{0.52}
	20	1.05 ^{0.47}	0.93 ^{0.41}	0.98 ^{0.39}	1.02 ^{0.41}	0.72 ^{0.35}
	50	1.01 ^{0.28}	0.97 ^{0.27}	0.99 ^{0.26}	1.01 ^{0.27}	0.93 ^{0.25}
2: Small upper limit: $\phi_u = 100$	10	1.19 ^{1.57}	0.89 ^{1.18}		0.87 ^{0.96}	
	20	1.03 ^{0.45}	0.91 ^{0.40}		0.92 ^{0.37}	
	50	1.01 ^{0.28}	0.97 ^{0.26}		0.98 ^{0.26}	
3: Large upper limit: $\phi_u = 2000$	10	1.26 ^{1.79}	0.95 ^{1.34}		1.20 ^{1.39}	
	20	1.05 ^{0.47}	0.93 ^{0.42}		1.07 ^{0.43}	
	50	1.02 ^{0.28}	0.97 ^{0.27}		1.03 ^{0.27}	
4: Trend is Gaussian Random field	10	1.29 ^{1.37}	0.97 ^{1.03}	0.96 ^{0.76}	1.07 ^{0.95}	0.51 ^{0.66}
	20	1.07 ^{0.46}	0.95 ^{0.41}	0.97 ^{0.37}	1.02 ^{0.40}	0.79 ^{0.38}
	50	1.03 ^{0.28}	0.99 ^{0.27}	1.00 ^{0.26}	1.02 ^{0.27}	0.96 ^{0.26}
5: No trend	10	1.26 ^{1.56}	0.98 ^{1.21}	0.97 ^{0.85}	1.09 ^{1.11}	0.54 ^{0.86}
	20	1.07 ^{0.47}	0.96 ^{0.42}	0.97 ^{0.37}	1.03 ^{0.41}	0.82 ^{0.39}
	50	1.03 ^{0.28}	0.98 ^{0.27}	0.99 ^{0.26}	1.02 ^{0.27}	0.96 ^{0.26}
6: Matérn simulation, smooth; $v_{\text{sim}} = 2$	10	0.14 ^{0.14}	0.10 ^{0.11}	0.10 ^{0.11}	0.10 ^{0.11}	0.05 ^{0.05}
	20	0.04 ^{0.03}	0.04 ^{0.03}	0.03 ^{0.02}	0.04 ^{0.03}	0.03 ^{0.02}
	50	0.01 ^{0.01}	0.01 ^{0.01}	0.01 ^{0.01}	0.01 ^{0.01}	0.01 ^{0.01}
7: Matérn simulation, unsmooth; $v_{\text{sim}} = .25$	10	2.78 ^{2.49}	2.09 ^{1.87}	2.42 ^{1.76}	2.56 ^{1.91}	0.89 ^{0.78}
	20	2.57 ^{1.40}	2.29 ^{1.24}	2.67 ^{1.14}	2.70 ^{1.21}	1.68 ^{1.09}
	50	2.51 ^{0.73}	2.41 ^{0.70}	2.73 ^{0.69}	2.68 ^{0.70}	2.25 ^{0.70}
8: Added nugget: $\tau^2 = 1$	10	3.45 ^{3.54}	2.59 ^{2.66}	2.78 ^{2.21}	2.99 ^{2.49}	1.09 ^{1.05}
	20	4.70 ^{2.18}	4.17 ^{1.94}	4.43 ^{1.76}	4.59 ^{1.87}	3.21 ^{1.68}
	50	7.19 ^{1.93}	6.89 ^{1.85}	7.34 ^{1.75}	7.36 ^{1.80}	6.57 ^{1.75}
9: Extreme short range in simulation: $\phi_{\text{sim}} = 5$	10	4.90 ^{6.04}	3.68 ^{4.53}	5.72 ^{4.24}	5.71 ^{4.55}	1.62 ^{1.62}
	20	2.39 ^{1.33}	2.13 ^{1.19}	3.24 ^{1.63}	3.02 ^{1.60}	1.01 ^{0.94}
	50	1.51 ^{0.40}	1.45 ^{0.38}	1.72 ^{0.48}	1.61 ^{0.45}	0.92 ^{0.42}
10: Extreme long range in simulation: $\phi_{\text{sim}} = 600$	10	1.01 ^{0.79}	0.76 ^{0.59}	0.77 ^{0.56}	0.85 ^{0.61}	0.32 ^{0.25}
	20	0.99 ^{0.42}	0.88 ^{0.37}	0.87 ^{0.35}	0.93 ^{0.37}	0.69 ^{0.31}
	50	1.00 ^{0.27}	0.96 ^{0.26}	0.95 ^{0.25}	0.98 ^{0.25}	0.91 ^{0.24}

In the online version of this paper, colours are used to highlight larger and smaller values for the readers' convenience

4.2.1 General Results

Referring to Online Resources C and E, the maximal difference with respect to the mass preserving property ($\max(\text{MPP})$) ranges between 0.09 and 0.28 in the case of the two-dimensional simulations. In the one-dimensional case, $\max(\text{MPP})$ is much smaller. With all approaches in all simulations, the ME was small. The RMSE was, for a given simulation, almost equal for all, but the baseline approach was, on average, larger. The main difference between the approaches was in the prediction uncertainty (assessed by StSE).

The standard session 1 in Table 3 shows that $m = 10$ caused REML to be optimistic, while the Bayesian-uniform approach was less optimistic, Bayesian-inverse-gamma was closest to one (perhaps due to the knowledge captured in the prior distribution for ϕ), MML was slightly conservative and Bayesian-reference very conservative.

With increasing m , all $\text{mean}(\text{StSE})$ approached one, while the corresponding $\text{sd}(\text{StSE})$ decreased. Even with $m = 20$, the differences between approaches and the deviation from one became small, except for the Bayesian-reference approach. The results for two-dimensional simulations were similar, although differences between approaches were a bit larger for $m = 9$ and deviations from one were often still substantial for $m = 25$.

4.2.2 Changing Uniform Prior for ϕ (Sessions 2 and 3)

Sessions 2 and 3 vary only in the upper bound of the uniform prior for ϕ (au = 100 and 2000, respectively) used as the basis for inference, rather than in the simulating model; for comparison, the same bounds in the REML and MML parameter searches were applied. Note that Bayesian-inverse-gamma and Bayesian-reference results (see Sect. 3.2.1), having their own bounds, are not repeated here. Also recall that the extent of the simulated data set was 300 (au). The seemingly arbitrary choice of the upper bound of the uniform prior for ϕ influenced the results, especially with few data (small m) and with the two-dimensional simulations (see Online Resource D).

Although MML and the Bayesian-uniform approach use the same range of possible ϕ values, MML was far less influenced by the upper bound for its parameter search. The proportion of $\hat{\phi}$ values (estimated by REML and MML, respectively) that were very close to the upper and lower bounds are also given (Online Resource C/E). Interestingly, in the case of a larger upper bound (session 3), the fraction of REML-estimated ϕ 's close to the unchanged lower bound was larger than in session 1, while the fraction of MML-estimated ϕ 's close to the lower bound stayed the same.

4.2.3 Varying Simulation Trend (Sessions 4 and 5)

The trend on the spatial coordinate (the standard) was also compared with a trend that was a simulated GRF itself ($\beta_1 = 0$, $\beta_2 = 2$, $\tau^2 = 0$, $\sigma^2 = 0.5$, $\phi = 30$), session 4, and with a constant mean, session 5. In both cases, the form of the trend (i.e., the design matrix) was assumed to be known for inference and prediction. The only difference between the sessions was the means: the simulated error signals for sessions 1–5 were identical. Compared with a trend on the coordinate, both the GRF trend and a constant mean gave only minor differences in $\text{mean}(\text{StSE})$; this also held for the two-dimensional simulations.

4.2.4 Misspecified Model (Sessions 6, 7 and 8)

In sessions 6 and 7, the error signal was simulated using a Matérn covariance function with large and small values for the smoothness parameter ν_{sim} (not to be confused with the degrees of freedom ν of a t -distribution used earlier). The inference in these sessions was still based on the exponential covariance model, which equals the Matérn model with $\nu_{\text{sim}} = 0.5$. These sessions were designed to provide a test of how the methods deal with a misspecified inferential model. The large ν_{sim} in session 6 caused all $\text{mean}(\text{StSE})$ to be far too conservative, with the Bayesian approaches slightly more conservative, and with average $\text{mean}(\text{StSE})$ values becoming smaller with increasing

m . In the two-dimensional simulations, the values stayed considerably closer to one. A small v_{sim} , as shown in session 7, caused almost all results to be optimistic. With increasing m , the mean(StSE) did not converge towards one, but rather seemed to stabilise at an optimistic value. With a nugget component added to the simulated data (session 8; with nugget-sill ratio 1/6), all approaches were optimistic (except Bayesian-reference and $m = 10$, and its two-dimensional counterpart with $m = 9$), and the average mean(StSE) increased with m in the one-dimensional simulations. In the two-dimensional simulations, the relation between m and mean(StSE) was ambiguous.

4.2.5 Simulation with Extreme Distance Parameter (Sessions 9 and 10)

If the distance parameter used for the simulations was very small in relation to the areas under consideration, such as in session 9, all approaches seemed to be quite optimistic, but this effect strongly decreased with increasing m . The worst performer was the Bayesian-inverse-gamma, where information encapsulated in $f_0(\phi)$ was now mismatched with the simulation model, although Bayesian-uniform also performed badly. The Bayesian-reference approach performed best. In the two-dimensional simulations, values were more extreme, especially for $m = 9$. When, as in session 10, the distance parameter was large compared with the total extent under consideration, REML performed almost perfectly while other approaches tended to be slightly or fairly conservative, but improved with increasing m .

5 Case Studies

5.1 Synthetic Case Study: Vegetation Index Data, with Validation on Point Support

To briefly investigate how REML, MML and full Bayesian would perform for a real-world data set, a remote-sensing vegetation index, CFAPAR-27, was used as the variable of interest. These data are used as a covariate in the real case study (spatial prediction of crop yield in Burkina Faso, Sect. 5.2) and therefore concisely described in Online Resource F. This spatial variable is, obviously, available on pixel support. The CFAPAR-27 data were masked using the crop yield mask (see also Sect. 5.2), and subsequently aggregated over the 45 provinces of Burkina Faso. As covariates for inference, two climate variables broadly representing rainfall and temperature (CRAIN-EC-27 and TMIN-EC-21) and one variable representing soil pH (PHAQ) were used. Gaussianity for all real-world variables of interest was assumed.

ATPK was applied using four approaches: (1) REML, (2) MML, (3) the full Bayesian approach using the uniform prior for ϕ , and (4) the full Bayesian approach using the reference prior for ϕ . For REML and MML, the parameter search for ϕ was bounded between 37 and 300 km, being roughly the smallest distance between the centres of any two areas, and one third of the largest extent of the region of interest, respectively. The same bounds defined the uniform prior for the full Bayesian approach.

The resulting mean(StSE) was 2.87, 2.73, 2.92 and 2.59 for the REML, MML, Bayesian-uniform and Bayesian-reference approaches, respectively, showing that prediction uncertainty was seriously underestimated by all approaches. The mean(StSE) of the Bayesian-uniform approach could be changed by several tenths by adjusting the bounds of the uniform prior. All RMSE values for the four approaches were around 6.19 (compared with the baseline approach RMSE of 16.54), indicating that they offered the same prediction quality and probably quite similar predictions.

5.2 Real Case Study: Crop Yield Data

As a real-world case study, this paper predicts yields of sorghum and millet, both cereal staple foods, in Burkina Faso, West Africa. The observation areas are the 45 provinces, for which only the average yields are known (averaged over the years 2000–2013, and provided by AGRHYMET), as shown in Fig. 6 for millet. Covariates for the trends as suggested by Brus et al. (2018) are used: for millet no covariates, and for sorghum four covariates are shown and briefly explained in Online Resource F.

REML, MML, Bayesian-uniform and Bayesian-reference approaches were applied, with similar settings to those of Sect. 5.1 in the previous analysis of the vegetation index data. The observed millet yields, and the resulting predictions and prediction uncertainties (standard deviations of the predictive distributions) when applying MML, are presented in Fig. 6; maps are presented first over the entire study region, then focused on a subregion to reveal more detail. Similar maps for sorghum are presented in Online Resource F.

Figure 7 shows the densities of the millet yield predictions and prediction standard deviations based on all four approaches, indicating that the Bayesian and, to a lesser extent, MML approaches generated larger prediction uncertainties than REML. For sorghum (see Online Resource F), the Bayesian-reference-calculated prediction diverted from the other approaches, due to the tendency of the distance parameter to move as close as possible to zero; the applied lower bound for the uniform prior for ϕ and for the REML and MML parameter searches imposed a limit on this effect. This shows again that a seemingly arbitrary choice of a uniform prior or of a parameter range for REML or MML might influence the resulting prediction uncertainty.

6 Discussion

6.1 Setting Uniform Prior

Both in the simulations (Table 3 and Online Resource C) and in the case studies, the choice of the upper and lower bounds of a uniform prior for ϕ can influence the prediction uncertainty, especially (but not exclusively) with smaller data sets and if the posterior mode of ϕ coincides with one of the bounds of the prior. This effect can also occur with REML and MML approaches, where the search for the optimum value of ϕ is bounded by the same limits. It should be stressed that, in this context, this ‘flat’ uniform prior cannot be considered uninformative. The fact that the posterior modes of ϕ (resulting from Bayesian approaches), or $\hat{\phi}$ (from the REML approach), often

Fig. 6 Millet case study. **a** Millet yields per area, averaged over the years 2000 to 2013; *source*: AGRHYMET (Traore et al. 2014). **b** Predicted yields and **c** prediction uncertainty, both from the MML approach. Subfigs. **d–f** Show corresponding map details from the south-western part of Burkina Faso

Fig. 7 Densities of predicted millet (**a**) and associated sd of prediction errors (**b**) in Burkina Faso, with restricted maximum likelihood (REML), maximum marginal likelihood (MML), Bayesian with a uniform prior for the distance parameter and Bayesian with the reference prior

coincided with one of the bounds (for example see the ' $\hat{\phi}$, $\text{mode}(\phi) \approx \min, \max$ ' columns in Online Resources C and E, but also sorghum in the case study) highlights the importance of carefully considering such prior or parameter search settings in geostatistical practice.

6.2 Reference Prior

According to the simulations, the reference prior did not perform well, being in many cases too conservative about prediction uncertainty, and pushing posterior distributions of the distance parameter too strongly towards zero (see also Online Resources C and D). In the case of small ν_{sim} or $\tau_{\text{sim}}^2 > 0$, this conservatism compensated to some extent for model misspecification. Berger et al. (2001) derived the form of the reference prior for analysis of spatial point-support data, and the same logic was applied—with area-to-area average correlations replacing the point-to-point correlations of Berger et al. (2001)—to justify a similar prior for analysis of areal-support data. However, although the logic to derive the form of the prior follows analogously, the authors are unsure of the analogous logic to ensure propriety of the resulting prior and posterior distributions. As such, even if the simulations would have demonstrated a strong advantage, it would require further work to derive the required proofs of propriety. With the simulation results not demonstrating strong advantages, other priors for ϕ are currently recommended.

6.3 Misspecified Model

Validation statistics about prediction uncertainties are very sensitive to the misspecification of variogram parameters that determine the smoothness of the spatial signal. Examples are the nugget parameter and the smoothness parameter of the Matérn covariance function, as demonstrated in the simulation sessions 6, 7 and 8 and, in the authors' opinion, in the vegetation index synthetic case study. Short-range spatial relationships are however difficult, if not impossible, to assess if only areal means are available. Situations with areal data combined with some high-density point data could improve the results, see for example Moraga et al. (2017); another approach would be to use prior information, such as expert opinions, for the nugget (Truong et al. 2014). In cases in which more summary data per area are available than only the mean, Orton et al. (2012) proposed a method for incorporating this information. In this situation, the exponential covariance model without a nugget was applied for convenience, as is often the case in comparable research; this is however a quite arbitrary choice, and given the results, a careful consideration of all model parameters that determine the smoothness of realisations is suggested for future research.

6.4 Number of Observations

In simulation sessions 6, 7 and 8 (very smooth or very rough simulated signals), $\text{mean}(\text{StSE})$ did not converge to one with an increasing number of observations m as

might be expected, and actually diverged from one in most cases. Therefore, more data do not alleviate a poor choice of model. Furthermore, in the simulation set-up, increasing m had the side effect of decreasing the size of the individual areas, which might also have influenced this behaviour due to short-range variations becoming better observable.

Very small data sets (nine or ten observations) were analysed in the simulations. The main point of interest was to see how the different approaches behave in such extreme situations, assessed by taking averages over many simulations. The authors stress that, even when using a Bayesian approach, any geostatistical conclusion based on nine or ten observations should be interpreted with caution, except perhaps if strong and honest prior information is available and can be incorporated.

6.5 One-Dimensional Versus Two-Dimensional Simulations

The effect of simulation choices and statistical inference approaches was quite similar in the one-dimensional and two-dimensional simulations. Differences might be due to different mutual spatial relationships. For example, for the two-dimensional simulations with nine observations, the closest pairs of units have centroids separated by 100 au. For the one-dimensional simulations with ten observations, neighbouring units' centroids are separated by 30 au. Therefore, despite there being an almost equal amount of data, there is much more short-range information in the one-dimensional data in the set-up used. This might explain the extremely large mean(StSE) values in session 9 of the two-dimensional study (up to 49.2) compared with the less extreme values in the one-dimensional study (up to 4.9).

The approximation of the average covariance matrices might have been less successful for the two-dimensional simulations. This would explain the relatively high max(MPP) and the unexpected irregular spatial pattern of the prediction error sd (see Online Resource D, Fig. D1 d).

6.6 The Algorithm

Although the authors did not compare the approach used here with more conventional MCMC methods, it proved an effective and efficient means of performing Bayesian (and also MML) spatial data analysis in the area-to-point context presented. As indicated in Sect. 3.2.3, several different methods can be used to approximate the average covariance matrices. For example, the Legendre–Gauss quadrature—as described by Orton et al. (2017)—is computationally and memory-wise much cheaper, but perhaps less accurate than the applied discretisation points method. Both methods, including some variations, are included in the code, as is area-to-area kriging. Future extensions might include directionality and point-to-point kriging.

The Integrated Nested Laplace Approximations (INLA) alternative to MCMC (Blangiardo et al. 2013) has some similarities to our partly analytical Bayesian approach, such as a gridded search in parameter space and numerical integration. However, it uses Laplace approximations of some integrals and is applicable to a much wider range of models, including hierarchical ones. Furthermore, its spatial implemen-

tation assumes the Markovian property on the spatial Gaussian random field (meaning that any point or area in the region is influenced only by its immediate neighbours), leading to sparse covariance matrices and thus reductions in computational costs. In our opinion, our approach offers specific and transparent insights into the Bayesian approach of model-based geostatistics. For future research, however, it would be interesting to redo the calculations with INLA, or to integrate some of the sophisticated and cost-reducing details of INLA into the code.

6.7 General Remarks

The general impression obtained from the ensemble of simulations is that REML tended to underestimate prediction uncertainty the most, followed by Bayesian-uniform. The Bayesian-reference approach tended to be more conservative, while MML was slightly conservative but seemed relatively stable. The differences between the approaches decreased with increasing m .

Given a covariance model that is more or less accurate in terms of short-range behaviour, the conclusion is that, for data sets of sufficient size, or if a slight underestimation of prediction uncertainty is allowed, the REML approach as demonstrated by Brus et al. (2018) should be sufficient. For smaller data sets with no prior information available, the most robust and in many cases best approach, albeit somewhat conservative, appeared to be MML. An additional advantage of MML is its relative insensitivity to arbitrary choices such as bounds on the correlation distance parameter. In several sessions, MML even outperformed Bayesian-inverse-gamma when the supplied prior information about ϕ was correct. Finally, MML has additional computational benefits for prediction over the fully Bayesian approach.

The authors suggest focusing future research on modelling short-range variation and including a smoothness parameter in the inferential models. Using honest and informative priors—depending on the research question at hand—might also yield interesting results. The Matérn smoothness parameter ν_M could be made an integral part of the Bayesian model, or alternatively incorporated as an extra model parameter to be optimised in an MML approach, which could then be used as a plug-in value for prediction.

7 Conclusions

All tested geostatistical approaches for ATPK (REML, MML, and Bayesian with different priors for the distance parameter) provided very similar predictions, but were different in the prediction uncertainties, with REML slightly underestimating the uncertainty in the case of very few data. Prediction uncertainties are quite sensitive to the parameters determining the smoothness of the spatial signal (i.e., nugget and smoothness parameter of the Matérn covariance function). Given correctly modelled short-range effects, for data sets of sufficient size, or if an underestimation of prediction uncertainty is allowed, the REML approach as demonstrated by Brus et al. (2018) is satisfactory. The MML approach (maximum likelihood with trend and variance integrated out) provided acceptable results while being relatively robust

to arbitrary settings for the parameter search. Also, this approach does not need a choice of prior for the distance parameter. A useful and robust full Bayesian approach could not be accomplished, perhaps due to the lack of a good uninformative prior for the distance parameter of the covariance function; the reference prior as proposed by Berger et al. (2001) overestimated the prediction uncertainty in most cases. For real-world case studies, the demonstrated algorithms (<https://doi.org/10.4121/uuid:1fe0c01e-7ff67-435b-a240-800579adc6e6>) can be used.

Acknowledgements This study was supported by the SIGMA European Collaborative Project (FP7-ENV-2013 SIGMA-Stimulating Innovation for Global Monitoring of Agriculture and its Impact on the Environment in support of GEOGLAM-project), and by the Grains Research and Development Corporation (GRDC) of Australia under Project PROC-9175385. The helpful comments and suggestions from the anonymous reviewers and from the editor are acknowledged with gratitude.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix–Abbreviations

au	Abstract units (length)
ATPK	Area-to-point-kriging
BAK	Bayesian areal kriging
BATPK	Bayesian area-to-point kriging
FIR	Fraction inside region
GRF	Gaussian random field
INLA	Integrated Nested Laplace Approximation
MCMC	Markov Chain Monte Carlo
ME	Mean error
ML	Maximum likelihood
MML	Maximum marginal likelihood
MPP	Mass preserving property
StSE	Standardised squared error
REML	Restricted maximum likelihood
RK	Regression kriging
RMSE	Root mean squared error
UK	Universal kriging

References

- Albert J (2009) Bayesian computation with R. Springer, New York. <https://doi.org/10.1007/978-0-387-92298-0>
- Berger JO, de Oliveira V, Sansó B (2001) Objective Bayesian analysis of spatially correlated data. *J Am Stat Assoc* 96(456):1361–1374
- Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with R-INLA. *Spat Spatio-temporal Epidemiol* 4:33–49. <https://doi.org/10.1016/j.sste.2012.12.001>

- Brus DJ, Orton TG, Walvoort DJJ, Reijneveld JA, Oenema O (2014) Disaggregation of soil testing data on organic matter by the summary statistics approach to area-to-point kriging. *Geoderma* 226–227:151–159. <https://doi.org/10.1016/j.geoderma.2014.02.011>
- Brus DJ, Boogaard H, Ceccarelli T, Orton TG, Traore S, Zhang M (2018) Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging. *Eur J Agron* 97(July):48–59. <https://doi.org/10.1016/j.eja.2018.05.003>
- Christensen OF, Roberts GO, Skld M (2006) Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *J Comput Graph Stat* 15(1):1–17. <https://doi.org/10.1198/106186006x100470>
- Diggle P, Ribeiro PJ (2007) *Model-based geostatistics*. Springer, Berlin. <https://doi.org/10.1007/978-0-387-48536-2>
- Goovaerts P (2006) Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *Int J Health Geogr* 5(1):52. <https://doi.org/10.1186/1476-072x-5-52>
- Goovaerts P (2008) Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Math Geosci* 40(1):101–128. <https://doi.org/10.1007/s11004-007-9129-1>
- Harville DA (1974) Bayesian inference for variance components using only error contrasts. *Biometrika* 61(2):383–385. <https://doi.org/10.2307/2334370>
- Horta A, Pereira MJ, Goncalves M, Ramos T, Soares A (2014) Spatial modelling of soil hydraulic properties integrating different supports. *J Hydrol* 511(Supplement C):1–9. <https://doi.org/10.1016/j.jhydrol.2014.01.027>
- Jansen MJ (1998) Prediction error through modelling concepts and uncertainty from basic data. *Nutr Cycl Agroecosyst* 50(1):247–253. <https://doi.org/10.1023/a:1009748529970>
- Kerry R, Goovaerts P, Rawlins BG, Marchant BP (2012) Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* 170(Supplement C):347–358. <https://doi.org/10.1016/j.geoderma.2011.10.007>
- Kitanidis PK (1983) Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resour Res* 19(4):909–921. <https://doi.org/10.1029/WR019i004p00909>
- Kitanidis PK (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resour Res* 22(4):499–507. <https://doi.org/10.1029/WR022i004p00499>
- Kyriakidis PC (2004) A geostatistical framework for area-to-point spatial interpolation. *Geogr Anal* 36(3):259–289. <https://doi.org/10.1111/j.1538-4632.2004.tb01135.x>
- Lark RM (2000) Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *Eur J Soil Sci* 51(4):717–728. <https://doi.org/10.1046/j.1365-2389.2000.00345.x>
- Lark RM, Cullis BR (2004) Model-based analysis using reml for inference from systematically sampled data on soil. *Eur J Soil Sci* 55(4):799–813. <https://doi.org/10.1111/j.1365-2389.2004.00637.x>
- Le ND, Zidek JV (1992) Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging. *J Multivar Anal* 43(2):351–374. [https://doi.org/10.1016/0047-259X\(92\)90040-M](https://doi.org/10.1016/0047-259X(92)90040-M)
- Lindley D (2004) That wretched prior. *Significance* 1(2):85–87. <https://doi.org/10.1111/j.1740-9713.2004.0026.x>
- Malone BP, McBratney AB, Minasny B, Laslett GM (2009) Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154(12):138–152. <https://doi.org/10.1016/j.geoderma.2009.10.007>
- Marchant BP, Lark RM (2004) Estimating variogram uncertainty. *Math Geol* 36(8):867–898. <https://doi.org/10.1023/B:MATG.0000048797.08986.a7>
- Marchant BP, Lark RM (2007) Optimized sample schemes for geostatistical surveys. *Math Geol* 39(1):113–134. <https://doi.org/10.1007/s11004-006-9069-1>
- McElreath R (2016) *Statistical rethinking: a Bayesian course with examples in R and Stan*. Chapman & Hall/CRC texts in statistical science series. CRC Press, Boca Raton
- Minasny B, Vrugt JA, McBratney AB (2011) Confronting uncertainty in model-based geostatistics using Markov Chain Monte Carlo simulation. *Geoderma* 163(34):150–162. <https://doi.org/10.1016/j.geoderma.2011.03.011>
- Moraga P, Cramb SM, Mengersen KL, Pagano M (2017) A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spat Stat* 21:27–41. <https://doi.org/10.1016/j.spasta.2017.04.006>
- Oliver M, Webster R (2014) A tutorial guide to geostatistics: computing and modelling variograms and kriging. *Catena* 113:56–69. <https://doi.org/10.1016/j.catena.2013.09.006>

- Orton TG, Saby NPA, Arrouays D, Walter C, Lemerrier B, Schvartz C, Lark RM (2012) Spatial prediction of soil organic carbon from data on large and variable spatial supports. I. Inventory and mapping. *Environmetrics* 23(2):129–147. <https://doi.org/10.1002/env.2136>
- Orton TG, Romn Dobarco M, Saby NPA (2017) Kriging based on areal summary statistics data: effects of within-unit variability on predictions and uncertainties. *Spat Stat* 19:42–67. <https://doi.org/10.1016/j.spasta.2016.11.003>
- Orton TG, Mallawaarachchi T, Pringle MJ, Menzies NW, Dalal RC, Kopittke PM, Searle R, Hochman Z, Dang YP (2018) Quantifying the economic impact of soil constraints on Australian agriculture: a case-study of wheat. *Land Degrad Dev* 29(11):3866–3875. <https://doi.org/10.1002/ldr.3130>
- Pardo-Iguzquiza E, Dowd P (2001) Variance–covariance matrix of the experimental variogram: assessing variogram uncertainty. *Math Geol* 33(4):397–419. <https://doi.org/10.1023/a:1011097228254>
- Park NW (2013) Spatial downscaling of trmm precipitation using geostatistics and fine scale environmental variables. *Adv Meteorol* 2013:9. <https://doi.org/10.1155/2013/237126>
- Roth M (2013) On the multivariate t distribution. Report, Department of Electrical Engineering, Linköpings universitet. <http://users.isy.liu.se/en/rt/roth/student.pdf>. Accessed 30 Nov 2018
- Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman and Hall, London
- Seaman JW, Seaman JW, Stamey JD (2012) Hidden dangers of specifying noninformative priors. *Am Stat* 66(2):77–84. <https://doi.org/10.1080/00031305.2012.695938>
- Steinbuch L, Orton TG, Brus DJ (2019) Source code in the R programming language, belonging with: model based geostatistics from a Bayesian perspective: investigating area to point kriging with small datasets [Dataset]. 4TU. Centre for Research Data. <https://doi.org/10.4121/uuid:1fe0c01e-7f67-435b-a240-800579adc6e6>
- Stigler SM (2007) The epic story of maximum likelihood. *Stat Sci* 22(4):598–620. <https://doi.org/10.1214/07-STS249>
- Traore SB, Ali A, Tinni SH, Samake M, Garba I, Maigari I, Alhassane A, Samba A, Diao MB, Atta S, Dieye PO, Nacro HB, Bouafou KGM (2014) Agrhymet: a drought monitoring and capacity building center in the west africa region. *Weather Clim Extremes* 3:22–30. <https://doi.org/10.1016/j.wace.2014.03.008>
- Truong PN, Heuvelink GBM, Pebesma E (2014) Bayesian area-to-point kriging using expert knowledge as informative priors. *Int J Appl Earth Obs Geoinf* 30:128–138. <https://doi.org/10.1016/j.jag.2014.01.019>
- Wackernagel H (2014) Geostatistics. Wiley StatsRef: Statistics Reference Online. <https://doi.org/10.1007/978-3-662-05294-5>
- Webster R, Oliver MA (2007) Geostatistics for environmental scientists. Wiley, Hoboken
- White P, Gelfand A, Utlaut T (2017) Prediction and model comparison for areal unit data. *Spat Stat* 22:89–106. <https://doi.org/10.1016/j.spasta.2017.09.002>
- You L, Wood S, Wood-Sichra U (2009) Generating plausible crop distribution maps for sub-saharan africa using a spatially disaggregated data fusion and optimization approach. *Agric Syst* 99(23):126–140. <https://doi.org/10.1016/j.agry.2008.11.003>