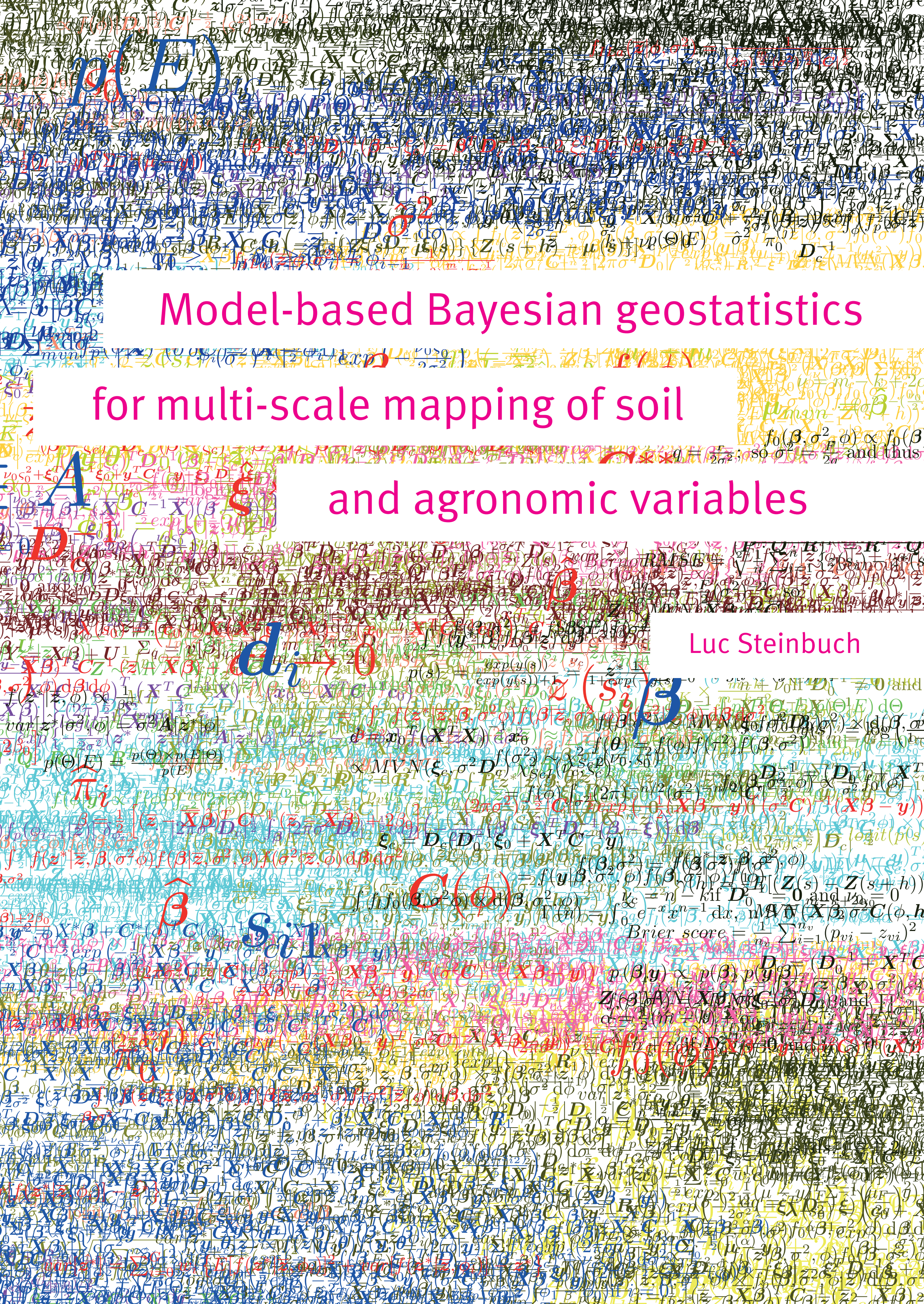


Model-based Bayesian geostatistics for multi-scale mapping of soil and agronomic variables

Luc Steinbuch

Propositions

1. The often used expression ‘flat prior’ for a low-informative prior is confusing (this thesis)

2. In this data-flooded era there is still need for the advancement of “small data science” (this thesis)

3. Presenting scientific concepts in a difficult way harms their scientific quality

4. All scientific models are useful

5. Good scientists make bad programmers

6. A manuscripts’ accompanying source code should undergo the same rigid review process as the manuscript itself

7. All members of society – not only Bayesians – should be aware of uncertainty in any form of data, information, knowledge and wisdom

8. The urgency for behavioural change to solve the climate crisis transcends by far the urgency of behavioural change regarding the COVID-19 crisis

Propositions belonging to the thesis, entitled:
Model-based Bayesian geostatistics for multi-scale mapping of soil and agronomic variables

Luc Steinbuch
Wageningen, 1 February 2021

Note: The public defence is postponed till Thursday 22 April 2021

Model-based Bayesian geostatistics for multi-scale mapping of soil and agronomic variables

Luc Steinbuch

Thesis committee

Promotor
Prof. Dr Gerard B.M. Heuvelink
Special Professor Pedometrics and Digital Soil Mapping
Soil Geography and Landscape group,
Wageningen University & Research
ISRIC – World Soil Information, Wageningen

Co-promotor
Dr Dick Brus
Researcher Biometris,
Wageningen University & Research

Other members
Prof. Dr Arnold K. Bregt, Wageningen University & Research
Prof. Dr Murray Lark, University of Nottingham, United Kingdom
Dr Laura Poggio, ISRIC – World Soil Information, Wageningen
Prof. Dr Patrick Bogaert, Université catholique de Louvain, Belgium

This research was conducted under the auspices of the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC).

Model-based Bayesian geostatistics for multi-scale mapping of soil and agronomic variables

Luc Steinbuch

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr A. P. J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 1 February 2021
at 4 p.m. in the Aula.

Note:
The public defence is postponed till
Thursday 22 April 2021, at 11 a.m.

Luc Steinbuch
Model-based Bayesian geostatistics for multi-scale mapping of soil and agronomic variables, 196 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)
With references, with summary in English

ISBN 978-94-6395-587-4
DOI 10.18174/532894

Table of contents

$\alpha_0 > 0, \beta_0$

$f_p(\beta_q|\overline{z})\ p$

$f(\boldsymbol{z}^*|\boldsymbol{z}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\phi})$

$f_p(\phi|\overline{\boldsymbol{z}})$

$\boldsymbol{D}_c = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T$

$\alpha = \frac{m-k+2\alpha_0}{2}$

$f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_{\boldsymbol{\beta},\phi} f_p(\boldsymbol{\beta}|\overline{\boldsymbol{z}})$

$= -\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T \boldsymbol{D}_0^{-1}\right.$

$\alpha = \frac{1}{2}(m-k) + \alpha_0$

$\boldsymbol{D}_c = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$

$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})}$

$f(\boldsymbol{z}^*|\boldsymbol{z}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\phi})$

$\gamma(h) = \frac{1}{2}E[(\boldsymbol{Z}(s) - \boldsymbol{Z}(s+h))^2]$

$\hat{\boldsymbol{\xi}} = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$

$\boldsymbol{D}_c = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$

$\boldsymbol{\xi}_c = \boldsymbol{D}_c(\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0$

$(2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}\,exp$

$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)$

$\alpha = \frac{m-k+2\alpha_0}{2}$

$(\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)$

$= -\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T\boldsymbol{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0) + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right.$

$\alpha = \frac{1}{2}(m-k) + \alpha_0$

$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})}$

$\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$

$f_{ILa}(\overline{\boldsymbol{z}}|\phi,\sigma^2) =$

$\int f(\boldsymbol{z}^*|$

$\gamma(h) = \frac{1}{2}E[(\boldsymbol{Z}(s) - \boldsymbol{Z}(s+h))^2]$

$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \propto\ p(\boldsymbol{\beta})\ p(\boldsymbol{y}|\boldsymbol{\beta})$

$MVN(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)$

$\hat{\boldsymbol{z}}^*|\sigma^2,\phi = \overline{\boldsymbol{C}}^*\ \boldsymbol{C}^{-1}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + \boldsymbol{X}^*\hat{\boldsymbol{\beta}}$

$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2)$

$\propto |2\pi\sigma^2\boldsymbol{C}|^{-\frac{1}{2}}\,exp$

$\int f(\boldsymbol{z}^*|\boldsymbol{z},\boldsymbol{\beta},\sigma^2,\hat{\phi}) \times f(\boldsymbol{\beta},\sigma^2|\boldsymbol{z}$

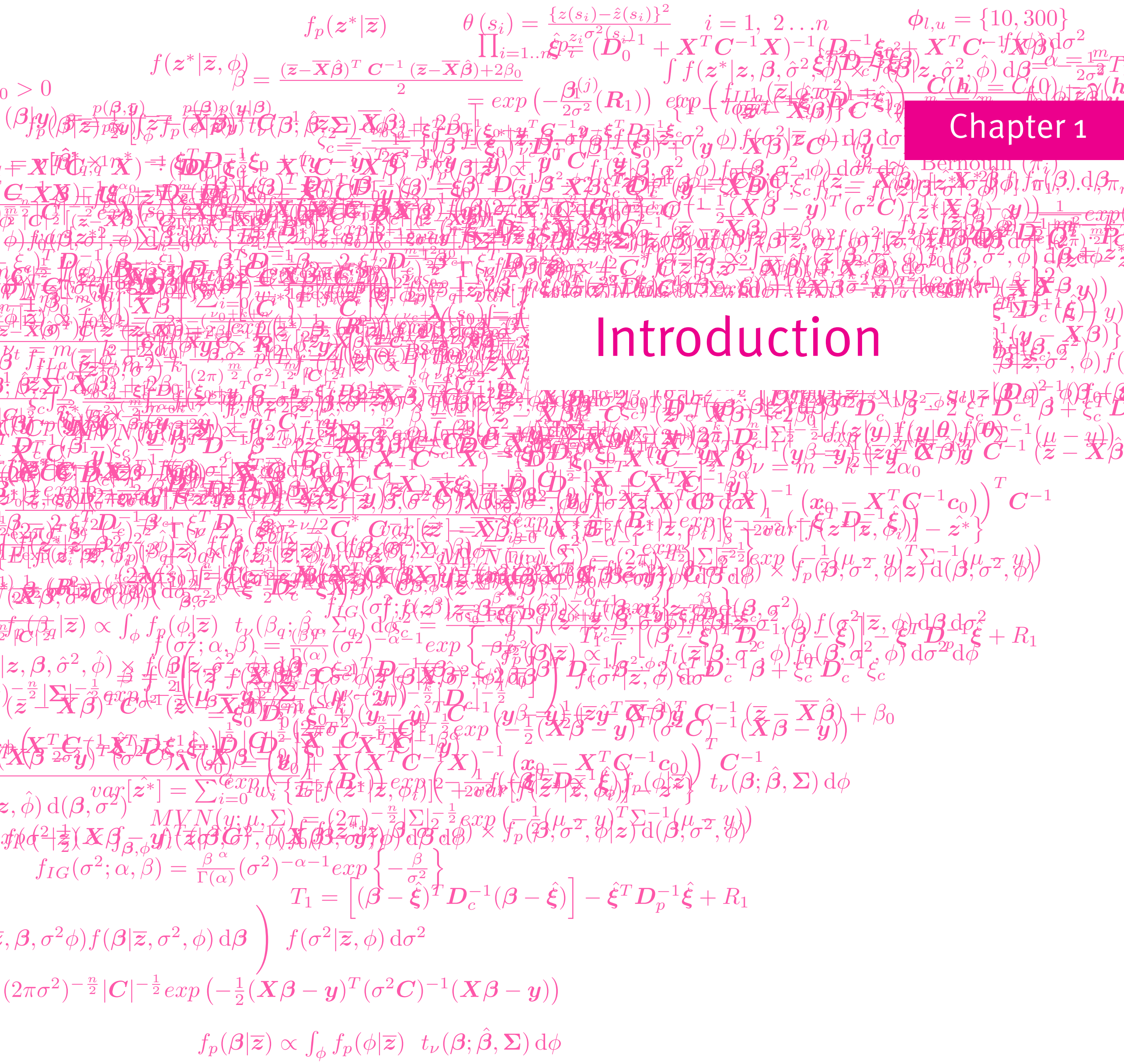
$(2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}e$

$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})}$

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

$= \int_{\sigma^2}\left(\int_{\boldsymbol{\beta}} f(\boldsymbol{z}^*|\overline{\boldsymbol{z}}\right.$

$= exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{R}_1)\right)\ exp\left(-\frac{1}{2\sigma^2}(-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}})\right)$


$f_p(z^*|\overline{z})$
$\theta\left(s_i\right) = \frac{\{z(s_i)-\hat{z}(s_i)\}^2}{2}$
$i = 1,\ 2 \ldots n$
$\phi_{l,u} = \{10, 300\}$
$var[\hat{z^*}] = \sum_{i=0}^{\infty} w_i$
$f_{IG}(\sigma^2;\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1} exp\left\{-\frac{\beta}{\sigma^2}\right\}$
$T_1 = \left[(\beta - \hat{\xi})^T D_c^{-1}(\beta - \hat{\xi})\right] - \hat{\xi}^T D_p^{-1}\hat{\xi} + R_1$
$f(\sigma^2|\overline{z},\phi)\,\mathrm{d}\sigma^2$
$(2\pi\sigma^2)^{-\frac{n}{2}}|C|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(X\beta - y)^T(\sigma^2 C)^{-1}(X\beta - y)\right)$
$f_p(\beta|\overline{z}) \propto \int_{\phi} f_p(\phi|\overline{z})\ \ t_{\nu}(\beta;\hat{\beta},\Sigma)\,\mathrm{d}\phi$

Chapter 1

Introduction

1.1 Mapping spatial soil- and agronomic variables

Many of today's global challenges, such as mitigating the climate crisis, protecting biodiversity and guaranteeing food safety partly rely on soils (Keesstra et al., 2016; McBratney et al., 2014), and thus on human's knowledge about soils. The research domain of *pedometrics* deals with "the development and application of statistical and mathematical methods applicable to data analysis problems in soil science" (McBratney et al., 2018). The focus of this thesis is on understanding soil spatial relationships, which are applied for mapping soils – the sub-domain of *digital soil mapping*, DSM (Zhang et al., 2017) – and also on mapping crop yields and yield gaps. Soil mapping quantifies soil properties over space and contributes to solutions for the above mentioned global challenges.

Soil and other agronomic variables can be mapped using various methods. The general focus is nowadays on machine learning (also called 'data mining') methods, relying on numerical approaches using abundant data, such as data from remote sensing (McBratney et al., 2000) and often also relying on abundant soil data or crop data itself. However, while machine learning methods undoubtedly have their benefits, using those methods risk finding irrelevant relationships (Wadoux et al., 2020) as well as obscuring causalities and natural laws (Spiegelhalter, 2019, Ch. 5). Also, machine learning methods might under-perform in correctly assessing the spatial uncertainty of a produced map (Fouedjio and Klump, 2019).

Even more, interesting and even indispensable data from other sources than remote sensing can still be very labour intensive or expensive to gather, such as observations from visually assessed soil profiles, chemical soil analysis, or crop yield gaps. Thus we still need methods to create maps based on relatively small data sets. Also it sometimes remains difficult to match data supply with knowledge demand (Hendriks et al., 2016; Grunwald et al., 2011); as an example, certain soil properties need to be quantified while only soil type is known. Another mismatch can be in spatial support: for instance, observations exist as point values but values concerning an area are needed, or vice versa. For such cases, machine learning methods might be too limited.

Therefore, to meet the needs of the future, we need to think about statistically explicit and consistent ways to combine and process data: data from different sources, with different uncertainty, with different spatial support, etc. We also need ways to combine new and legacy data, including existing beliefs in a more qualitative formulation (Truong, 2014). *Model-based geostatistics* provides tools to deal smartly, consciously and statistically soundly with spatial data (Magalhães et al., 2011; Diggle and Ribeiro, 2007), and enables – unlike most machine learning methods – the integration of varying spatial support data into one stochastic model (McKinley and Atkinson, 2020). Besides, using methods based on a statistical model that is a realistic representation of reality, one can assess the model functioning, unlike many data-driven approaches.

This thesis aims to feature current developments in model-based geostatistics to be used for the needs of today and the future, and also bring these developments in line with contemporary computational possibilities. This thesis explores the methods and limitations of model-based geostatistics in the context of mapping soil properties and crop yields, applied on the knowledge gaps defined later in this chapter.

1.2 Model-based geostatistics

Geostatistics is part of the broader field of spatial statistics, and was originally designed for geological and other environmental data, having its own terminology and workflow. Geostatistics describes spatial phenomena with the help of spatial correlation, beside more general descriptions such as overall variance, overall mean, and dependency on eventual independent variables. Spatial correlation is modelled by considering observed reality as one realisation of a spatial random process[1]. In our context, spatial random processes are characterised by decreasing covariance over separation distance (Webster and Oliver, 2007), or in the words of Tobler's first law of Geography: "Everything is related to everything else. But near things are more related than distant things" (citing Miller, 2004). The relation between covariance and separation distance is visually represented in a correlogram or semi-variogram, and summarised into a geostatistical model with a few parameters. Then, this geostatistical model enables spatial interpolation (also called *kriging*) of location-specific observed values into a prediction map, together with a map of the associated prediction uncertainty (Fouedjio and Klump, 2019).

Note that the concept of geostatistics also includes temporal processes, extending the methodology to *space-time models* (Cressie and Wikle, 2015; van Zoest et al., 2019), as well as *directionality* – the latter meaning that the covariance depends not only on separation distance but also on direction (Allard et al., 2016); both extensions are outside the scope of this thesis. Regarding the two other main fields of spatial statistics (Schabenberger and Pierce, 2002; Cressie and Wikle, 2015): we do not consider *spatial point processes* where the location of the object of interest is a random variable in itself; and we will only slightly touch concepts regarding *lattice data*, which deals with spatial data related to an area divided into a set of fixed and discrete (thus countable) locations – an example is provinces in a country.

Model-based geostatistics aims at explicitly applying formal statistical methods into the field of geostatistics (Diggle and Ribeiro, 2007). For example, traditional geostatistics (Journel and Huijbregts, 1978) produces a variogram cloud – plotted as a point cloud – showing the observed semi-variance depending on distance. Using arbitrary choices, this variogram cloud is summarised into an empirical variogram. Then, the geostatistical model parameters are estimated by fitting a line through the empirical variogram points, using a pragmatic numerical algorithm (Banerjee et al., 2004) or even by visually guided tuning (Oliver and Webster, 2014). On the contrary, using model-based geostatistics one can estimate or infer model parameters as well as predictions using sound methods developed in general statistics, where – apart from model choice and related assumptions – no subjective choices have to be made (Diggle and Ribeiro, 2002). Also *Bayesian geostatistical models* are part of the repertoire of model-based geostatistics.

[1] 'Random' in this context means the same as 'stochastic': containing an element of randomness.

1.3 Bayesian statistics in the context of model-based geostatistics

An overview of *Bayesian statistics* from a science history perspective and discussion of its place in modern statistics, research and society is given by McGrayne (2011), while McElreath (2016) provides a mathematical and also intuitive treatment of the subject. Bayesian statisticians use probability to express their knowledge – or inversely formulated: their ignorance – about the world; this view is one of the three main statistical paradigms[2] (Spiegelhalter, 2019). By gathering data (observing events, doing experiments, etc.) our knowledge and thus the associated probability changes – hence the former name ‘inverse probability’. More mathematically, the combination of pre-observation knowledge with gathered data will produce a probability distribution of the variables of interest, eventually combined with the prediction of unobserved quantities, also as probability distributions. According to Diggle and Ribeiro (2002), the Bayesian extension of model-based geostatistics facilitates an accurate estimation of parameter and prediction uncertainty because of the consistent underlying model, while allowing incorporation of pre-observation knowledge into the final results.

Even more, Bayesian statistics provides a coherent framework to build a hierarchical statistical model, consisting of several layers. It does so by modelling variables which are related, while these relationships are expressed in other variables in another layer, but conveniently being part of the same model (Gelman et al., 2013). Note that all those variables are considered random variables, that is: having a probability distribution rather than being a single, deterministic value. This approach facilitates *hierarchical model-based geostatistics*, where the spatial random field forms one layer in the model, and the geostatistical model parameters form another layer (Banerjee et al., 2004; Diggle and Ribeiro, 2007). The realisation of the random field is conditional on the observations, while the pre-observation knowledge is connected to the geostatistical and regression parameters.

In practice, using such a hierarchical model can be cumbersome because there is often no analytical solution. Because of technological and mathematical developments in the last decades, it is nowadays possible to numerically infer the parameters of many Bayesian hierarchical models, although some models are computationally very expensive (Arab et al., 2017).

[2] The others are Fisher’s tests of significance and Neyman-Pearson tests of acceptance. Both are considered approaches from the *frequentist* school, where probabilities are defined as proportions as they appear in – often imaginary – very large repeatable experiments (McElreath, 2016).

1.4 Knowledge gaps

Given the rich toolbox offered by model-based geostatistics, I identified the following knowledge gaps related – among others – to food security and purposeful land use:

i In the research domain of agronomy, consciousness and experience is lagging behind how to translate potential yields, according to a plant growth model and known for certain point locations only, to numbers on regional level (so-called *spatial aggregation*), using a statistically correct way and including uncertainty quantification.

ii Combining legacy and new data in soil science is lacking a statistically sound basis.

iii Predicting on point level while the available observations have areal support (*spatial disaggregation*) is often based on only a few areas. This means that sparse data are available for model calibration. In such cases, the accuracy of the prediction as well the accuracy of the prediction uncertainty deserves extra attention.

iv The use of Bayesian statistics is increasing everywhere in science, because of mathematical, numerical and computational developments, but its application in pedometrics is lagging behind, perhaps because of unfamiliarity and usability issues.

1.5 Research objective, research questions

This thesis aims to explore the application of model-based geostatistics and Bayesian statistics in a spatial context, as well as their combination, to spatially predict soil- and agronomical properties including assessment of prediction uncertainties. Therefore, in this thesis I will address the following research questions:

1. What is the added value of model-based geostatistics vs. the usual climate zone approach in case of yield gap prediction, including inferences per country based on few point data?

2. Can legacy data be re-used in case of a generalised linear regression model to predict a binary soil property?

3. Does the use of Bayesian statistics for spatial prediction of crop yields on national scale, with only regional data available, provide more accurate predictions and prediction uncertainties?

4. When mapping a binary soil property, is the prediction map more accurate when using Bayesian hierarchical model-based geostatistics instead of Bayesian generalised linear regression?

1.6 Thesis outline

Table 1.1 shows how the main chapters of this thesis, the knowledge gaps and the research questions are related.

Table 1.1: Overview main thesis chapters and their relation with knowledge gaps and research questions.

Ch.	Applied method	Knowledge gap	Research question	Case study
2	Model-based geostatistics	i	1	Crop yield gap, West-Africa
3	Bayesian statistics	ii, iv	2	Clay ripening, Netherlands
4	Bayesian model-based geostatistics	iii, iv	3	Crop yields, Burkina Faso
5	Bayesian model-based geostatistics, hierarchical	iv	4	Depth Pleistocene sand layer, Netherlands

Chapter 2, *Geostatistical interpolation and aggregation of crop growth model outputs*, addresses the added value of using model-based geostatistics, and the prediction (including uncertainty) of crop yield gaps based on point data, based on a case study with sorghum and millet in West-Africa.

Chapter 3, *Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors*, explores the use of legacy information to improve the accuracy of the prediction map. As motivating example we chose to map a binary soil property, clay ripening, in the west of The Netherlands.

Chapter 4, *Model-based geostatistics from a Bayesian perspective: investigating area-to-point kriging with small data sets*, investigates the accuracy of prediction uncertainties in case of sparse data when model-based geostatistics is applied on an area-to-point kriging situation. This is illustrated with disaggregating millet crop yields in Burkina-Faso.

In Chapter 5, *Mapping depth to Pleistocene sand with Bayesian generalised linear geostatistical models*, we thoroughly explain an existing implementation of a Bayesian generalised linear geostatistcal model and explore possible issues and their solutions. Then, using the depth of the Pleistocene sand layer in the province of Flevoland, reduced to a binary variable, we check if this approach provides more accurate maps than less complicated standard alternatives.

The final Chapter 6, *Synthesis and general discussion* discusses the achievements as presented in this thesis, together with general recommendations for future research.

$\boldsymbol{C}(\boldsymbol{h}) = E[\{\boldsymbol{Z}($

$E[\cdot]$

$\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{\beta}_0 + \sum_{i=1}^{p} \beta_i \ \times \ x_i(s) + \varepsilon(s) = X(s)^T \times \beta +$

$(\boldsymbol{X}^T \boldsymbol{C}$

$s\epsilon D, \boldsymbol{Z}(s)$

x_i

$\boldsymbol{\mu}(s)$

$\hat{\boldsymbol{V}} = diag($

$\boldsymbol{\varepsilon}(\boldsymbol{s})$

β_0

$\varepsilon(\boldsymbol{s}_1)$

$f_p(\boldsymbol{z}^*$

R^2_{adj}

$\boldsymbol{\mu}(\boldsymbol{s}) = E[Z(s)] = \boldsymbol{X}(\boldsymbol{s})$

$\beta_i \ (i = 1 \cdots k)$

$f(\boldsymbol{\beta}, \sigma^2|\phi,$

$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{C}$

$BIC =$

$\boldsymbol{C}(\boldsymbol{h}) = C(0) - \boldsymbol{\gamma}(\boldsymbol{h})$

$\theta(s_i) = \frac{\{z(s_i) - \hat{z}(s_i)\}^2}{\sigma^2(s_i)}$

$i = 1, \ 2 \ldots n$

$p(P|E) = \int p(P$

$f_{IG}(\sigma^2; \alpha,$

$|h_{12}|$

$\varepsilon(s)$

$\boldsymbol{s}_0$

$\gamma(h) = \frac{1}{2}E[(\boldsymbol{Z}(s) - \boldsymbol{Z}($

σ^2

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{C}$

$\boldsymbol{c}_0$

$\boldsymbol{\lambda}(s_0)$

$\boldsymbol{x}(s_i$

$\hat{z}(\boldsymbol{s}_0) = \ \sum_{i=1}^{n} \lambda_i \ z(\boldsymbol{s}_i)$

$\boldsymbol{\gamma}(\boldsymbol{h}) = constant \ for \ |h|$

$n + p + 1$

$n \times n$

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z(s_i) - \hat{z}}$

s_i

$\hat{z}(s_i)$

$z(s_i)$

$\widehat{\boldsymbol{\beta}}$

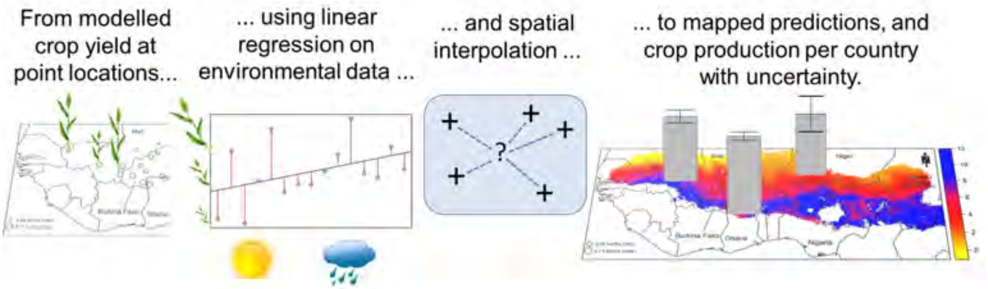
λ_i

Chapter 2

Geostatistical interpolation and aggregation of crop growth model outputs

Based on:
Luc Steinbuch, Dick J Brus, Lenny GJ van Bussel, Gerard BM Heuvelink, 2016. Geostatistical interpolation and aggregation of crop growth model outputs. *European Journal of Agronomy*, **v77**, pp.111-121.

Abstract Many crop growth models require daily meteorological data. Consequently, model simulations can be obtained only at a limited number of locations, i.e. at weather stations with long-term records of daily data. To estimate the potential crop production at country level, we present in this study a geostatistical approach for spatial interpolation and aggregation of crop growth model outputs. As case study, we interpolated, simulated and aggregated crop growth model outputs of sorghum and millet in West-Africa. We used crop growth model outputs to calibrate a linear regression model using environmental covariates as predictors. The spatial regression residuals were investigated for spatial correlation. The linear regression model and the spatial correlation of residuals together were used to predict theoretical crop yield at all locations using kriging with external drift. A spatial standard deviation comes along with this prediction, indicating the uncertainty of the prediction. In combination with land use data and country borders, we summed the crop yield predictions to determine an area total. With spatial stochastic simulation, we estimated the uncertainty of that total production potential as well as the spatial cumulative distribution function. We compared our results with the prevailing agro-ecological Climate Zones approach used for spatial aggregation. Linear regression could explain up to 70% of the spatial variation of the yield. In three out of four cases the regression residuals showed spatial correlation. The potential crop production per country according to the Climate Zones approach was in all countries and cases except one within the 95% prediction interval as obtained after yield aggregation. We concluded that the geostatistical approach can estimate a country's crop production, including a quantification of uncertainty. In addition, we stress the importance of the use of geostatics to create tools for crop modelling scientists to explore relationships between yields and spatial environmental variables and to assist policy makers with tangible results on yield gaps at multiple levels of spatial aggregation.


From modelled crop yield at point locations...
... using linear regression on environmental data ...
... and spatial interpolation ...
?
... to mapped predictions, and crop production per country with uncertainty.
Mali
Burkina Faso
Ghana
Niger
Nigeria

Abbreviations

CZ: agro-ecological climate zones; GYGA: global yield gap atlas; KED: kriging with external drift; LOOCV: leave one out cross validation; REML: restricted maximal likelihood estimation; RVH: regressor variable hull; RWS: reference weather station, reference weather station location; SCDF: spatial cumulative distribution function; sd: standard deviation, kriging standard deviation; se: standard error; SPAM: spatial plant allocation model; WOFOST: world food studies; Yp: yield potential; Yw: water-limited yield potential

2.1 Introduction

To support research and decision making related to global food security, mechanistic crop growth models are frequently used to calculate the potential yield of a food crop in a certain area and context. These models describe the build-up of harvestable biomass as a result of the interaction between plant physiology and environment (Roudier et al., 2011; van Ittersum et al., 2013). Many of these models require accurate daily meteorological data, preferably observations, instead of interpolated grid based data, due to the non-linearity of many weather-crop relationships (van Bussel et al., 2011; van Wart et al., 2013b) In addition, detailed and locally relevant information about crop management and soil information are required for accurate crop growth simulations (van Ittersum et al., 2013). Consequently, model simulations can only be obtained on a limited number of locations, i.e. close to weather stations with long-term records.

In several studies average crop estimates for large areas have been obtained by spatially aggregating location-specific crop model simulations, see e.g. Rosenzweig and Parry (1994), Wolf and Diepen (1995) and Alexandrov et al. (2002). A recently implemented approach is based on so-called agro-ecological Climate Zones (CZ) (van Wart et al., 2013a), applied e.g. in the Global Yield Gap Atlas (GYGA; www.yieldgap.org) (van Bussel et al., 2015). In this approach it is assumed that CZ are regions that are homogeneous with respect to climate conditions. The CZ approach is a straightforward and clear example of the calculate > interpolate > aggregate class of spatial aggregation approaches. An important drawback of this approach is that it ignores spatial variation of crop growth simulations within the climate zones, i.e. within these zones the simulated crop growth is assumed constant. Incorporating spatial variation could improve the spatial resolution and accuracy of the final results and thus help supporting national and local policy decisions, prioritizing investment strategies of fertilizer and seed companies and NGOs'. The CZ approach also fails to quantify the uncertainties associated with the interpolation and aggregation steps, which is essential information to guide accuracy improvement strategies (van Bussel et al., 2016). In this study we therefore explore whether the drawbacks of the CZ approach can be overcome with the help of a geostatistical approach. Geostatistics provides tools for a coherent quantification of site as well as aggregated modelled crop yield predictions. It produces continuous spatial maps that provide valuable location-specific information for crop modellers as well as decision makers and yields graphs that indicate areal proportions below or above a potential yield level threshold for regions or countries. It also offers means to explore the relationships between calculated yields and explanatory environmental variables.

The aims of this study are to present a state-of-the art model-based geostatistical method for spatial interpolation and aggregation of simulated yields, to illustrate it with a case study, and to compare the results with those of the common CZ approach. More specifically, we use kriging with external drift (KED), supported by restricted maximum likelihood parameter estimation (REML; Lark, 2000; Diggle and Ribeiro, 2007). Additionally, we use spatial stochastic simulation to predict aggregated crop production at country level and its associated uncertainty. As a case study, we interpolate and aggregate modelled yields of sorghum (*Sorghum bicolor*) and millet (e.g. *Pennisetum glaucum*, *Eleusine coracana*) in West Africa, according to as provided by the crop growth model WOFOST (Wolf et al., 2011; Supit et al., 2012).

2.2 Materials and Methods

2.2.1 Study area

This study has been carried out in West Africa, focussing on Burkina Faso, Mali, Ghana, Niger and Nigeria. Most of this area consists of a low plateau of maximal 500 metre above sea level, with some mountainous areas up to 2040 metres. The daily mean temperature is almost always and everywhere (except at high altitudes) above 18°C and relatively stable during the year. The most dynamic weather pattern is precipitation, dictated by dry winds from the Sahara in the north, dominant from November until February, and by the moist southwest marine wind, dominant in July (von Kaufmann et al., 1983).

2.2.2 Modelled crop yield data

Modelled crop yields for sorghum at 38 (Fig. 2.1) and millet at 37 Reference Weather Station locations (RWS) were obtained from the Global Yield Gap Atlas (www.yieldgap.org). Two yield levels, yield potential (Yp) and water-limited yield potential (Yw), were simulated using the crop growth model WOFOST version 7.1.3 (release March 2011) (Wolf et al., 2011; Supit et al., 2012). The yield potential is determined by solar radiation, temperature and carbon dioxide concentration; there are no limitations due to water stress, low soil fertility, weeds, pests, etc. The yield potential is further influenced by management practices like sowing date and cultivar choice. The water-limited yield potential, i.e. rainfed yield, is defined similar as Yp, except that possible water stress is taken into account (Evans, 1996; van Ittersum and Rabbinge, 1997).

The 38 and 37 locations used in the crop yield modelling were selected on: 1) the basis of proximity of weather stations with high-quality weather data, which are located in areas with high crop densities as indicated by You et al. (2006), You et al. (2009, see also http://mapspam.info) and 2) the dominant representation of the crop growing conditions in terms of weather, soils and cropping system for the countries of interest. Sorghum and millet share the same location 32 times. The final numbers of Yw on each location are area-weighted means of several simulations for dominant soil types, and both Yp and Yw are averaged over multiple years of simulation (Grassini et al., 2015; van Bussel et al., 2015)[1]. Summary statistics of simulated Yp and Yw for sorghum and millet are provided in Table 2.1.

2.2.3 Trend model covariates

In the kriging procedure described hereafter, we used grid maps of environmental and meteorological variables that are expected to be related to the simulated crop yield. To

[1] More detailed information about site selection and spatial support for the simulated point data can be found on www.yieldgap.org, section “Methods”

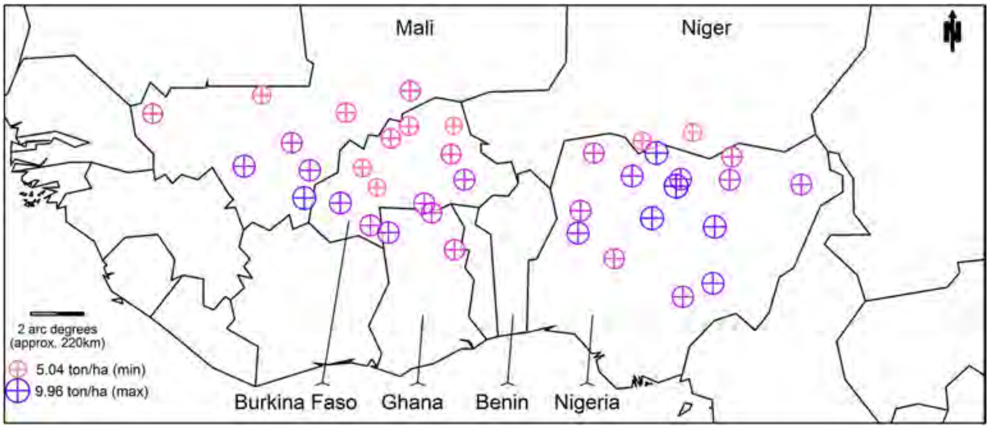


Figure 2.1: The yield potential (Yp; indicated by both circle area and colour scale) for sorghum at 38 locations as provided by the Global Yield Gap Atlas. Distance is indicated by geographical arc degrees; due to the map projection, the corresponding distance in km differs slightly across the map.

Table 2.1: Summary statistics of crop growth model yields Yp and Yw, for sorghum and millet. ‘n’ is the number of observations, i.e. the number of modelled crop yields per case. ‘skewness’ refers to the asymmetry of the dataset values. sd = standard deviation.

		Research cases			
	units	**Sorghum Yp**	**Sorghum Yw**	**Millet Yp**	**Millet Yw**
n	[-]	38	38	37	37
mean	[ton/ha]	7.50	6.21	4.23	3.02
sd	[ton/ha]	1.17	2.00	1.02	1.40
skewness	[-]	–0.15	–0.45	–0.66	0.09
min	[ton/ha]	5.04	2.04	1.21	0.56
max	[ton/ha]	9.96	9.71	5.90	5.80

Table 2.2: Meteorological variables (long-term averages) used in the trend models. The geographical resolution of all grid maps is 30 arc seconds. We derived the monthly temperatures from www.worldclim.org, accessed July 22nd, 2014.

Covariate	Units, description	Source
Aridity index	dimensionless; higher value means more water available for vegetation.	www.cgiar-csi.org Accessed Aug. 29, 2014
Degree days	°C × number of days; cumulative	Calculated from monthly temperature according to van Wart et al. (2013a).
Temperature seasonality	°C; a measure for temperature differences over the year	Calculated from monthly temperature according to van Wart et al. (2013a).

stay as close as possible to the agro-ecological Climate Zones method, we decided to use for all four cases a trend model with the three covariates used in the CZ approach (Table 2.2).

2.2.4 Geostatistical modelling

2.2.4.1 General framework: the geostatistical model

To build a geostatistical model, we first need to introduce the idea of a random field. A random field is a set of random variables indexed by location (Plant, 2012). Additional to the statistical model of a random variable, a random field has parameters describing its spatial correlation.

In this chapter, we build a statistical model of a random field for each of the four cases defined in Section 2.2.2. Thus, the crop growth model outputs for each case are considered realisations of four separate random fields. Our general statistical model of each random field is denoted by $\boldsymbol{Z} = \{\boldsymbol{Z}(\boldsymbol{s}),\ \boldsymbol{s} \in D\}$ (unit: ton/ha; $\boldsymbol{s}$ is a two-dimensional vector, representing geographic location, D is the geographic domain of interest). At each location $s \epsilon D$, $\boldsymbol{Z}(s)$ is modelled as the sum of a spatial trend (a linear regression part) and a stochastic residual (a random variable):

$$\boldsymbol{Z}(\boldsymbol{s}) = \boldsymbol{\beta}_0 + \sum_{i=1}^{k} \beta_i \times x_i(s) + \varepsilon(s) = X(s)^T \times \beta + \varepsilon(s) \tag{2.1}$$

where β_0 is the regression intercept, β_i $(i = 1 \cdots k)$ are regression coefficients associated with the covariates, the x_i are k environmental covariates[2] and $\varepsilon(s)$ is the stochastic residual. In matrix notation (the last part of Eq 2.1) the spatial trend is written as $\boldsymbol{X}(s)^T \times \boldsymbol{\beta}$, where $\boldsymbol{X}(s)^T$ is the transpose of a location specific column vector $\boldsymbol{X}(s)$ of size $k+1$ composed of $[1,\ x_1(s),\ \ldots,\ x_p(s)]$ and $\boldsymbol{\beta}$ is the $k+1$ vector $[\beta_0,\ \ldots, \beta_k]$. To allow comparison between the estimated regression coefficients, all covariates are scaled ('normalised') such that the means are zero and the standard deviations are one (Montgomery et al., 2001). The stochastic residual $\varepsilon(\boldsymbol{s})$ is assumed to be normally distributed with zero mean and constant variance, independent of $\boldsymbol{s}$. However, unlike

[2] Note that in other chapters of this thesis k includes the intercept

standard linear regression, in geostatistics ε is allowed to be spatially correlated. This spatial correlation is expressed in the covariance function $\boldsymbol{C}(\boldsymbol{h})$, where $\boldsymbol{h}$ refers to geographic distances. Here we assume that the covariance $C(h_{12})$ between $\varepsilon(\boldsymbol{s}_1)$ and $\varepsilon(\boldsymbol{s}_2)$ at any two locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ depends only on the separation vector $h_{12} = \boldsymbol{s}_1 - \boldsymbol{s}_2$. Under this assumption, the spatial dependence structure of $\boldsymbol{Z}$ is fully characterised by the covariance function, which is formally defined as (for readability, the multiplication sign × is omitted for the rest of this chapter):

$$\boldsymbol{C}(\boldsymbol{h}) = E\left[\{\boldsymbol{Z}(\boldsymbol{s}) - \boldsymbol{\mu}(\boldsymbol{s})\}\{\boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h}) - \boldsymbol{\mu}(\boldsymbol{s})\}\right] \tag{2.2}$$

where $E[\cdot]$ is mathematical expectation and $\boldsymbol{\mu}(s)$ is the mean of $\boldsymbol{Z}(s)$ (i.e., $\boldsymbol{\mu}(\boldsymbol{s}) = E[Z(s)] = \boldsymbol{X}(\boldsymbol{s})^T\boldsymbol{\beta}$). Alternatively, the spatial dependence structure of $Z(s)$ can be characterised by the variogram, which is a basic concept in geostatistics:

$$\gamma(h) = \frac{1}{2}E\left[(\boldsymbol{Z}(s) - \boldsymbol{Z}(s+h))^2\right] \tag{2.3}$$

With the assumption of a constant mean and a covariance that only depends on the separation vector, the variogram and covariance function are related by the identity:

$$\boldsymbol{C}(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h}) \tag{2.4}$$

where $C(0)$ is the covariance at distance 0, i.e. the variance of $\boldsymbol{Z}$. Note that $\boldsymbol{h}$ is a vector. A further simplification is to assume that the covariance function and corresponding variogram only depend on the length of the vector, the Euclidean distance $|h_{12}|$ between any s_1 and s_2.

Commonly used variogram models, such as the exponential and spherical model, are defined in Diggle and Ribeiro (2007) and Webster and Oliver (2007). The variogram model $\gamma(\boldsymbol{h})$= *constant for* $|h| > 0$ is referred to as a pure nugget model. With this model the residuals are spatially independent (i.e., the covariance equals 0 for distances greater than 0). In other words, the residuals show no spatial structure and consequently all spatial structure is captured by the trend. Estimation of a variogram from available data is discussed in Section 2.2.4.2.

2.2.4.2 Trend model calibration, variogram model selection and calibration

For all four cases the modelled crop yields and covariate values at the n input point data locations were used for fitting the trend and selecting and fitting a variogram model. For spatial models with a trend (as in this research), the variogram parameters can best be estimated by restricted maximum likelihood (REML) (Lark et al., 2006; Webster and Oliver, 2007). The REML procedure optimizes the parameters for a chosen variogram model directly from the data, by filtering out the trend in an analytical way. As REML maximizing algorithm we used differential evolution (Storn and Price, 1997), implemented in the R package `DEoptim` (Mullen et al., 2011).

We manually tried several variogram models, all with added nugget, and chose the model with the largest REML value. Given the REML estimates of the variogram model parameters, from which a covariance matrix $\boldsymbol{C}$ of residuals at observation locations is derived, the regression coefficients were next estimated by Generalised Least Squares (Webster and Oliver, 2007):

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{z} \tag{2.5}$$

The trend estimation algorithm also provides an R^2_{adj}, a measure for the goodness of fit of the trend model adjusted for the number of included covariates, and with standard errors for each of the elements of $\widehat{\boldsymbol{\beta}}$.

2.2.4.3 Defining the suitable area for prediction by limiting extrapolation

The combination of covariate values at the reference weather station locations (RWS), given by matrix $\boldsymbol{X}$, is the calibration domain. Care should be taken when the combination of covariate values of a prediction location is outside this calibration domain. An example: the aridity index (see Table 2.2) of the RWS in this research ranges approximately from 1,700 to 7,000. However, on the map of West-Africa the aridity index ranges from 25 (deep into the Sahara desert) to 30,000 (some coastal areas). Because of severe extrapolation, using these extreme values might produce unrealistic yield predictions (e.g., a negative sorghum yield) and the associated prediction uncertainty may underestimate the actual uncertainty, as this uncertainty is based on the assumption that the trend relation remains linear outside the calibration domain. Therefore, we cannot safely predict sorghum yields in the Sahara or at the coast. The same holds for a combination of covariate values outside the calibration domain, e.g., the combination of aridity index and degree days values, even if the aridity index values and the degree days values separately are each inside the calibration domain; this is known as ‘hidden extrapolation’.

The above means that the geographical area which is suitable for prediction should be restricted to locations with a combination of covariate values inside the calibration domain. Hence, we introduce the concept of ‘covariate space’: a k-dimensional mathematical space where each dimension is a covariate of the trend model. The covariate values of the observation dataset can be visualised as a point cloud in this covariate space. The scaled distance from the centre of this point cloud to any combination of covariate values $\tilde{x}_0$ (a vector, indicating a single point in the covariate space) is given by d (Montgomery et al., 2001):

$$d = \tilde{x}_0^T\left(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\right)^{-1}\tilde{x}_0 \tag{2.6}$$

where $\tilde{\boldsymbol{X}}$ equals design matrix $\boldsymbol{X}$ without the leading column of ones. The point cloud can be encapsulated within an ellipsoid, a k-dimensional ellipse, referred to as the regressor variable hull (RVH) by Montgomery et al. (2001). The axes of the RVH are scaled in every dimension, according to the distribution of the point cloud. The scaled distance from the RVH to the RVH-centre is d_{max}. We choose this threshold d_{max} such that the RVH touches the outside border of the observation point cloud. All covariate value combinations with d larger than d_{max} are considered an extrapolation in covariate space; thus their corresponding locations on the geographical map are considered as not suitable for yield prediction.

2.2.5 Spatial prediction: kriging with external drift (KED)

'Kriging' is a geostatistical interpolation method used to predict the value at any location based on a number of observations in the same area. 'External drift' is another expression for trend model, with covariates other than coordinates. In this research, the crop yield at a point location $\boldsymbol{s}_0$ in the study area is predicted as a weighted average of the crop growth model outputs provided by WOFOST at n point data locations:

$$\hat{z}\left(\boldsymbol{s}_0\right) = \sum_{i=1}^{n} \lambda_i \, z\left(\boldsymbol{s}_i\right), \tag{2.7}$$

where $z(s_i)$ is the crop growth model output at 'observation' location s_i and λ_i a kriging weight. The kriging weights are obtained by solving a system of $n+k+1$ linear equations (Webster and Oliver, 2007). In case of kriging with external drift (KED) these kriging weights are a function of: 1) the covariances, as discussed in the section 2.2.4.1, between the regression residuals at the observation locations and the prediction location: a vector $\boldsymbol{c}_0$ of length n; 2) the covariances between the residuals at each pair of observation locations: an $n \times n$ matrix $\boldsymbol{C}$; 3) the covariate values at the observation locations: an $n \times (k + 1)$ matrix $\boldsymbol{X}$, a composition of n vectors $\boldsymbol{x}(s_i)$; and 4) the covariate values at the prediction location with a leading one: a vector $\boldsymbol{x}_0$ of length $k + 1$. The kriging weights are computed by solving the matrix equation

$$\boldsymbol{\lambda}\left(s_0\right) = \left(\boldsymbol{c}_0 + \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{x}_0 - \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{c}_0\right)\right)^T \boldsymbol{C}^{-1} \tag{2.8}$$

where $\boldsymbol{\lambda}\left(s_0\right)$ is the vector of all λ_i for location $\boldsymbol{s}_0$. The variance of the prediction error (a measure for the uncertainty of the prediction within the geostatistical framework), also known as the kriging variance, at location $\boldsymbol{s}_0$ is given by

$$\sigma^2\left(\boldsymbol{s}_0\right) = \boldsymbol{C}(0) - \boldsymbol{c}_0^T\boldsymbol{C}^{-1}\boldsymbol{c}_0 + \left(\boldsymbol{x}_0 - \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{c}_0\right)^T\left(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{x}_0 - \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{c}_0\right) \tag{2.9}$$

The kriging variance contains both the uncertainty in the interpolated residuals (first two terms in Eq (2.9) and the uncertainty in the estimated mean (according to the trend model; third term in Eq. (2.9) (Brus and Heuvelink, 2007)). The square root of the kriging variance is known as the kriging standard deviation and commonly mapped alongside with the kriging predictions. It is taken as a summary measure for the uncertainty of the kriging prediction. For kriging, and for the related functions of cross-validation and spatial stochastic simulation (both explained later) we used the `gstat` package (Pebesma, 2004) in the R programming environment.

2.2.6 Spatial aggregation: total production and the spatial cumulative distribution function (SCDF) of production per country

To obtain a total production potential for a defined area, we sum up the predicted crop production over all suitable grid cells: spatial aggregation. For this we need to incorporate land use. The Spatial Plant Allocation Model (SPAM) dataset, version 2005, provides a spatially explicit estimate of the area cultivated by the crop per grid cell (You et al., 2009). For millet, we summed the SPAM maps for the two available kinds of this crop (millet pearl and millet small). Secondly, the geographical area where crop yield can be predicted by KED is limited as explained in Section 2.2.4.3. Therefore we restrict predictions to the suitable geographical area. Multiplying the predicted yield per grid cell, for all suitable locations, with the cultivated area per grid cell (as provided by SPAM) gives the predicted production per grid cell in ton. Summing the predicted production per grid cell, over all grid cells within a country finally gives the predicted total crop production potential within that country in tons.

However, unlike the prediction itself, the variance of the predicted total crop production potential per country cannot be calculated by summing the weighted variances per grid cell, because the prediction errors are correlated. Therefore we turned to a spatial stochastic simulation approach (Webster and Oliver, 2007). In this approach, a pseudo-random number generator is used to simulate a large number (we used 1000) ‘possible realities’ of crop yield maps based on the calibrated geostatistical model. For each grid cell, the average of the simulated crop yield approximates the predicted crop yield as obtained by kriging, while the variance of the simulated crop yield maps approximates the kriging variance. To perform the spatial aggregation with quantification of uncertainty, for every simulated crop yield map the total crop production in a country is computed by the same procedure as used for the predicted crop production, i.e. by summing up the simulated yield multiplied by the SPAM surface fraction for all suitable grid cells in the country. This results in as many possible values for the total crop production per country as the number of possible realities that had been generated. This set of values may be interpreted as a random sample of the probability distribution of the (weighted) aggregated crop production of the country (Heuvelink and Pebesma, 1999). Thus, the variance of this set of simulated values is an approximation of the variance of the predicted total crop production per country. Moreover, by sorting the simulated total crop production values the lower and upper bound of prediction intervals can easily be computed. For instance, by taking the 2.5 and 97.5 quantiles, the bounds of a 95% prediction interval are obtained.

The maps with simulated crop yield were also used to predict the Spatial Cumulative Distribution Function (SCDF; de Gruijter et al., 2006) of the crop production per country. From a SCDF one can read the proportion of land in a country for which the crop yield is below or above a given threshold, and the maximum crop yield for a given fraction of the country.

2.2.7 Cross-validation to quantify the modelling and calibration accuracy

The uncertainty associated with the predictions is quantified through the kriging variance and additionally by leave-one-out cross-validation (LOOCV). For LOOCV each location of the crop growth model outputs is taken out one-by-one, and its value is predicted based on the remaining crop growth model output locations, using the already estimated parameters for the trend and spatial correlation. The advantage of LOOCV is that we do not rely on modelling assumptions, i.e. it also provides realistic estimates of the quality of predictions if the assumptions made in spatial modelling are violated.

Four LOOCV statistics were computed. First, the mean error (ME) (Oliver and Webster, 2014) that quantifies the systematic error:

$$ME = \frac{1}{n}\sum_{i=1}^{n}\left(z\left(s_i\right) - \hat{z}\left(s_i\right)\right) \tag{2.10}$$

where $z\left(s_i\right)$ is the modelled crop yield (true value), and $\hat{z}\left(s_i\right)$ the geostatistically predicted crop yield at the LOOCV location s_i. In case of unbiased prediction, ME equals zero. The second LOOCV statistic is the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(z\left(s_i\right) - \hat{z}\left(s_i\right)\right)^2} \tag{2.11}$$

A smaller RMSE indicates more accurate predictions (Oliver and Webster, 2014). As a third validation measure, we calculated the correlation coefficient r between the $z\left(s_i\right)$ and $\hat{z}\left(s_i\right)$.

ME and RMSE are a function of the prediction errors only, not of the kriging variances. To verify whether the kriging variance is a proper measure of the prediction error, Lark (2000) proposed to compute the standardised squared prediction error at each cross-validation location. This fourth validation statistic is a function of the prediction error and the kriging variance σ^2:

$$\theta\left(s_i\right) = \frac{\left\{z\left(s_i\right) - \hat{z}\left(s_i\right)\right\}^2}{\sigma^2\left(s_i\right)} \quad i = 1,\ 2\ldots n \tag{2.12}$$

If the errors are normally distributed and the kriging variance is a correct assessment of the expected squared prediction error, this quantity has a Chi-square distribution with one degree of freedom. Hence, the mean of the $\theta\left(s_i\right)$ should be close to one. Lark

(2000) proposed to compute also the median of $\theta(s_i)$, as a median is less sensitive to outliers than a mean. Ideally the median of $\theta(s_i)$ should be close to 0.455. If the median and/or mean are close to their ideal values, on average the kriging variance is an unbiased quantification of the accuracy of the predictions.

2.3 Results

2.3.1 Model choice and calibration

2.3.1.1 Calibrated trend models

The calibrated spatial trend models for the sorghum and millet cases are given in Table 2.3. Recall that the estimated regression coefficients $\widehat{\beta}$ and the se (standard error) are based on scaled covariates, indicating the relative importance of the covariate and the precision of the estimated coefficient, respectively (Montgomery et al., 2001). Between 21% and 70% of the total variation was explained by the spatial trend; the Yw cases variation was better explained than the Yp cases variation. Every case had at least one covariate where the standard error was relatively low compared to the estimated coefficient values, indicating that the corresponding regression coefficient differed significantly (p-value ≤ 0.1) from zero.

Table 2.3: Scaled trend model coefficients for sorghum and millet.

covariate	Sorghum Yp R^2_{adj}: 0.53		Sorghum Yw R$^2{}_{adj}$: 0.70		Millet Yp R$^2{}_{adj}$: 0.21		Millet Yw R$^2{}_{adj}$: 0.58	
	$\widehat{\beta}$	se	$\widehat{\beta}$	se	$\widehat{\beta}$	se	$\widehat{\beta}$	se
(Intercept)	7.50	0.13	6.21	0.18	4.23	0.15	3.05	0.15
Aridity Index	0.44	0.44	1.60	0.60	0.17	0.49	0.20	0.49
Degree Days	-0.62	0.17	-0.53	0.23	0.38	0.19	0.13	0.19
Temperature Seasonability	0.03	0.41	0.20	0.56	-0.35	0.46	-0.94	0.46

2.3.1.2 Selected and calibrated variogram models

Table 2.4 presents the variogram models fitted by REML. In the case of sorghum Yp, the residuals from the spatial trend showed almost no spatial structure. In other words, almost all spatial structure was captured by the covariates. Because of this lack of spatial correlation, we cannot profit from spatial interpolation of the residuals, in terms of a reduction of the prediction error variance compared to the variance obtained with multiple linear regression.

The three other cases showed a spatial structure in the trend residuals; the range parameters were optimized to the maximal value as dictated by the settings of the REML maximizing algorithm.

Table 2.4: Variogram models fitted by REML for the four cases. Variogram type and variogram parameters are explained by Diggle and Ribeiro (2007) and Webster and Oliver (2007).

Case	**variogram type**	**variogram parameters**		
		nugget [(ton/ha)^2]	partial sill [(ton/ha)^2]	range [arc degrees]
Sorghum Yp	Nugget + Spherical	0.56	0.09	2.2
Sorghum Yw	Nugget + Exponential	0.87	1.48	20.0
Millet Yp	Nugget + Exponential	0.13	2.26	20.0
Millet Yw	Nugget + Exponential	0.32	1.39	20.0

2.3.2 Spatial prediction

Fig. 2.2 shows the predicted yield and kriging standard deviations (sd) for sorghum Yp and sorghum Yw. The maps of predicted millet Yp and millet Yw (and the corresponding sd) are provided as supplementary materials[3], together with summary statistics. At the unmapped, white parts of the land surface, the covariates are too different from the conditions at the RWS where crop yield was modelled; therefore these locations are outside the suitable area, as explained in Section 2.2.4.3, and no yield predictions were made at these locations. Note that the mapped areas for sorghum and millet are partly the same (for example in northern Mali and northern Niger), and partly different (for example in southern Ghana). Remember that the RWS where sorghum yield was modelled were partly the same, but also partly different from the RWS used for millet; this explains the similarities as well as the differences in the mapped area. For the same crop, the RWS for Yp and Yw were identical, thus those mapped areas are exactly the same.

Both sorghum yield maps (2.2a and 2.2c) show a clear north-south influence, with higher yields in the south. With millet, this tendency is rotated to higher yields in the south-west, decreasing towards north-east. The maps with kriging standard deviations show relatively large values far away from the RWS and near the border of the mapped area. Near the border several of the covariates have extreme values, so that the uncertainty about the mean (spatial trend) was relatively large. In case of sorghum Yp, no effect of distance to RWS can be seen. This is because the trend residuals showed almost no spatial structure. As a consequence the kriging variance (sd^2) was nearly constant throughout the area, almost equal to the nugget of the variogram (Table 2.4). The three other cases had a spatial structure in the trend residuals, which shows in the kriging sd being smaller near RWS locations. This effect is clearly visualized in the detail map in the supplementary materials, Fig. A.5.

[3] The supplementary materials, Appendix A, can be downloaded from the journal version of this chapter: Luc Steinbuch, Dick J Brus, Lenny GJ van Bussel, Gerard BM Heuvelink, 2016. Geostatistical interpolation and aggregation of crop growth model outputs. *European Journal of Agronomy*, **v77**, pp.111-121. https://doi.org/10.1016/j.eja.2016.03.007

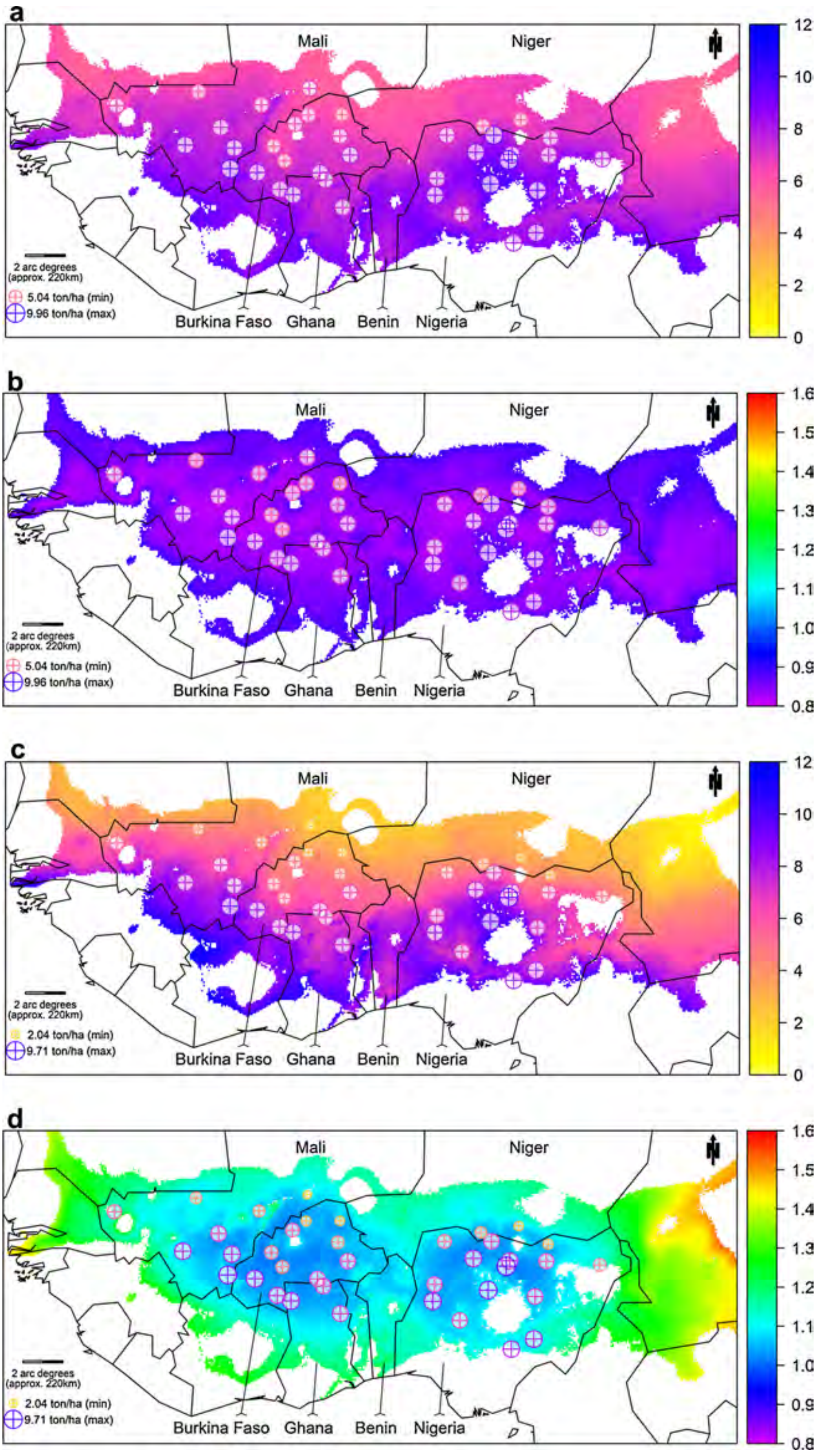


Figure 2.2: KED (kriging with external drift): prediction of sorghum yield potential (Yp; a) and its corresponding standard deviation (b); prediction of sorghum water-limited yield potential (Yw; c) and its corresponding standard deviation (d). All maps are in ton/ha. The colours of the circles indicate the crop modelled yield according to the legend of a and c. The circle diameter indicates the same property.

2.3.3 Spatial aggregation

Fig. 2.3 shows the geostatistically predicted average sorghum and millet yields (potential and water limited) per country. We calculated the average sorghum yield by dividing the predicted total crop production per country by the area tilled with the crop within the country; within each country, at least 70% of the SPAM surface is covered by the predictions. The error bars indicate the 95% prediction interval for the calculated yields. The figure also shows the corresponding yields according to the agro-ecological Climate Zone (CZ) approach, together with the actual yield per country as provided by the GYGA project. Because Benin is not considered by the GYGA project, we only have its geostatistical predictions. For all cases in all countries, the geostatistical approach yields were close to the CZ yields. Except for sorghum Yp in Niger, the CZ values were within the 95% prediction interval obtained with the geostatistical approach. In Niger, the difference between Yp and Yw was largest, for both crops and approaches. In Ghana, those differences were minimal. Except for Ghana, the differences between Yp and Yw were more pronounced for millet than for sorghum. The size of the prediction interval varied considerably, as examples: ±0.3 ton/ha (around a mean of 7.6) for sorghum Yp in Mali, contrasting ±0.7 ton/ha (around a mean of 4.1) for millet Yp in Benin.

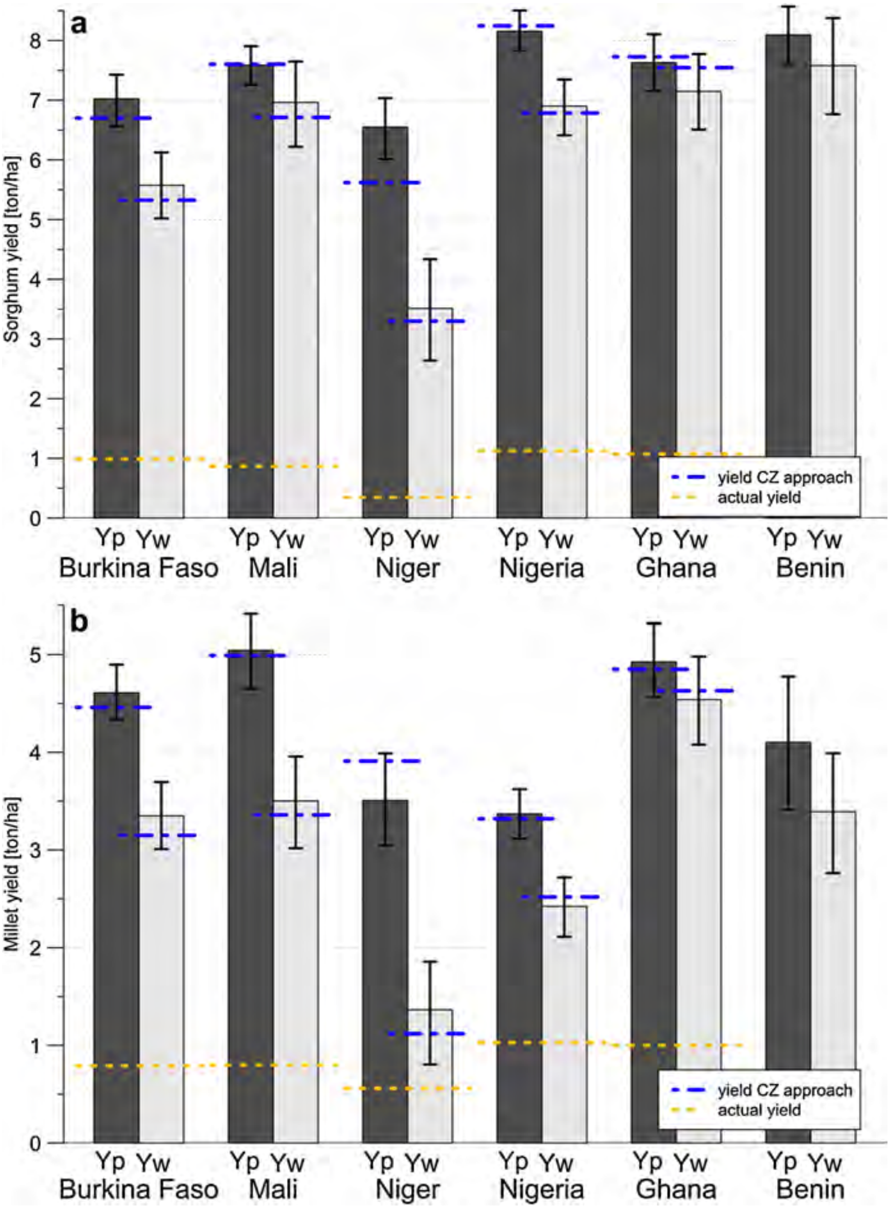


Figure 2.3: Predicted average sorghum (a) and millet (b) yield per country [ton/ha]. Yp: yield potential; Yw: water-limited yield potential. The bars indicate average yield according to the geostatistical approach, including the 95% prediction interval. For comparison, the horizontal lines show Yp and Yw according to the CZ approach, and an estimation of the actual yield (source: www.yieldgap.org/web/guest/sub-saharan-africa, accessed Oct 22, 2015).

The predicted spatial cumulative distribution function in Mali is visualized for sorghum Yp and Yw in Fig. 2.4. Following arrow (A) as interpretation example, Fig. 2.4a reads as: “It is estimated that for 50% of the cultivated area in Mali, Yp is 7.6 ton/ha or lower; with a probability of 95%, this estimated yield is between 7.3 and 7.9 ton/ha”. Following arrow (B): “It is estimated that a yield potential of 8.5 ton/ha or higher is achievable for 100-82 = 18% of the cultivated area in Mali; with a probability of 95%, this estimated area is between 100-91=9% and 100-72=28%”.

Comparison of Fig. 2.4a and b shows how accounting for water limitation decreases crop yield. In case of Yw, 50% of the area has a crop yield of just 7.0 ton/ha or higher, instead of the threshold of 7.6 ton/ha for Yp. Note also that accounting for water limitation heightens the left long tail visible in Fig. 2.4b, which represents marginal growing conditions. For example, in case of Yw, ca. 12% of the sorghum cultivated land in Mali has a predicted yield below 5 ton/ha; in case of Yp the area fraction with predicted yield below 5 ton/ha is practically zero.

Note that 50% of the area corresponds to the median, which is in both cases close to the means given in Fig. 2.3, indicating fairly symmetrical distributions.

2.3.4 Cross validation

The results of leave one out cross validation (LOOCV) are presented in Table 2.5. In all four cases the *ME* was very small showing the absence of bias. The *RMSE* ranged from 0.63 to 1.08. Compared to the standard deviation of the modelled crop yield at the 38 or 37 RWS (Table 1), the RMSE values were lower in all cases, with ratios varying between 0.53 (millet Yw) and 0.74 (sorghum Yp). This also agrees with the correlation coefficients r, which show that a substantial portion of the spatial variation in the modelled yields was explained by the KED model. The mean of the standardised squared prediction errors (θ) was close to the ideal value of one for all cases. Based on the median of θ, the squared prediction error appears to be overestimated by the kriging variance for sorghum Yp and millet Yp, and underestimated for sorghum Yw and millet Yw.

Table 2.5: Cross-validation results values for the four cases.

Case	ME	RMSE	r	mean(θ)	median(θ)
Sorghum Yp	-0.005	0.86	0.68	1.007	0.300
Sorghum Yw	-0.031	1.08	0.84	0.986	0.812
Millet Yp	-0.025	0.63	0.78	1.038	0.378
Millet Yw	-0.014	0.74	0.85	1.007	0.553
Ideal value	0.000	0.00	1.00	1.000	0.455

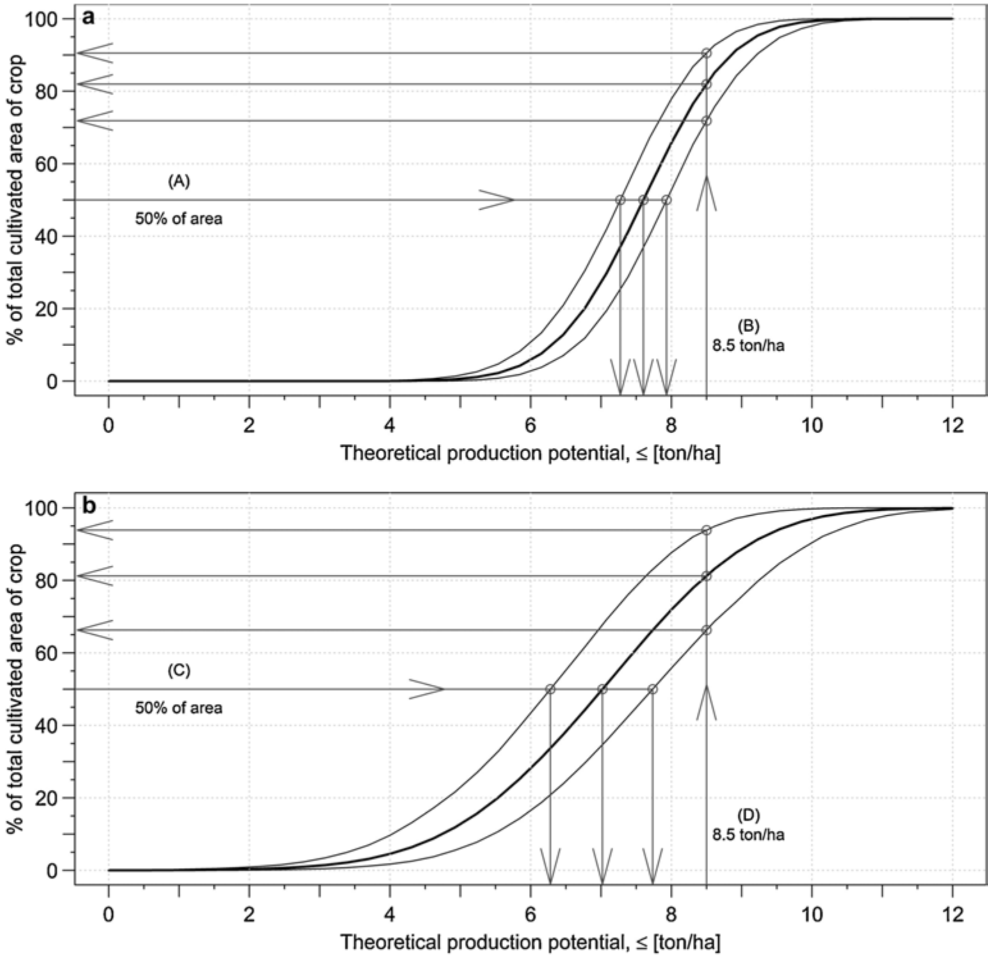


Figure 2.4: Predicted spatial cumulative distribution function of sorghum yield potential Yp (a) and sorghum water-limited yield potential Yw (b) over the sorghum cultivated area of Mali, with lower and upper limits of a 95% prediction envelope. See main text for explanation of arrows (A) and (B).

2.4 Discussion

In this chapter, we applied model-based geostatistical methods for spatial interpolation (i.e., prediction) and aggregation of crop yield model outputs. As case study we investigated sorghum and millet in West Africa, each with potential (Yp) as well as water-limited potential (Yw) yields. We built a spatial stochastic model (a fixed trend model combined with a variogram model), and used this model for spatial prediction and thereafter for spatial aggregation.

2.4.1 Model choice and calibration

2.4.1.1 Calibrated trend models

Investigating the calibrated trend models enables to explore some of the presumed key drivers of plant growth. According to Table 2.3, *Degree days* has a negative effect on the sorghum yields (-0.62 for Yp and -0.53 for Yw), a positive effect on millet Yp (0.38) and no significant (defined here as se larger than $\widehat{\beta}$) effect on millet Yw. This might be due to heat stress during the growing period, or to the relation between *Degree days* and growing period length; or perhaps *Degree days* is an indirect indicator for factors like total amount of precipitation during the growing season that are not incorporated in our trend model but play a role in the crop growth modelling itself. Because water availability is part of the modelling of sorghum Yw, we understand why the aridity index has a positive influence. However, for millet Yw there seems to be no such relationship. Actually, the majority of regression coefficients are not significantly different from zero. This is largely due to the limited number of observations. However, recall that we deliberately chose the same three covariates that define the CZ's. The findings presented in Table 2.3 hint that the definition and validity of the CZ's itself can be more thoroughly investigated with the help of spatial linear regression. For example, perhaps it is worth considering to use different CZ definitions for different crops and/or different CZ definitions for potential yield vs. water-limited potential yield.

The choice for a trend model, i.e. which covariates to use, is often arbitrary. Soil properties have not been included in the set of candidate covariates, although these are important inputs of the WOFOST model, especially for Yw. Because the modelled crop yield was calculated by an area-weighted mix of model outcomes using the soils nearby the reference weather station, rather than using the soil on the exact spot of the weather station, there was no obvious original soil property to offer to the trend model. A solution would be to use an average soil property (for example the total available water holding capacity) over a circular neighbourhood near the weather station. Recently produced high-resolution soil maps of Africa (Hengl et al., 2015) are already used by GYGA and would be very useful for this extension as well.

Unlike the CZ approach, the geostatistical approach did not take differences between countries into account. This would be a useful addition, because management-measures vary (and are included in the WOFOST model), but it would need much more reference weather station locations (RWS), in every included country.

Also other covariates like latitude, altitude etc. can be included. However, incorporating too many covariates – compared to the number of RWS - might result in ‘overfitting’: the trend model is essentially explaining noise, rather than the crop yield modelling mechanistics (Mundry and Nunn, 2009). Covariates should be selected with care.

2.4.1.2 Selected and calibrated variogram models

In this research, the variogram model describes the spatial correlation, or the lack of it, of the yields as far as this has not yet been explained by the calibrated trend models. From Table 2.1, and by visual inspection of the yield maps (Fig. 2.2), we can see that the spatial variation of sorghum Yp is the smallest. The calibrated trend model explains about half of the spatial variation (see the R^2_{adj} in Table 2.3); the rest of the variation has no detectable spatial structure. Despite the high values of the corresponding R^2_{adj}’s, the residual yields of sorghum Yw, millet Yp and millet Yw have a significant spatial structure (Table 2.4). The three fitted variograms with spatial correlation (sorghum Yw, millet Yp, millet Yw) have large variogram ranges that are larger than the spatial extent of the study area. This indicates presence of a geographical trend, which may be modelled by taking latitude and longitude as additional covariates into account. The presence of a residual spatial correlation and a geographical trend hints at additional causal covariates that have not been included in the trend model. For instance, latitude might be related to ‘hours of sunshine during growing season’.

2.4.1.3 Location and number of reference weather stations

The number and locations of the RWS used in this study were originally designed for the CZ-approach. For a geostatistical approach, the number of used RWS in this chapter (37 or 38) could be considered small. For a classical variogram parameter estimation, by method-of-moments, at least 100 observations are needed (Oliver and Webster, 2014); for state of the art likelihood based methods such as REML, Kerry and Oliver (2007) suggest a minimum of 50 observations. Although we succeeded in selecting, calibrating and validating geostatistical models with these small datasets, geostatistical theory and practices encourage to increase the sample size in future research. For this study area this was not possible because it lacks trustable weather stations with long-term data.

2.4.2 Spatial prediction

The prediction maps (Fig. 2.2 - a and c) show how the predicted yield potentials are affected by climatological variables, some of which follow common sense. For example, the sorghum yield potential (Yp) decreases towards the north, the dry Sahara desert. As expected, this effect is even stronger for the sorghum water-limited yield potential (Yw). The blank enclosure in central Nigeria indicates the elevated Jos Plateau: the elevation related climate variables place this area outside suitable area according to our extrapolation limitations measures. In the two Yw cases, the min and max values found by prediction (Table A.1 in the supplementary materials) cover a wider range than the minimum and maximum values of the modelled crop yields (Table 2.1), indicating an extrapolation in the prediction, despite our extrapolation limiting measures. For sorghum

Yp, the prediction yield range is just inside the modelled crop yield range (5.25-9.82 vs. 5.04-9.96 ton/ha). This shows that the prediction locations are not exactly the RWS locations with the most extreme values; due to the nugget effect, there might be substantial differences in prediction on short distances. For millet Yp, the smallest predicted yield (1.28 ton/ha) is a bit larger than the smallest modelled crop yield (1.21 ton/ha), but the maximal predicted yield is larger (8.16 ton/ha) than the maximal modelled crop yield (5.90 ton/ha).

The standard deviation (*sd*) of the prediction errors is for large parts of the maps small relative to the prediction: therefore we consider the prediction maps reliable. However, at places with marginal growing conditions *sd* is substantial compared to the predicted yield, such as in the north of the prediction area for the Yw cases.

2.4.3 Spatial aggregation

Predicted yields aggregated per country as obtained with the geostatistical approach and CZ-approach were comparable (Fig. 2.3). Compared to the uncertainty of the yield predictions at point-locations, the uncertainty in the aggregated yield is much lower. This is due to the averaging-out effect of aggregation, as element-wise dependent uncertainties partly cancel each other out when taken together. This might explain why the uncertainties of millet Yp and millet Yw are relatively large in Benin (Fig. 2.3b). The mapped millet yield area in Benin, taking the cultivated area into account, is very small compared to the mapped areas in all other countries (and also compared to sorghum in Benin). Hence the averaging-out effect for millet in Benin is smaller.

The supplied example of the spatial cumulative distribution function of sorghum in Mali (Fig. 2.4) shows that the geostatistical approach enables to answer detailed questions about e.g. the effect of irrigation on the total production in a country. Depending on the development of crop growth models, in the future similar approaches can be used for visualising the effect of other yield-limiting factors, for example fertilization.

Considering the outcomes (predicted yield, predicted production per country), there is more uncertainty than accounted for in this work, for example in crop growth modelling itself (Asseng et al., 2013), in the SPAM mask, and perhaps in the covariates. Additionally, yield plateaus in irrigated daily practice are about 80% of Yp (Cassman, 1999) and the actual land use, represented by the SPAM mask can easily change. To keep things simple we did not make use of available information about the uncertainty of the crop model outcomes. Research into how this information can be incorporated in the geostatistical approach is left for future work.

2.5 Conclusions

The model-based geostatistical approach presented and applied in this study has a number of benefits for spatial prediction and aggregation of modelled crop yield at weather stations:

- it provides a high-resolution map depicting predicted yield that varies continuously over the area of interest;
- it incorporates information contained in correlated environmental variables to improve spatial interpolation, and takes maximal use of the spatial correlation in trend residuals;
- it quantifies the prediction uncertainties, as well as the uncertainty of the aggregated production
- it supplies a spatial cumulative distribution function of the yield for countries or any other spatial region, including an uncertainty envelope;
- it uses a systematic and reproducible approach that involves few ad hoc decisions.

Because of these additional features, we stress the importance of applying geostatistical approaches in future crop yield mapping and aggregation.

Supplementary materials

The supplementary materials, Appendix A, can be downloaded from the journal version of this chapter: Luc Steinbuch, Dick J Brus, Lenny GJ van Bussel, Gerard BM Heuvelink, 2016. Geostatistical interpolation and aggregation of crop growth model outputs. *European Journal of Agronomy*, **v77**, pp.111-121; https://doi.org/10.1016/j.eja.2016.03.007

$$\mathcal{L}\left(\boldsymbol{\beta}\right)$$

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})}$$

π_i $\quad i = 1 \ldots n$ $\qquad \mathcal{L}\left(\boldsymbol{\beta}\right) = p\left(\boldsymbol{y}|\boldsymbol{\beta}\right) =$

$$\text{logit}\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{d}_i^T\boldsymbol{\beta} \qquad \widehat{\pi}_i$$

$$\pi_i = \text{logit}^{-1}\left(d_i^T\boldsymbol{\beta}\right) = \frac{\exp(d_i^T\boldsymbol{\beta})}{1+\exp(d_i^T\boldsymbol{\beta})} \qquad \boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\beta})$$

$$\propto \int_\phi \frac{k-1}{}$$

$$\hat{\boldsymbol{V}} = diag(\widehat{\sigma}_1^2,\ \widehat{\sigma}_2^2,\ \ldots,\ \widehat{\sigma}_n^2)$$

$$f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}$$

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = \left(X^T\hat{\boldsymbol{V}}X\right)$$

$$BIC = -2\log\left(\mathcal{L}\left(\hat{\boldsymbol{\beta}}\right)\right)$$

$$f_p(\beta|\boldsymbol{\Sigma}_\beta) \propto$$

$$MVt_\nu(\boldsymbol{y};\mu,\boldsymbol{\Sigma}) =$$

$i = 1 \ldots n$

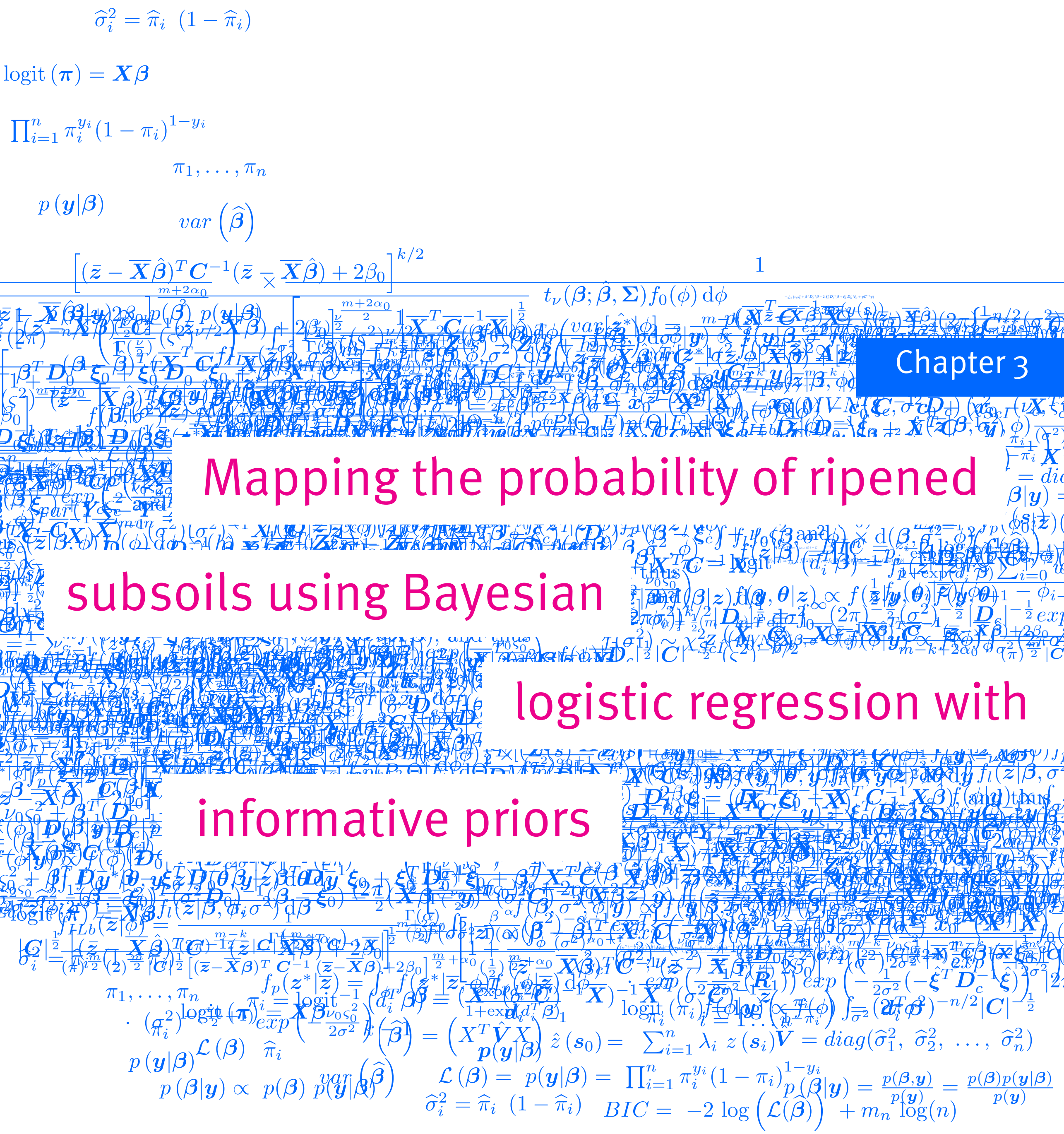
$$= (2\pi)^{-\frac{k}{2}}\frac{\left(\frac{\nu_c}{2}\right)^{\frac{\nu_c}{2}}}{\Gamma\left(\frac{\nu_c}{2}\right)}|\boldsymbol{D}_c|^{-\frac{1}{2}}$$

$$\cdot\ |2\pi\sigma^2\boldsymbol{D}_c|^{\frac{1}{2}}\int_{\boldsymbol{\beta}}|2\pi\sigma^2\boldsymbol{D}_c|$$

$$\propto \int_\phi$$

$$\text{var}\left(\widehat{\boldsymbol{\beta}}\right) = \left(X^T\hat{\boldsymbol{V}}X\right)$$

$$\hat{\boldsymbol{V}} = diag($$

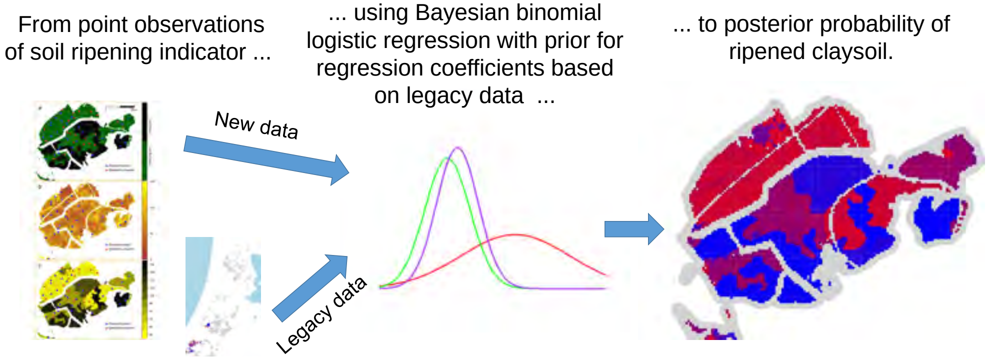


Chapter 3

Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors

Based on:
Luc Steinbuch, Dick J Brus, Gerard BM Heuvelink, 2018. Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors. *Geoderma*, **v316**, pp.56-69.

Abstract One of the first soil forming processes in marine and fluviatile clay soils is ripening, the irreversible change of physical and chemical soil properties, especially consistency, under influence of air. We used Bayesian binomial logistic regression (BBLR) to update the map showing unripened subsoils for a reclamation area in the west of The Netherlands. Similar to conventional binomial logistic regression (BLR), in BBLR the binary target variable (the subsoil is ripened or unripened) is modelled by a Bernoulli distribution. The logit transform of the 'probability of success' parameter of the Bernoulli distribution was modelled as a linear combination of the covariates soil type, freeboard (the desired water level in the ditches, compared to surface level) and mean lowest groundwater table. To capture all available information, Bayesian statistics combines legacy data summarized in a 'prior' probability distribution for the regression coefficients with actual observations. Our research focused on quantifying the influence of priors with different information levels, in combination with different sample sizes, on the resulting parameters and maps. We combined subsamples of different size (ranging from 5% to 50% of the original dataset of 676 observations) with priors representing different levels of trust in legacy data and investigated the effect of sample size and prior distribution on map accuracy. The resulting posterior parameter distributions, calculated by Markov chain Monte Carlo simulation, vary in centrality as well as in dispersion, especially for the smaller datasets. More informative priors decreased dispersion and pushed posterior central values towards prior central values. Interestingly, the resulting probability maps were almost similar. However, the associated uncertainty maps were different: a more informative prior decreased prediction uncertainty. Based on the 'overall accuracy' validation metric we found – for this case specific – an optimal value for the prior information level: the standard deviation of the legacy data regression parameters should be multiplied by 10. This effect is only detectable for smaller datasets. The Area Under Curve validation statistic did not provide a meaningful optimal multiplier for the standard deviation. Bayesian binomial logistic regression proved to be a flexible mapping tool but the accuracy gain compared to conventional logistic regression was marginal and may not outweigh the extra modelling and computing effort.



From point observations of soil ripening indicator ...

... using Bayesian binomial logistic regression with prior for regression coefficients based on legacy data ...

... to posterior probability of ripened claysoil.

Abbreviations

AUC: Area under curve; BBLR: Bayesian binomial logistic regression; BIC: Bayesian Information Criterion; BLR: Binomial logistic regression; FPR: False Positive Rate; GLM: Generalized Linear Model; ML: Maximum likelihood; MLE: Maximum likelihood estimator; MLR: Multinomial logistic regression; MLW: Mean lowest ground water table; ROC: Receiver Operating Characteristics; TPR: True Positive Rate; UMF: Uncertainty multiplication factor.

3.1 Introduction

One of the first soil forming processes in marine and fluviatile clay soils is ripening, which is the irreversible change of physical and chemical soil properties, such as consistency, under influence of air. The ripening stage is an important factor in determining land use suitability. Moreover, it is also an indicator for forecasting soil shrinkage (Pons and Zonneveld, 1965). In the central western part of The Netherlands, clay soils have been waterlogged almost since deposition, and part of these soils are thus still ripening. The ripening process is ongoing, and as a result the current maps, created between 1960 and 1995, are getting outdated. These maps must be updated to accurately represent the current situation.

Soil ripening is mapped as a binary property, i.e. on each location, the soil is considered either ‘ripened’ or ‘unripened’. It is unripened if any part of the profile (0-80cm) contains unripened clay. If point observations of soil ripening and maps of covariates related to soil ripening are available, a map of the probability of a ripened subsoil can be obtained by binomial logistic regression (BLR). In BLR, the logit transform of the ‘probability of success’ parameter of the Bernoulli distribution which represents in our case the probability that the soil is ripened, is modelled as a linear combination of covariates. With more than two classes the data follow a multinomial distribution and a similar approach, multinomial logistic regression (MLR), can be applied to map class probabilities. Kempen et al. (2009) and Vasques et al. (2014) applied MLR to map probabilities of multiple soil classes. Collard et al. (2014) compared MLR with classification trees and random forests. They found that MLR performed remarkably well for predicting soil classes. In contrast, Heung et al. (2016) showed MLR to perform worse for predicting soil classes in a comparison of ten machine learning approaches (e.g. logistic model trees, artificial neural networks).

BLR and MLR only use the observations of the variable of interest at the sampling points and the maps of the covariates. Models might better reflect reality and give more accurate predictions if we were able to exploit all available information in the model calibration process, especially in situations with scarce data. In particular, we may think of ‘prior’ knowledge about the regression coefficients of the BLR (MLR) model, which is not used in BLR (MLR) calibration. Bayesian statistics is equipped to capture all available knowledge by combining multiple information streams, i.e. information summarized in a ‘prior’ probability distribution of the model parameters, and information contained in the actual observations. For instance, Stanaway et al. (2011) used knowledge of plant properties and observation accuracy in Bayesian mapping of the risk of invasive plant species in Australia. Frigessi and Stander (1994) used deterministic terrain data to support Bayesian classification of satellite spectral images. Truong et al. (2014) used expert guesses of point-support variogram parameters to support Bayesian area-to-point kriging for remotely sensed air temperature.

To our best knowledge, Bayesian logistic regression has not yet been used to create soil property maps. In this research, we extensively explain, and apply, Bayesian binomial logistic regression (BBLR) for mapping clay soil ripening probability. In particular, we assess the added value of incorporating prior information derived from case-related legacy data. We investigate the added value of prior information with different degrees of information level in combination with different sample sizes of recent soil ripening data. Furthermore, this work includes a brief explanation of Bayesian generalized regression, the Metropolis algorithm and the validation statistics ‘overall accuracy’ and ‘area under curve’, with the purpose to familiarize soil scientists with these concepts.

3.2 Theory

3.2.1 The binomial logistic regression (BLR) model

Binary responses on discrete or continuous covariates can be modelled with the binomial logistic regression (BLR) model, which is an instance of the Generalized Linear Models family.

Let $y_i,\ i = 1\ldots n$ be observations of a binary target variable, where each y_i equals 1 or 0 and n is the number of sampling locations. In BLR, the data are modelled as independent draws from a Bernoulli probability distribution:

$$y_i \sim \text{Bernoulli}\,(\pi_i) \tag{3.1}$$

with π_i the ‘probability of success’ parameter at the i-th sampling location. The logit transform of π_i is modelled by a linear combination of covariates:

$$\text{logit}\,(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{d}_i^T\boldsymbol{\beta} \tag{3.2}$$

where $\boldsymbol{d}_i$ is an k-size vector, the first element of which equals 1 and the remaining elements of which contain the values of $k-1$ covariates at the i-th sampling location, and $\boldsymbol{\beta}$ is a vector of regression coefficients, including an intercept term. The inverse logit is written as:

$$\pi_i = \text{logit}^{-1}\left(d_i^T\boldsymbol{\beta}\right) = \frac{\exp(d_i^T\boldsymbol{\beta})}{1+\exp(d_i^T\boldsymbol{\beta})}. \tag{3.3}$$

For all locations together, Eqn. 3.2 can be written as Eqn. 3.4 with $\boldsymbol{\pi}$ a column vector of $\pi_1,\ldots,\pi_n$ and $\boldsymbol{X}$ the design matrix, which contains the k covariates at the n sampling locations, including a column of leading ones:

$$\text{logit}\,(\boldsymbol{\pi}) = \boldsymbol{X}\boldsymbol{\beta}. \tag{3.4}$$

Having described $\boldsymbol{\pi}$ as a function of a vector of regression parameters $\boldsymbol{\beta}$, we obtain an estimate of $\boldsymbol{\beta}$ that fits the data best, and use this calibrated BLR model for estimating the probability of a ripened subsoil at new locations. Note that we assume that the regression residuals are independent. In other words, we assume that the spatial structure in parameter $\boldsymbol{\pi}$ is fully captured by the covariates.

3.2.1.1 Estimation of regression parameters using maximum likelihood

Likelihood is a central concept in statistical model calibration, selection and comparison. In the scope of this paper, the likelihood $\mathcal{L}(\boldsymbol{\beta})$ equals the probability of the observations $\boldsymbol{y}$ as a function of the regression coefficients vector $\boldsymbol{\beta}$, given in Eqn. 3.5 (Collet, 1991):

$$\mathcal{L}(\boldsymbol{\beta}) = p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{3.5}$$

Note that parameter π_i is a function of $\boldsymbol{\beta}$ as given in Eqn. 3.3. Note also that $\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\beta})$ is a proper probability distribution when considered a function of y, i.e. it integrates to a finite value (actually to one) over all possible values for $\boldsymbol{y}$, but it is a likelihood when considered a function of $\boldsymbol{\beta}$.

We calibrate a given model structure, i.e. a model with a given combination of covariates, on the data by finding the estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ that maximises the likelihood. Analytical solutions are not always available and numerical, iterative search algorithms are used instead (Collet, 1991). The uncertainty in $\widehat{\boldsymbol{\beta}}$ is expressed by its variance-covariance matrix:

$$\mathrm{var}\left(\widehat{\boldsymbol{\beta}}\right) = \left(X^T \hat{\boldsymbol{V}} X\right)^{-1} \tag{3.6}$$

with $\hat{\boldsymbol{V}} = diag(\widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \ldots, \widehat{\sigma}_n^2)$, where $\widehat{\sigma}_i^2 = \widehat{\pi}_i\ (1-\widehat{\pi}_i)$ and $\widehat{\pi}_i$ the estimate for π_i resulting from plugging in $\widehat{\boldsymbol{\beta}}$ in Eqn. 3.4. The diagonal of $var\left(\widehat{\boldsymbol{\beta}}\right)$ contains the squared standard errors, i.e. the modelling variance of $\widehat{\boldsymbol{\beta}}$.

3.2.1.2 Estimation of probability of ripened subsoil at new locations

Point estimates for the model parameter $\widehat{\boldsymbol{\pi}}$ at a new location can be obtained by substituting $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ and $\boldsymbol{x}_0$ for $\boldsymbol{X}$ in Eqn. 3.3, with $\boldsymbol{x}_0$ the covariate values at the new location. The modelling uncertainty in $\widehat{\boldsymbol{\pi}}$ at a new location as result of uncertainty in $\widehat{\boldsymbol{\beta}}$ can be investigated by the Monte Carlo method: simulate a large number of independent vectors with regression coefficients (using $\widehat{\boldsymbol{\beta}}$, $var(\widehat{\boldsymbol{\beta}})$ and a pseudo-random number generator, while assuming $\widehat{\boldsymbol{\beta}}$ has a multivariate normal distribution) and calculate the corresponding $\pi_0^{(j)}$ using Eqn. 3.4, where (j) indicates the iteration number, $(j) = 1,\ 2..r$ of r iterations. The resulting empirical distribution at the new location can be visualised by a histogram of all simulated $\pi_0^{(j)}$.

3.2.1.3 Selecting model structure

The regression model structure, i.e. the combination of covariates, may be chosen by minimizing the Bayesian Information Criterion (BIC; Neath and Cavanaugh, 2012). Model selection criteria such as BIC favour models that explain the data well – quantified with a high maximum likelihood – but penalizes for model complexity, expressed as the number of model parameters m_n, which equals the number of covariates plus interactions if included:

$$BIC = \ -2\log\left(\mathcal{L}(\widehat{\boldsymbol{\beta}})\right)\ + m_n\ \log(n) \tag{3.7}$$

3.2.2 The Bayesian binomial logistic regression (BBLR) model

3.2.2.1 Bayesian statistics

Using notation introduced in the previous section, Bayes' theorem follows from the definitions of conditional and joint probabilities, and states that

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})\, p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})} \tag{3.8}$$

Within the Bayesian framework, $\boldsymbol{\beta}$ can represent any collection of parameters – or even a more abstract concept such as 'a hypothesis'. These parameters are implicitly part of a statistical model definition. Observations $\boldsymbol{y}$ can also be named 'evidence'. Using this terminology, $p(\boldsymbol{\beta}, \boldsymbol{y})$ is the joint probability function of the parameters and the evidence, while $p(\boldsymbol{y})$ is the probability of the evidence itself. In the third part of the equation, $p(\boldsymbol{\beta})$ is the prior probability or 'prior belief', the probability function assigned to the parameters, independent of the evidence. The conditional probability $p(\boldsymbol{y}|\boldsymbol{\beta})$, already introduced in the previous section, is the probability of the evidence given (also called 'conditional on') the parameters. The resulting $p(\boldsymbol{\beta}|\boldsymbol{y})$ is called the posterior probability, the measure of our belief in the parameters given the evidence (Bernardo and Smith, 2009; Pruim, 2011; Gelman et al., 2013).

Evaluation of Eqn. 3.8 is usually complicated because the evidence probability $p(\boldsymbol{y})$ is unknown. One solution, explained below, is to make use of Markov chain Monte Carlo simulation. This only requires the ratio of different posteriors, all conditional on the same, fixed evidence but with different parameters. Markov chain Monte Carlo simulation makes use of the fact that, given fixed evidence $\boldsymbol{y}$, the posterior of $\boldsymbol{\beta}$ is proportional to the product of the prior and likelihood (Gelman et al., 2013):

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \; p(\boldsymbol{\beta})\; p(\boldsymbol{y}|\boldsymbol{\beta}) \tag{3.9}$$

The prior belief can range from no belief to very strong belief, in other words from a non-informative prior (although some authors reflect that a non-informative prior does not exist, because a prior always contains some information) to a very informative prior. Note that in the Bayesian model building process, it is justified to adjust a prior after data collection so that the posterior makes sense (Bernardo and Smith, 2009; Gelman et al., 2013). However, a prior should not repeat information that is already captured by the likelihood; all information sources can be used only once.

3.2.2.2 Sampling the posterior with Markov chain Monte Carlo

As indicated above, we use a numerical simulation algorithm, the Metropolis algorithm, to sample from the BBLR posterior distribution (as illustrated in Fig. 3.1), rather than finding an analytical solution to the posterior by applying a so-called 'conjugate prior'. This latter option is shortly discussed in Section 3.4.4.2. The Metropolis algorithm generates a Markov chain of model parameter vectors (i.e., regression coefficients). A Markov chain has the property that given the current state, the next and previous state are independent; the current state depends only on the previous state. Simulating values from a distribution by generating a Markov chain is referred to as Markov chain Monte Carlo simulation (MCMC). The Metropolis algorithm is one specific example of MCMC. For more information about MCMC see e.g. Chib and Greenberg (1995), Christensen et al. (2010), and Gelman et al. (2013).

An important indicator of the proper functioning of the Metropolis algorithm is the acceptance rate, i.e. the number of accepted proposals divided by the total number of proposals. For a sufficient and efficient exploration of the parameter space, this number should be between 0.25 and 0.5 (Rosenthal, 2011). The acceptance rate is largely controlled by the size of the jumps in parameter space. The jump from $\boldsymbol{\beta}^{(j)}$ to $\boldsymbol{\beta}^{*}$ is generated by the proposal probability distribution (also called jumping distribution) using $\boldsymbol{\beta}^{(j)}$ as input. We tune this proposal distribution by manually setting a tuning parameter, often a variance scaling factor, to attain the desired acceptance rate. Furthermore, the first set of $\boldsymbol{\beta}^{(j)}$'s, the warm-up phase (also called 'burn-in'), that is highly influenced by the starting point, should be removed. The consecutive $\boldsymbol{\beta}^{(j)}$'s of the remaining chain are correlated. To speed up subsequent computations and save storage space, 'thinning' is applied by subsampling the Markov chain systematically. Chain convergence is checked by comparing different parts of the chain, by comparing different chains (for example chains with different settings), or by a combination of both (Brooks et al., 2011; Gelman et al., 2013). If the different parts or different chains produce similar posterior distributions then this indicates that the parameter space has been explored sufficiently.

A (thinned) Markov chain of regression coefficient vectors forms a random sample from the joint posterior distribution of those coefficients. We can easily obtain the marginal density for individual coefficients by considering the sample of only that coefficient (Fig. 3.1c).

The probability of success parameter $\boldsymbol{\pi}$ is computed from the regression coefficients in the same way as described for BLR in Section 3.2.1.2, but note that this provides a sample from the posterior distribution of $\boldsymbol{\pi}$. Typically, the mean of this sample is used to summarise information.

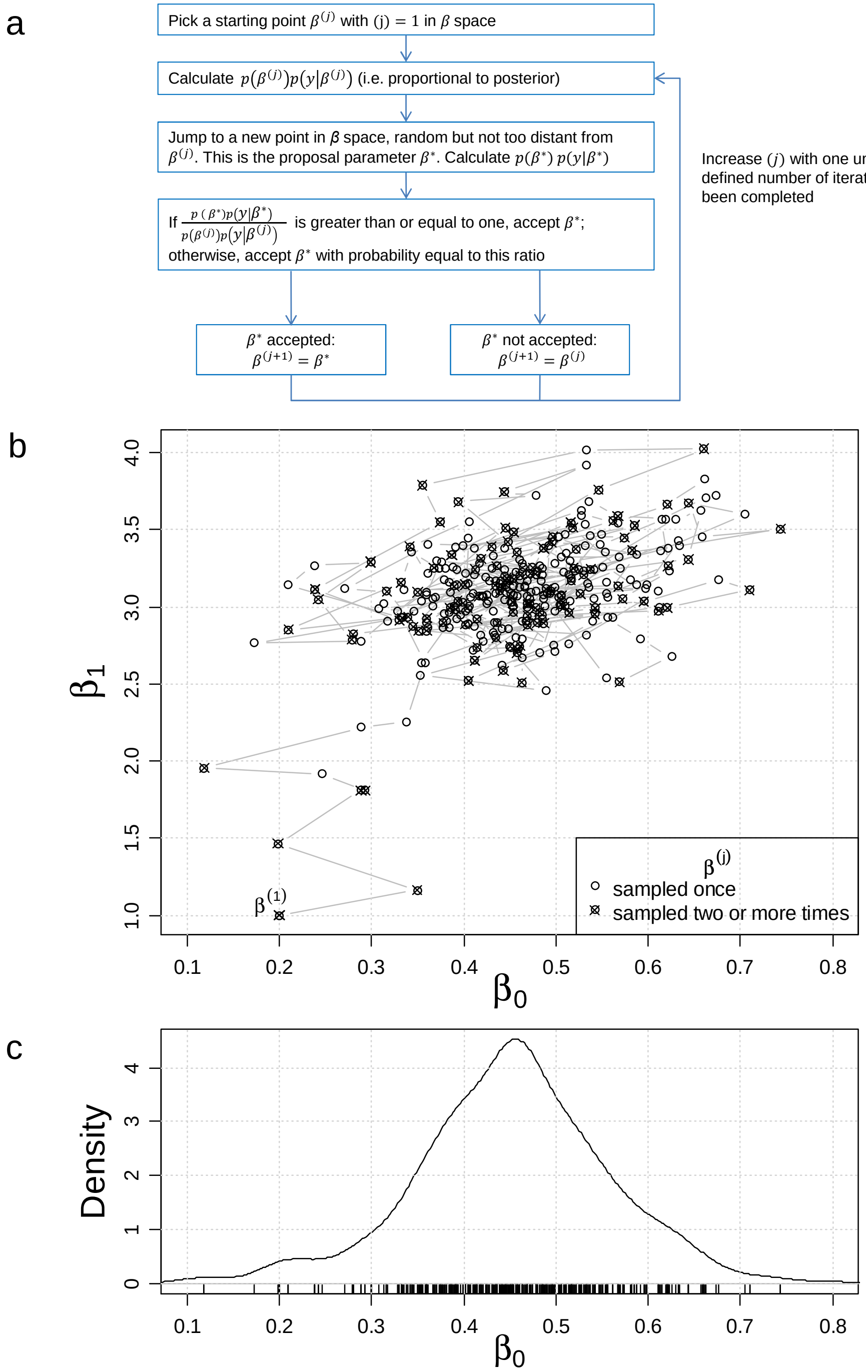
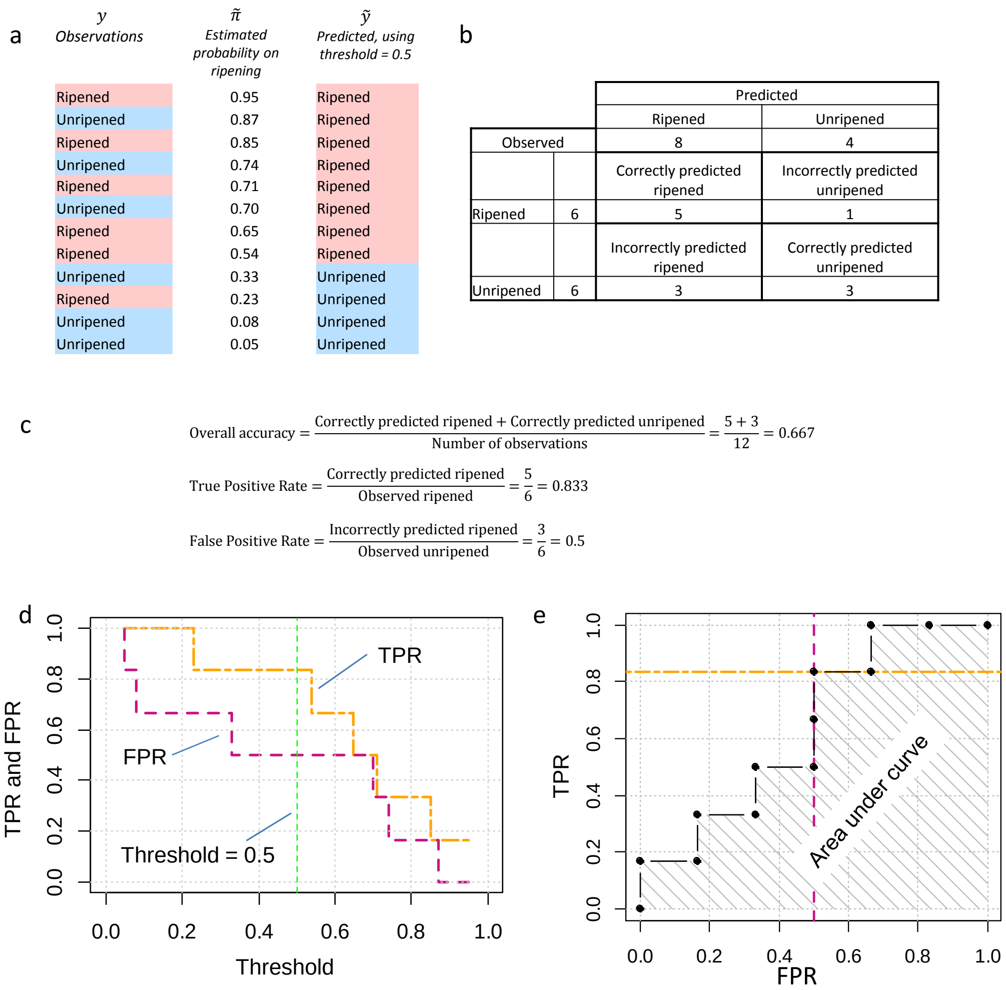


Figure 3.1: Flowchart of the Metropolis algorithm (a) and a simplified example of a resulting Markov chain in two-dimensional parameter space (b). After starting at $\beta^{(1)}$= [0.2, 1.0], the Markov chain gradually jumps to the area where the joint posterior probability density of β is concentrated – which appears to be around [0.45, 3.25] – and starts exploring it. In this example, we show the first 250 iterations. From the Markov chain, we can extract the marginal probability density of each model parameter; this is shown for β_0 (c). Note that we did not apply warm-up removal nor thinning before calculating and plotting the marginal density.

3.2.3 Validation of estimated probability by overall accuracy and area under curve

We validated the BLR and BBLR models by comparing the estimated probabilities of ripened subsoil at validation locations (continuous from zero to one) with the observations (binary: zero or one). Note that the probabilities themselves cannot be validated because these are only model constructs and are not observable. One possibility to deal with this is to set a probability threshold value, for example 0.5, that transforms the probabilities into zeroes and ones, and calculate the confusion matrix from comparison of observed and predicted ripened soil. This is illustrated in Figs 3.2a and 3.2b. From the confusion matrix, various metrics can be extracted, such as map unit purities, class representations and, as applied in this research, the overall accuracy (Fig. 3.2c; Brus et al., 2011).

The choice of the probability threshold value is often arbitrary. For example, if a prediction that is wrongfully classified ‘ripened’ has more serious consequences and should be penalized more than wrongfully classifying a soil as ‘unripened’, it may be useful to increase the threshold, so that a smaller part of the area is wrongfully classified as ripened (at the expense of increasing the area that is wrongfully classified as unripened). The Receiver Operating Characteristics (ROC) quantifies the estimation performance without the need for setting a threshold (Fig. 3.2de). ROC calculates the ‘True Positive Rate’ (the proportion of observed ripened that is correctly predicted as ripened) as well as the ‘False Positive Rate’ (the proportion of observed ripened that is incorrectly predicted as ripened) for all thresholds. The resulting graph is summarized in one performance statistic, the area under curve (AUC) (Fawcett, 2006). Ideally, AUC equals one, while an AUC of 0.5 indicates no predictive value at all.



y *Observations*	$\tilde{\pi}$ *Estimated probability on ripening*	$\tilde{y}$ *Predicted, using threshold = 0.5*
Ripened	0.95	Ripened
Unripened	0.87	Ripened
Ripened	0.85	Ripened
Unripened	0.74	Ripened
Ripened	0.71	Ripened
Unripened	0.70	Ripened
Ripened	0.65	Ripened
Ripened	0.54	Ripened
Unripened	0.33	Unripened
Ripened	0.23	Unripened
Unripened	0.08	Unripened
Unripened	0.05	Unripened

		Predicted	
		Ripened	Unripened
Observed		8	4
		Correctly predicted ripened	Incorrectly predicted unripened
Ripened	6	5	1
		Incorrectly predicted ripened	Correctly predicted unripened
Unripened	6	3	3

$$\text{Overall accuracy} = \frac{\text{Correctly predicted ripened} + \text{Correctly predicted unripened}}{\text{Number of observations}} = \frac{5+3}{12} = 0.667$$

$$\text{True Positive Rate} = \frac{\text{Correctly predicted ripened}}{\text{Observed ripened}} = \frac{5}{6} = 0.833$$

$$\text{False Positive Rate} = \frac{\text{Incorrectly predicted ripened}}{\text{Observed unripened}} = \frac{3}{6} = 0.5$$

Figure 3.2: Illustration of the ‘overall accuracy’ and ‘area under curve’ (AUC) statistics. Starting from a fictitious dataset of observations y, corresponding estimates $\tilde{\pi}$ and an arbitrary probability threshold (in the example taken as 0.5), predictions of the ripening indicator $\tilde{y}$ are obtained. By comparing y and $\tilde{y}$, the confusion matrix is constructed (b), and the overall accuracy is calculated, as well as the True Positive Rate (TPR) and False Positive Rate (FPR) (c). For AUC, TPR and FPR are calculated for all probability thresholds ranging from 0 to 1 (d). Next, TPR is plotted against FPR, resulting in the Receiver Operating Characteristic (ROC) (e); the TPR and FPR according to threshold=0.5 are shown. Finally, AUC summarizes the ROC into a single number. A good separating performance would show a large TPR while FPR is small, for all thresholds. In such case the ROC curve is steep at first and passes close by the upper left corner, resulting in an AUC close to 1. In this example, AUC equals 0.639.

3.3 Case study

3.3.1 Study area

The marine clay soils of interest in this study are situated in the western and north-western part of The Netherlands (Fig. 3.3a). The clay material was deposited between 6,000 and 1,000 BC, during a period of increased sea level (Jongmans et al., 2013). In a later stage organic material accumulated under wet conditions and the land became covered with peat. Then, because of water erosion and peat excavation, lakes emerged. Reclamation of the lakes started around 1500 AD and continued until at least 1958. Until reclamation, the soils were continuously submerged. The groundwater tables in the reclaimed clay soils commonly are shallow, and, as a result, locally these soils are still in the ripening stage (Stichting voor Bodemkartering Wageningen, 1969). This study is limited to the clay soils in these former lakes[1] classified as unripened in the previous mapping survey, which was gradually performed in the timespan 1960-1995 – but mainly between 1960 and 1980. The soils of interest cover 211km^2 and are highly fragmented in a rectangular area of 60 by 100 km (Fig. 3.3a). Land use consists mainly of agricultural permanent grasslands, and to a lesser extent arable farming and nature. In this densely populated area there are also many buildings and roads (Stichting voor Bodemkartering Wageningen, 1969; de Vries et al., 2017). The climate can be described as maritime with a warm summer, without any regular dry period (Peel et al., 2007).

3.3.2 Data

3.3.2.1 Observations

To determine the current ripening stage, 676 sampling locations were selected by spatial coverage sampling (Brus et al., 2006), using the R package spcosa (Walvoort et al., 2010). Sampling was done, in 2016, using an Edelman or a gouge auger. The soil profiles, up to at least 150 cm depth, were systematically described in the field, and later classified into the new data observations ‘ripened’ or ‘not-ripened’, considering the upper 80 cm soil layer (Fig. 3.3b) (de Vries et al., 2017). The sample fraction with ripened soils was 56%. Originating from earlier surveys, legacy data on the ripening stage of 1319 soil profiles sampled in the timespan 1985-2005 are shown in Fig. 3.3c. Here, 61% of the sample had ripened soils.

3.3.2.2 Subsamples of new data

To determine the effect of sample size on the posterior distributions and probability maps when using BBLR, subsamples from the complete dataset were selected. This was done by clustering the 676 sampling locations into clusters of equal size by k-means clustering (Hartigan and Wong, 1979), using the spatial coordinates of the sampling locations as clustering attributes. We computed clusters of size 2, 4, 10 and 20. By selecting randomly one point per cluster, two subsamples of size 676/2, four subsamples

[1] In Dutch: “Droogmakerijen”

of 676/4 et cetera were obtained. Each subsample was used for calibration. The observations not included in a calibration subsample were used for validation. An overview of the resulting subsamples is provided in Table 3.1.

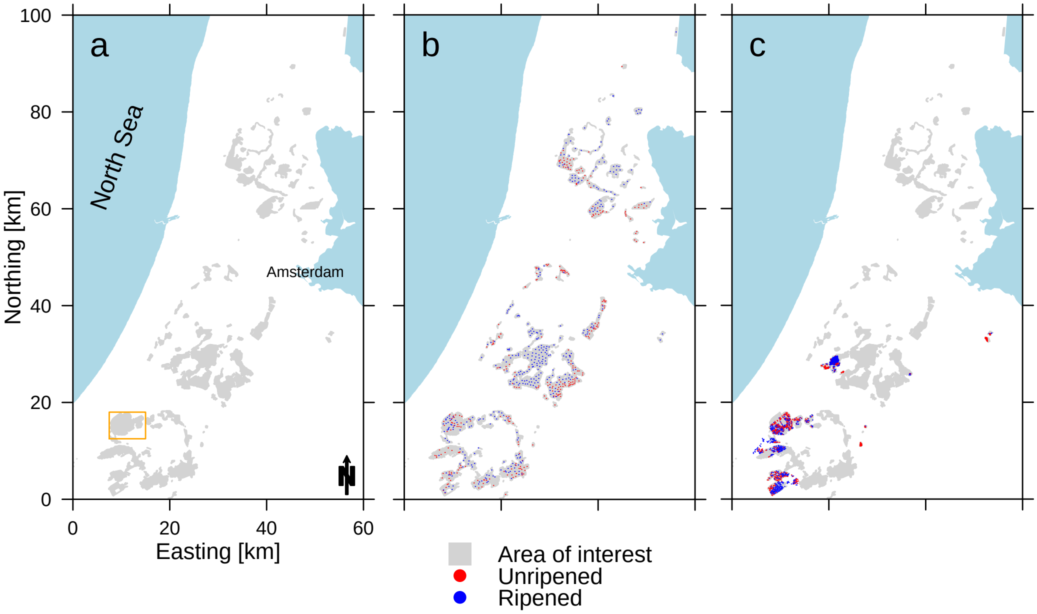


Figure 3.3: The area of interest in the central-western part of The Netherlands, mapped as unripened subsoils on the current (but outdated) Soil Map of the Netherlands 1:50,000 (a). Recently collected data on subsoil ripening stage ('new data') (b). Data collected on soil ripening stage in the years 1985-2005 ('legacy data')(c). The orange square in (a) indicates a subarea, which we will focus on.

Table 3.1: Overview of new data subsamples obtained by k-means clustering of the original dataset of 676 observations.

	Number of subsamples	**Observation locations per subsample**	**Validation locations per subsample**
Subsamples 50%	2	338	338
Subsamples 25%	4	169	507
Subsamples 10%	10	66, 67 or 68	608, 609 or 610
Subsamples 5%	20	33 or 34	642 or 643

3.3.2.3 Covariates

The set of covariates to be proposed as regressors was based on expert knowledge, i.e. we selected covariates that, based on pedological knowledge, could influence the ripening process. As our aim was to construct a map of the probability of a ripened subsoil, only covariates of which a map was available were considered. An overview of all 12 candidate covariates is given in de (de Vries et al., 2017), and summarized in Table 3.2.

Table 3.2: Candidate covariates.

Candidate covariate	Type
Distance to dike	Categorical
Elevation	Continuous
Soil type	Categorical
Freeboard	Continuous
Mean lowest water table	Categorical
Seepage and infiltration	Categorical
Land use	Categorical
Relative elevation, 100m radius	Continuous
Relative elevation, 250m radius	Continuous
Relative elevation, 500m radius	Continuous
Relative elevation, 750m radius	Continuous
Relative elevation, 1000m radius	Continuous

We did not consider interactions. We shifted and scaled the covariates according to Gelman et al. (2008), which means all means became zero and standard deviations became 0.5, so that continuous variables have the same scale as symmetric binary inputs. We used the full sample of new observations to select a subset of covariates used as regressors in the models. This subset of covariates was used in all models calibrated on the subsamples. As will be explained in Section 3.3.4.1, three covariates where finally selected: *soil type*, *freeboard* and *mean lowest water table* (MLW). Soil type is a categorical variable where two soil types are distinguished: a ‘peaty earth soils with unripened subsoil’[2], which we indicate here as ‘peaty clay soil’, and ‘clayey hydro-earth soils with unripened subsoil’[3], which we indicate shortly as ‘clayey soil’. The peaty clay soil contains an organic layer of at least 10cm, starting within 40cm from surface level, containing a substantial amount (>10%) of organic material (Bakker et al., 1989). Freeboard[4] refers to the regulated water level below surface as desired by the local water board (de Vries et al., 2017). The majority of the freeboard values are between 25 and 120 cm below surface level. MLW is explained on detail level by Hoogland et al. (2014) and de Vries et al. (2017). In our study area MLW is a discrete variable taking on 18 values, ranging from 54 to 137 cm below surface level.

[2] In Dutch: “Plaseerdgrond”

[3] In Dutch: “Tochteerdgrond”

[4] In Dutch: “Drooglegging”

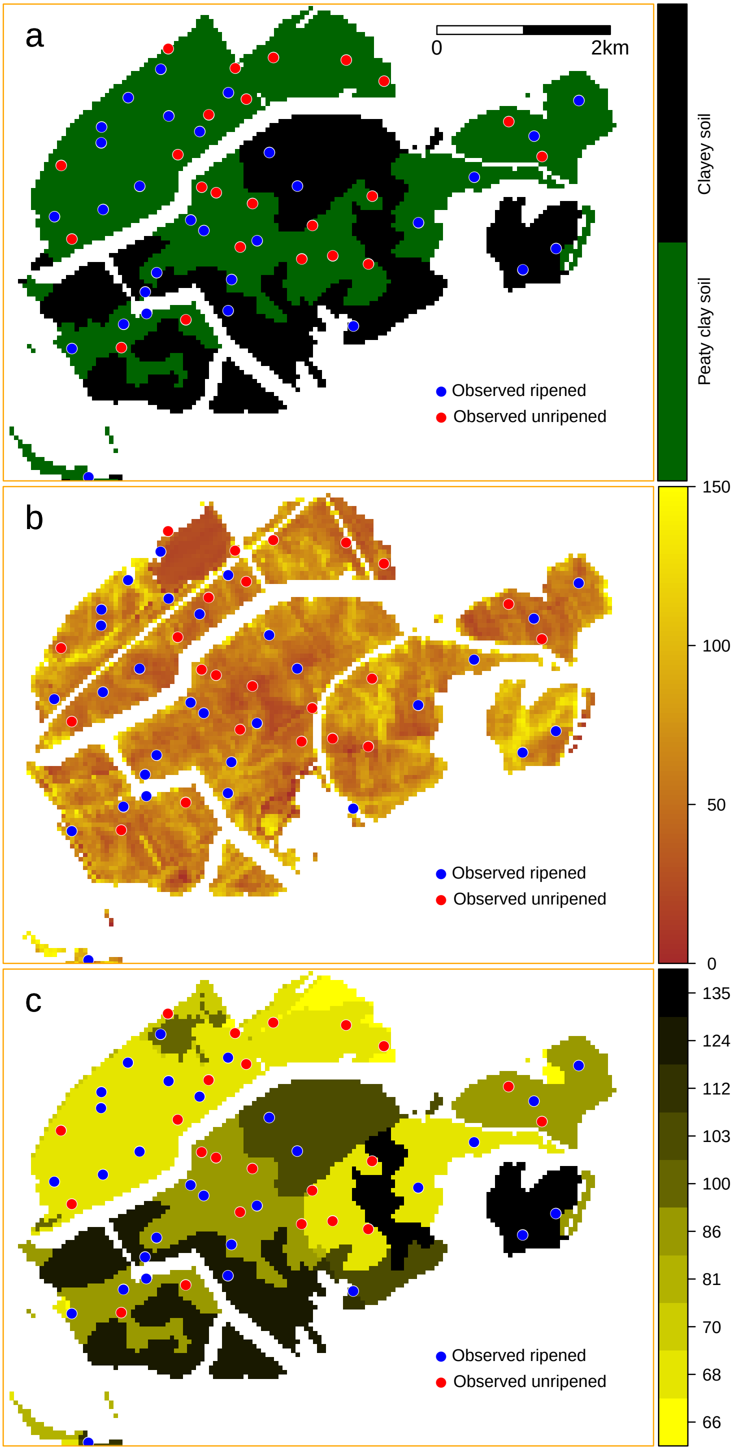


Figure 3.4: Maps of selected covariates for the subarea indicated in Fig.3.3a, including new data observations. Soil type, distinguishing two types of clay soil (a). Freeboard, in cm below surface level (b). Outliers (for example elevated roads) are removed, which explains the thin white lines not present on the maps shown in (a) and (c). Mean lowest water table (MLW) in cm under surface level (c). In this area, 10 discrete steps are present.

3.3.3 Bayesian and non-Bayesian binomial logistic regression implementation

3.3.3.1 Priors

We constructed several priors for the Bayesian models using the legacy data. We fitted a (non-Bayesian) BLR model to these data, using the selected covariates (as will be discussed in Section 3.3.4.1) as regressors. As these ripening observations were gathered in the past and are strongly spatially clustered (Fig. 3.3c), it would be unwise to simply add these to the new observations and treat them as if they were new observations. However, we consider them appropriate for constructing an informative case-related prior to be used in a Bayesian BLR.

We quantified our trust in the BLR model calibrated with legacy data by multiplying the variance-covariance matrix of the regression coefficients with an uncertainty multiplication factor (UMF). We only considered UMFs greater than one because the legacy data are less informative about current subsoil ripening than the new data. Note that UMF is a multiplier for the variance, thus the standard deviations are multiplied by the square root of UMF. We used a range of values for UMF and evaluated which UMF value gave the best results through validation. For further comparison, we added a data-unrelated prior, being a multivariate normal distribution with zero means and covariances and an extremely large value (10E20) for all variances. In the context of BBLR, this prior can be considered extremely low-informative. We also applied non-Bayesian binomial logistic regression using the new data only, and non-Bayesian binomial logistic regression with coefficients derived from the legacy data. An overview of all statistical models is given in Table 3.3.

Table 3.3: Overview of applied statistical models.

Binomial logistic regression using new data	Bayesian binomial logistic regression (BBLR)								Binomial logistic regression using legacy data
	Data-unrelated, non-informative prior	Prior based on old data with uncertainty multiplication factor (UMF)							
BLR	Non_inf	10000	1000	200	100	50	10	5	BLR_on_legacy

All statistical models were calibrated on all 36 subsamples presented in Table 3.1, giving a total of 360 calibrated models. Recall that all 360 models used the same subset of covariates as regressors.

3.3.3.2 MCMC settings and computation costs

For the MCMC Metropolis algorithm, we used the same settings for all cases: the total number of iterations was 15,500 of which 500 are warm-up; the thinning factor was 15, resulting in MCMC samples of size 1,000; starting values for all elements of β were zero. The tuning parameter for the jump distribution for this specific algorithm, `MCMClogit` in the R package `MCMCpack` (Martin et al., 2011) was 0.9, resulting in an acceptance rate between 0.27 and 0.47 for all statistical models, except for two subsamples with quasi-complete separation, i.e. subsamples for which the model (almost) perfectly predicts the observations. An example would be in Fig. 3.4a (if we would have sampled only this area): there is no observed unripened clayey soil, so a model with soil type as covariate perfectly predicts the observations in the clayey soil section of the map. The consequences of quasi-complete separation will be discussed later. Table A.3 in the Supplementary materials[5] shows all acceptance rates. We applied the Gelman diagnostic (Gelman et al., 2013), as implemented in the diagnostic `coda` R package (Plummer et al., 2006) to verify chain convergence. Calculating one Markov chain took between 0.5 and 4 seconds – increasing with subsample size – using a present-day office computer. Calculating one map with estimated probabilities of a ripened subsoil with a total of 81,000 grid points, based on 1,000 simulated vectors with regression coefficients, took about two minutes.

3.3.3.3 Binomial logistic regression

The non-Bayesian BLR parameters were estimated using the `glm` function, part of the base R-package `stats` (R Core Team, 2017).

[5] The supplementary materials, Appendix A., can be downloaded from the journal version of this chapter: Luc Steinbuch, Dick J Brus, Gerard BM Heuvelink, 2018. Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors. *Geoderma*, **v316**, pp.56-69; https://doi.org/10.1016/j.geoderma.2017.12.010.

3.3.4 Results

3.3.4.1 Regression model structure and BLR estimates

Using the full new dataset, we selected a model structure through stepwise selection, with BIC as model performance criterion. This resulted in a model with the covariates soil type, freeboard and MLW as predictors. For reference purposes, the maximum likelihood estimates (MLE) of β, using the full new dataset, are provided in the Supplementary materials, Table A.1. The MLE of β calibrated with the legacy data, which is used to construct the priors, is provided in Table A.2. The MLEs of β for the 36 subsamples are included in the overview in Table A.6 to Table A.9 of the Supplementary materials.

3.3.4.2 Posterior distribution of regression coefficients

We found as general trend that the probability of a ripened subsoil is larger for clayey soils than for peaty clay soils and increases with freeboard and MLW, which is in line with our knowledge of soil forming processes. This holds for all but a few of the smaller subsamples, as shown in Table A.6 to Table A.9.

Fig. 5 shows as an example of results for the MLW regression coefficient. Note that, as expected, the larger the sample size and the more informative the prior (i.e., a smaller UMF), the more peaked the posterior distribution, indicating a smaller modelling uncertainty about the regression coefficient. Also, a larger sample size results in a larger difference between the modes of the prior and the posterior distribution. This is especially true for UMF = 5. The smaller the UMF, the larger the overlap of the posterior distributions.

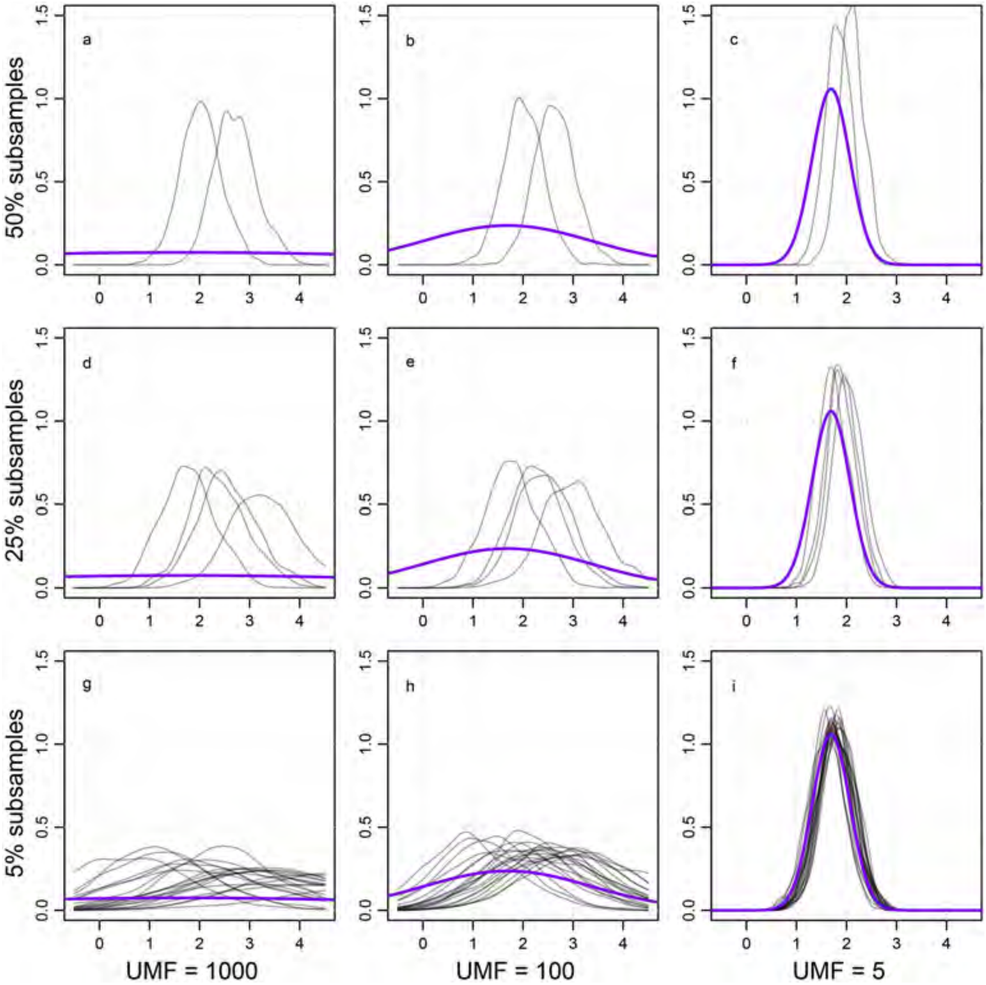


Figure 3.5: Marginal posterior probability distributions of the MLW coefficient (grey lines), for three sample sizes (rows) and three UMF values (columns). A smaller UMF value means a more informative prior. Prior distributions are in bold purple.

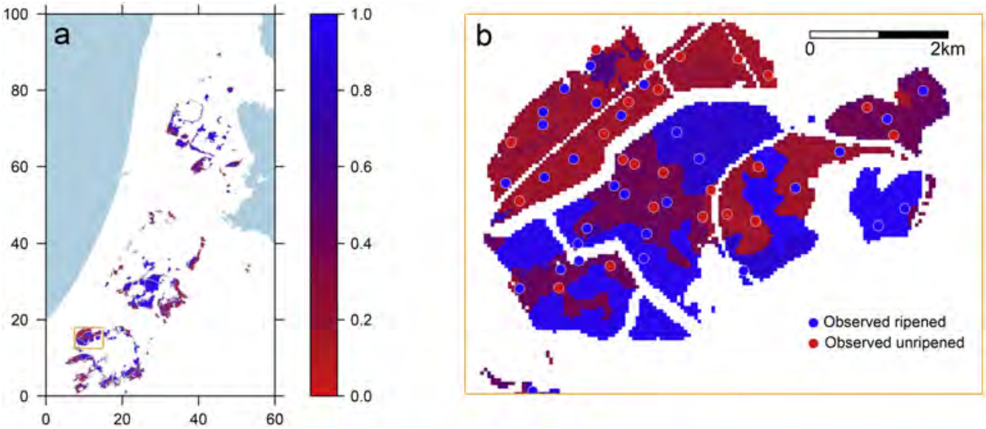


Figure 3.6: Estimated probability of ripened subsoil obtained with the entire new sample and a prior with UMF = 100 (a). Subarea with observations (b).

3.3.4.3 Probability maps of ripened subsoil

Fig. 3.6a shows a map of the estimated probability of a ripened subsoil. If we apply a probability threshold of 0.5, about 60% of the area, previously mapped as ‘unripened’, is predicted as ripened. Fig. 3.6b zooms in on the subarea. Note that in the area of Fig. 3.6b quite a few ripened subsoils are observed in areas with small estimated probabilities of a ripened subsoil (mainly in the north-west part). The sharp boundaries in the probability map are caused by sharp boundaries in the maps of the covariates soil type and MLW (Fig. 3.4a and c).

Fig. 3.7 shows probability maps for different combinations of sample size (for each sample size, one sample was randomly selected from all samples of that size) and prior distribution for the subarea. The spatial pattern is quite similar for all maps.

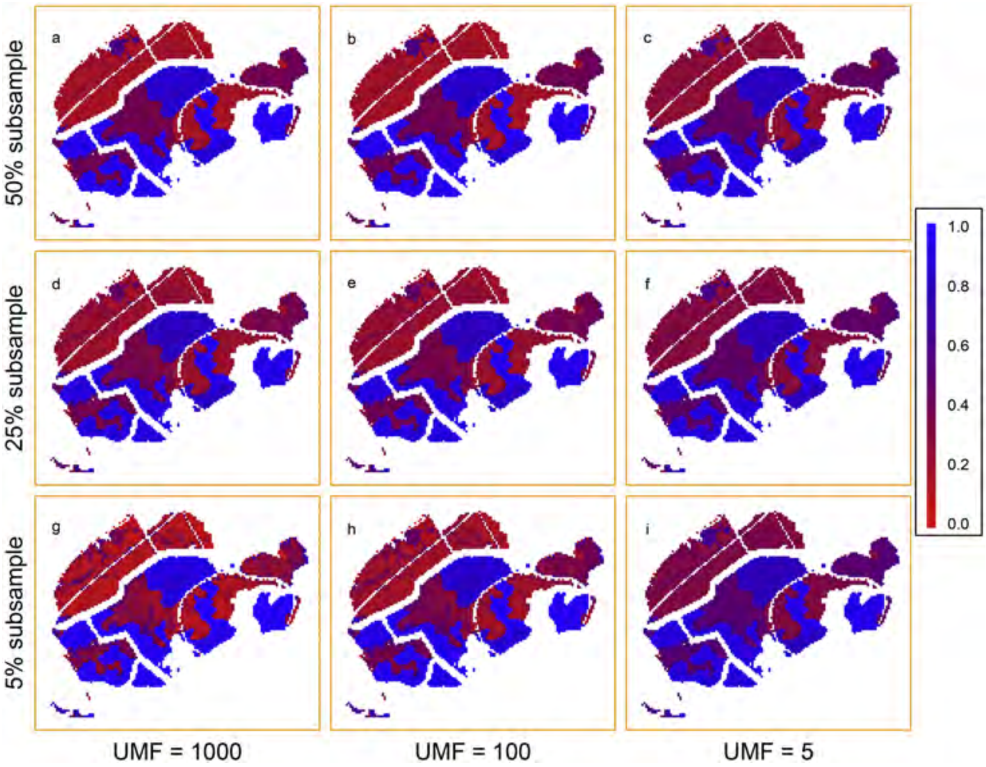


Figure 3.7: Maps of estimated probability of ripened subsoil for three subsample sizes (rows) and three UMF values (columns).

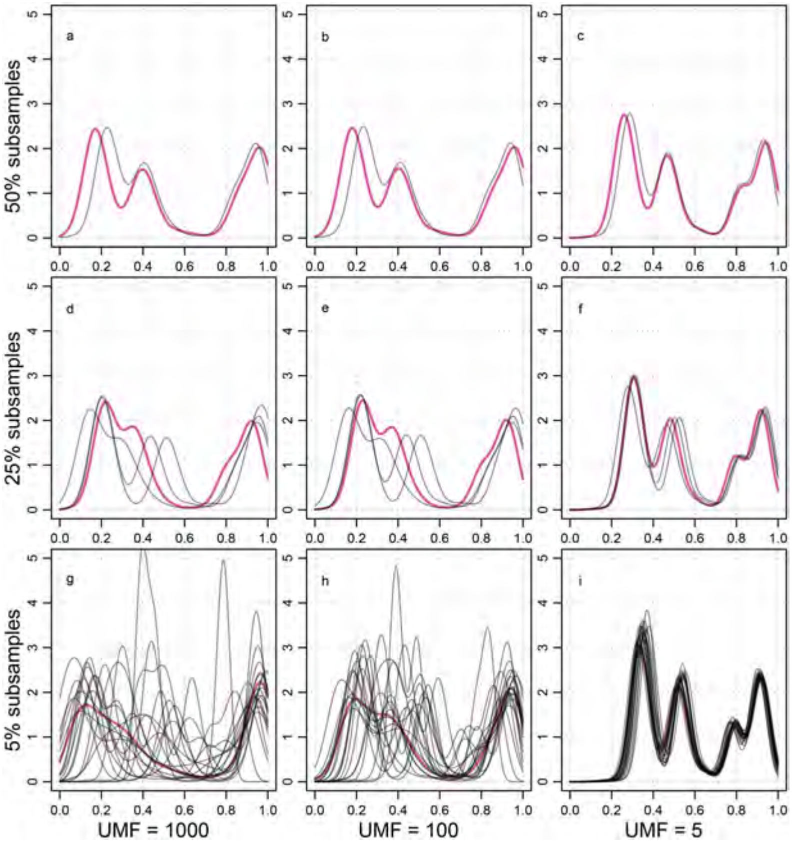


Figure 3.8: Density plots of the estimated probabilities of ripened subsoil in the subarea, for three sample sizes (50%, 25% and 5%) and three UMF values. The bold pink line corresponds with the probabilities of the maps shown in Fig. 3.7.

Fig. 3.8 shows density plots (smoothed and standardized histograms) of the estimated probabilities shown on the maps of Fig. 3.7. For comparison, density plots of all other subsamples of the same size are added. The density plots are multimodal (i.e. have several local maxima). For example, for the 50% and 25% subsamples three modes appear at probability levels near 0.2, 0.4 and 0.9. With UMF = 5 (strong informative prior) a fourth local mode at 0.8 appears, most pronounced for the 5% and 25% subsamples. For a given sample size, the densities at the modes for UMF = 5 are larger than at the corresponding modes for UMF = 1000 and 100, which means that for UMF = 5 we have larger areas with estimated probabilities near 0.2, 0.4, 0.8 and 0.9. For the 5% subsamples and large UMF (1000 and 100) there are large differences in density plots, indicating a substantial sampling uncertainty in the estimated probabilities of ripened subsoil.

Fig. 3.9 shows maps of the modelling uncertainty expressed as interquartile range (IQR), based on 1000 simulations for each of the nine shown cases (as explained in Section 3.2.1.2). The modelling uncertainties about the probability of ripened subsoil are highest for models calibrated with a small dataset and a low informative prior (such as Fig. 3.9g). The IQR decreases with increasing sample size and decreasing UMF (more informative prior), although not necessarily at every single location.

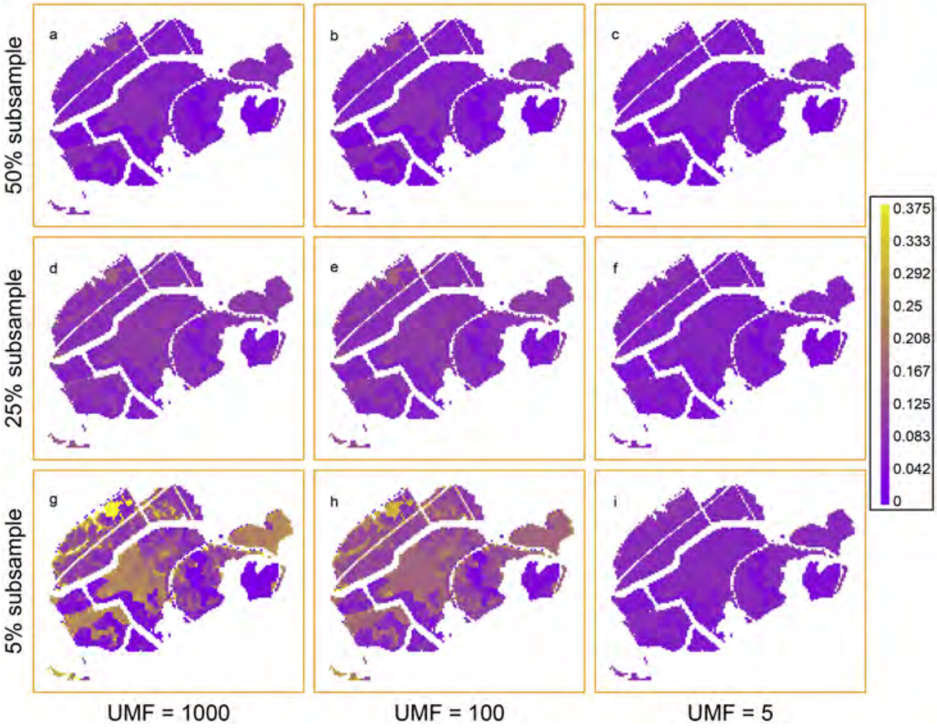


Figure 3.9: Modelling uncertainty as quantified by the interquartile range obtained with three sample sizes and three UMF values.

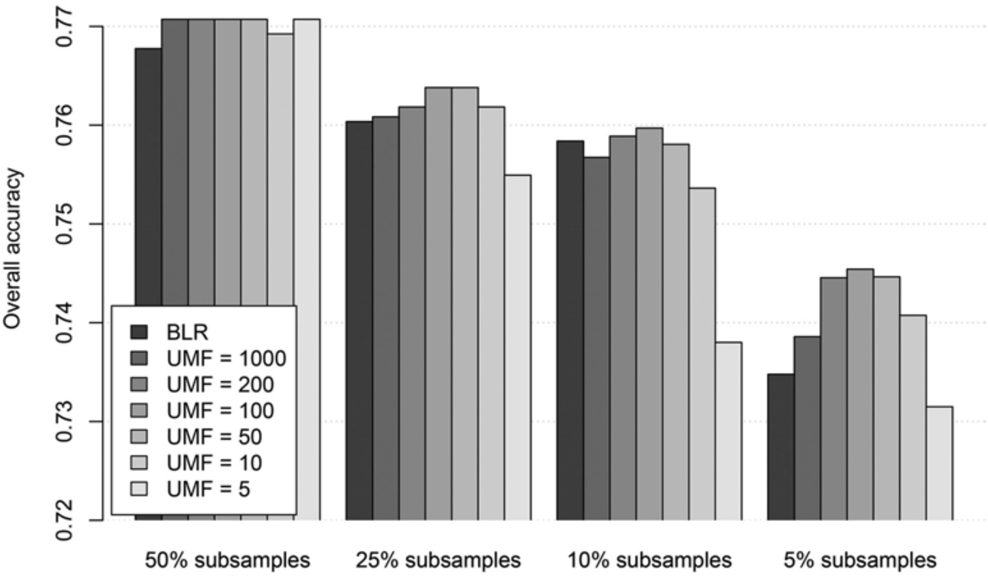


Figure 3.10: Overall accuracy for a probability threshold of 0.5, for four sample sizes and BLR + six UMF values. Note the vertical axis origin.

3.3.4.4 Validation

As explained in Section 3.2.3, we analysed the effect of increasingly informative priors on the quality of the maps by comparing the validation metrics ‘overall accuracy’ (Fig. 3.10) and ‘area under curve’ (AUC; Fig. 3.11 page 57) as obtained with BLR and with BBLR using different values for UMF. The validation statistics were computed from all subsamples of a given size. For example, for the four “subsamples 25%” (see Table 3.1), we have in total of $4 \times 507 = 2028$ validation points with an estimated probability and an observed ripening indicator. All 2028 estimation-observation pairs were used to estimate the overall accuracy and AUC for this sample size, for each statistical model. The threshold used to calculate the overall accuracy was set to 0.5. The two 5% subsamples with quasi-complete separation were not included in the computation of the validation statistics.

The overall accuracy suggests an optimal value for the UMF around 100 (Fig. 3.10 and Table A.4), independent of sample size, apart from the 50% subsamples size that show almost no dependency on the UMF. For the 25% subsample, both UMF=100 and UMF=50 seem to be optimal. Contrary to the overall accuracy, the AUC kept increasing (while levelling out) with the information level of the prior (Fig. 3.11, page 57). This even happened with unrealistically small UMF values, such as 1, 0.1 and 0.01 (results not shown) and with the legacy data based β (see Supplementary materials Table A.5). For the 50% subsamples, there is an unexpected jump in overall accuracy between the non-Bayesian BLR and all Bayesian models (Fig. 3.10), probably due to the numerical

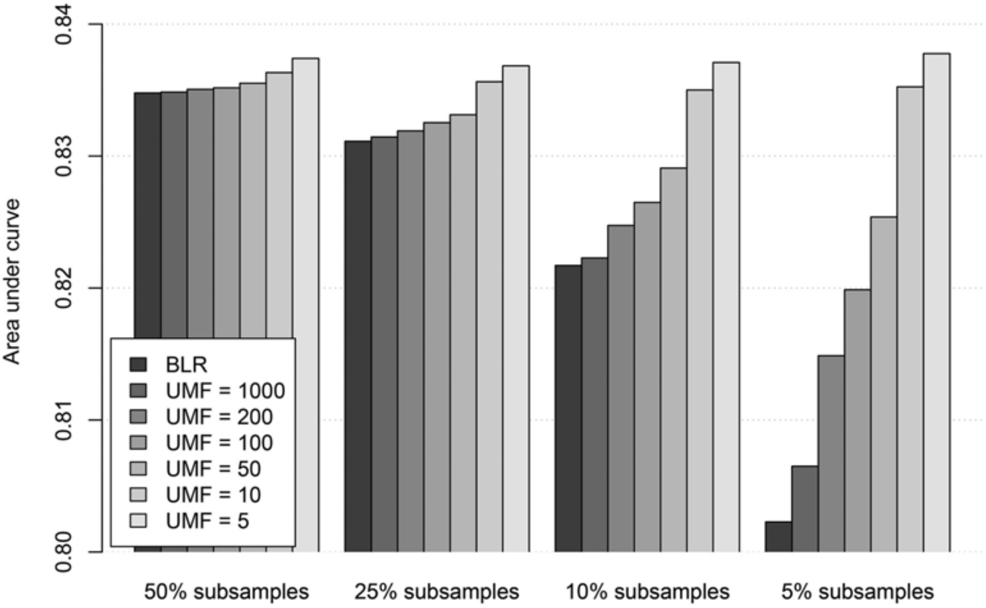


Figure 3.11: Area under curve (AUC) for four sample sizes and BLR + six UMF values. Results based on all subsamples (except the quasi-complete separation) of the given size. Note the vertical axis origin.

implementation rather than to the underlying mathematics[6]. For a quick indication of the added value of using BLR or BBLR, we also calculated the overall accuracy if we would use the sample mean of the observations as an estimate of the probability of ripened subsoil. This implies that the estimated probability is constant throughout the study area, and all subsoils are predicted either as ripened or unripened. For the different samples, overall accuracy ranged from 0.431 to 0.568, showing that there is a clear gain in overall accuracy thanks to the logistic modelling.

We also investigated the effect of UMF on the overall accuracy and AUC for each separate subsample, see Fig. A.1 and Fig. A.2 in the Supplementary materials, resp. Especially for the 5% subsamples (Fig. A.1d) the influence of a more informative prior – the validation statistics become more alike – is clear. However, there is no graphical indication of an optimal value for the UMF based on overall accuracy. With a single exception, AUC increases (and levels out) with increasingly informative priors for each individual subsample (Fig. A.2).

[6] Perhaps, by mistake the BLR model validation is based on the median of β while all other models are using its mean.

3.4 Discussion

In this research, we applied Bayesian binomial logistic regression (BBLR) for mapping the probability of ripened subsoils and the ripening indicator, and compared the added value of several priors, ranging from low-informative to high-informative. We investigated how these different priors influenced results for different numbers of new observations.

3.4.1 Mapping ripened clay soils

Fig. 3.3b and Fig. 3.6a indicate the progress of subsoil ripening in recent decades. Recall that the whole area of interest was unripened on the current (but outdated, several decades old) soil map of the Netherlands. On the updated map 60% of the area has ripened clay soils (if we apply a probability threshold of 0.5), so updating the soil map has proved relevant. It should be noted that during sampling, incidentally land owners did not permit access because the soil was too wet; therefore we might slightly have underestimated the area of unripened soils.

Bayesian BLR did not offer advantages compared to non-Bayesian BLR when using all new data: validation results (quantified as overall accuracy statistic and area under curve) were almost comparable. However, when only a small subsample of the new data was used, overall accuracy and AUC improved slightly when the prior was chosen well. In our research, the ‘uncertainty multiplication factor’ (UMF), expressing our trust in the legacy data, was optimal around 100 (Fig. 3.10), which means the standard deviation of the legacy data β should by multiplied by 10. In practice, it will be case-dependent whether or not BBLR with a prior based on legacy data can outperform BLR and if so, which UMF value yields the best results. When no validation data are available, the optimal UMF value can possibly be approximated by leave-one-out cross validation.

When comparing the results obtained with the separate samples of a given size, priors with stronger information content (smaller UMF) make the posterior β’s more similar and thus more robust against outliers in the subsample (see for example Fig. 3.5, and Supplementary results Fig. A.1 and Fig. A.2). With a very small UMF (strong informative prior), the legacy data start to dominate the new data, thus introducing a systematic difference with the β as estimated from all new data.

The observations actually consist of the ripening stage in five grades per soil horizon. These observations are used to classify each soil profile into ripened or unripened. In this study we used this binary variable. Further research into how we can make better use of the three-dimensional data on the ripening stage in five grades is welcomed.

3.4.2 AUC and overall accuracy as validation statistics

To our initial surprise, AUC kept increasing with increasing prior information level (Fig. 3.11). Lobo et al. (2008), Hanczar et al. (2010) warn to draw conclusions on the quality of the result based on AUC. Marzban (2004) discourages the use of AUC for model calibration and model fine-tuning. Being a threshold-independent metric, AUC is insensitive to changes that affect all estimated probabilities in a comparable way, for example if all estimates collectively increase. To clarify this, let us name the observed ripening indicator $\boldsymbol{y}$ and the corresponding estimated probability $\widetilde{\pi}$ together an 'estimation-observation pair'. These pairs are sorted on the estimated probability. For illustration see columns "$\boldsymbol{y}$" and "$\widetilde{\boldsymbol{\pi}}$" in Fig. 3.2a. As long as the order of the estimation-observation pairs remains the same, the ROC graph and thus AUC will not change, whatever collective changes occur in the estimated probabilities. However, if individual estimates change, the order in the estimation-observation pairs might change as well and thus influence the AUC. Perhaps, in our situation, a more informative prior smoothens out extreme (and unlikely) estimates, thereby increasing AUC. But the overall shift in posterior $\boldsymbol{\beta}$, from new data-dominated to legacy data-dominated, might leave the order of estimation-observation pairs intact, thus not penalizing AUC for an extreme trust in the legacy data. A possibility to prove this hypothesis would be to try a prior in which β strongly diverges from the MLE $\boldsymbol{\beta}$, for example with one regression coefficient being negative. This might cause a smaller AUC in case of a more informative prior. We conclude that, in our situation, AUC has limitations for model comparison, despite (or perhaps: because of) the advantage of being threshold independent. Unlike AUC, overall accuracy decreased if the trust in the legacy data was large (Fig. 3.10). Overall accuracy is maximized for an UMF around 100 – although users can make other choices depending on priorities. For example, if only a small new data set would be available, the trust in the legacy data may be increased to lower the risk of unrealistic estimated probabilities on a ripened subsoil. Based on Fig. A.2d in the Supplementary materials, a UMF between five and ten seems appropriate.

3.4.3 Quasi complete separation

Quasi-complete separation, affecting two of the twenty 5% subsamples, means that a covariate, or a combination of covariates, can almost perfectly predict the observations; this results in a badly defined maximum likelihood estimator (MLE; Rainey, 2016). The `MCMClogit` algorithm calculates the binomial logistic regression MLE variance-covariance matrix and multiplies it by a scalar (the jump size setting parameter) to derive the jumping distribution (Martin et al., 2011). The variance-covariance matrices of the MLE for both 5% subsamples contained extreme values (i.e. >1E6), probably resulting in a jumping proposal that is unbalanced for exploring simultaneously the dimensions of the $\boldsymbol{\beta}$ parameter space. For the low-informative prior (see for its definition Section 3.3.3.1), this quasi-complete separation itself, or the unbalanced jumping proposal, or the combination of the two caused an acceptance rate of about 0.05 (Table A.3 in Supplementary results), which is much too small and leads to results that are not meaningful. Decreasing the jump size setting increased the acceptance rate but the mixing – visually inspected by plotting the Markov chain for each element of $\boldsymbol{\beta}$, a so-called 'traceplot', was still unsatisfactory. A prior with some information increased both acceptance rate and mixing, while using the standard jump size setting of 0.9. Trial-and-error revealed that a not-case-related prior (Gaussian with means and covariances of β set to zero), with all variances equal and $\leq$ 20, or our case related prior (Gaussian, see Table A.2 in Supplementary materials) with UMF $\leq$ 10,000, contain enough information for the `MCMClogit` algorithm to mix well. Thus, using a prior with a very modest (or stronger) information level enables estimation of a meaningful β for samples with quasi-complete separation, where a (non-Bayesian) maximal likelihood algorithm fails. Related, but not similar is the proposal from Gelman et al. (2008) to use a "weakly informative prior", based on general rules-of-thumb regarding regression coefficients in binomial logistic regression to stabilize a maximum likelihood finding algorithm in case of complete separation.

3.4.4 Bayesian perspective

3.4.4.1 Bayesian paradigm

In this research, we focussed on a practical application of Bayesian statistics. The theoretical and philosophical considerations of the Bayesian paradigm, related to the perceived nature of probability, are outside the scope of this work; see for example Bernardo and Smith (2009) and Lindley (2004). But, to our opinion, the general principle that the Bayesian paradigm allows to exploit all sources of information while recognising their information content is highly relevant to soil mapping.

3.4.4.2 Using conjugate priors

We used an MCMC-algorithm to sample from the posterior β. Another often used approach in Bayesian statistics for finding the posterior is to apply conjugate priors. A conjugate prior distribution is chosen in such a way that, in combination with the likelihood function, the posterior distribution follows the same parametric form as the prior distribution. The parameters of this posterior distribution can be computed analytically, so that there is no need for sampling from the posterior. In other words, conjugancy between prior and posterior provides an analytical rather than a numerical solution for a posterior distribution, at the cost of limiting the possible choices for a prior distribution. Samaniego (2010) discusses the added value of using conjugate priors in the context of increasing computer power and software developments. For conjugate prior strategies for binomial logistic regression we refer to Chen and Ibrahim (2003).

3.4.4.3 Using new data to adjust the prior

One might question the validity of adjusting the prior through the 'uncertainty multiplication factor' on the basis of the new data. This might seem counterintuitive given the concept 'prior belief'. As we indicated in Section 3.2.2.1, literature recognizes that a prior might be adjusted after the collection of the data (which can be considered adding additional knowledge), but it should not repeat data that is captured in the likelihood. Some Bayesian approaches promote an even closer relationship between data and prior, leading to empirical Bayes and hierarchical Bayesian analysis. In our research, we allowed the uncertainty multiplication factor to depend on the new data, while we made a clear distinction between calibration new data (which appears in the likelihood), and validation new data. To arrive at our statement that – in our context – an uncertainty multiplication factor of 100 is optimal, we used the validation new data.

3.4.4.4 Added value Bayesian approach

This research showed a potential for recycling (legacy) information using a Bayesian approach, so that costs can be saved through collecting fewer new data. However, expressed in overall accuracy, large new datasets still performed best. Note that we used the full new dataset to select an optimal regression model structure. If only a small dataset is available, other model structure selection methods have to be applied. Note also that the differences between the models were often small on the scales of the validation statistics. We did not have a probability-based validation dataset and could not compute confidence intervals, which made it impossible to analyse whether observed differences were statistically significant.

In follow up research[7] we will include a spatially correlated term in the linear predictor for the logit transform of the probability (Christensen et al., 2006; Diggle and Ribeiro, 2007). If there is spatial structure that is not explained by the covariates than this approach is expected to result in more accurate maps.

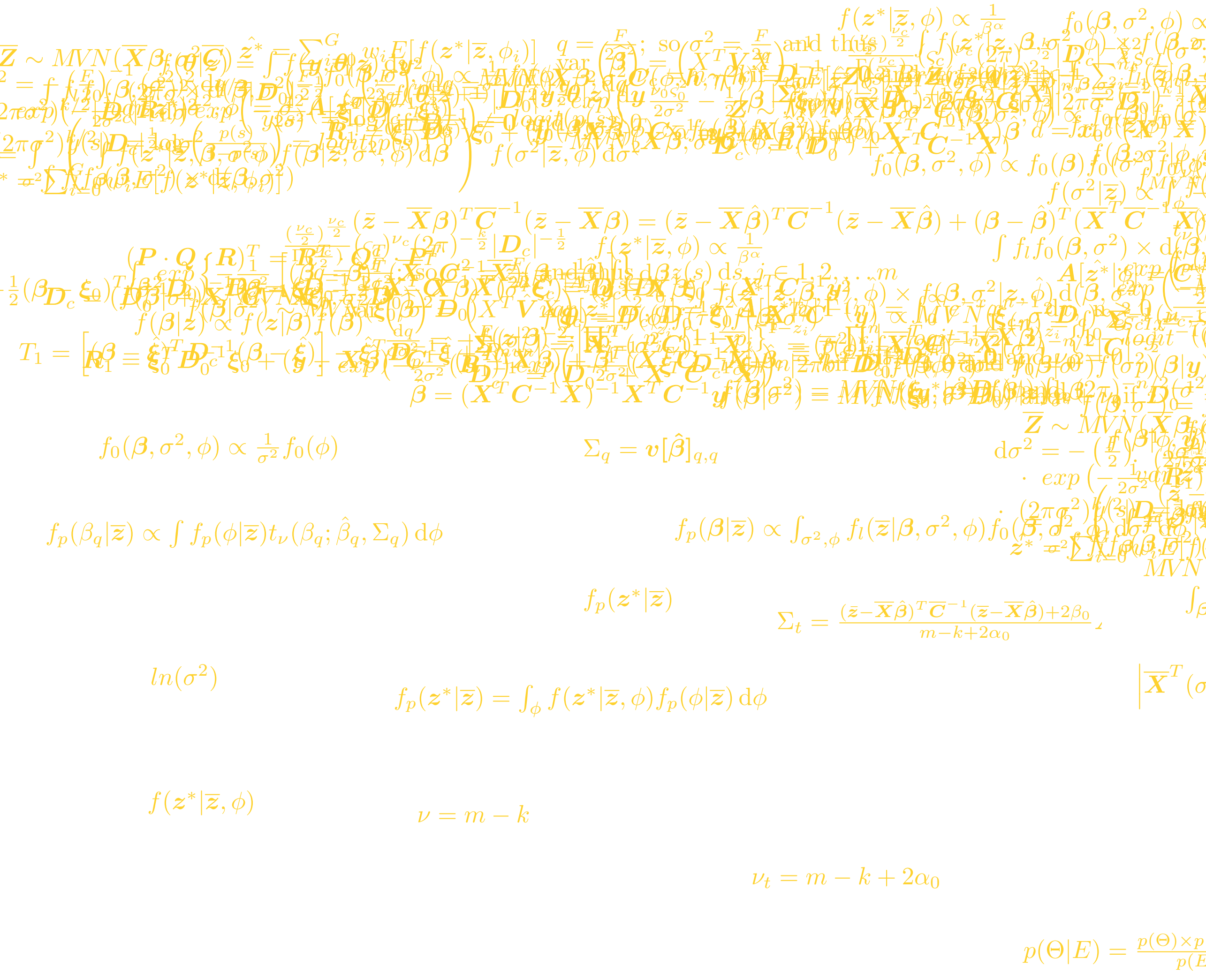
3.5 Conclusions

- Bayesian statistics allows combination of information sources by adding existing information in the form of a prior distribution to new data.
- For mapping clay soil ripening in the west of The Netherlands, Bayesian binomial logistic regression (BBLR) with a legacy data based prior yielded slightly more accurate maps than binomial logistic regression (BLR), but only in case of a limited amount of recent information (new data) available.
- To be able to use legacy data as prior, we introduced an ‘Uncertainty Multiplication Factor’, expressing our trust in the legacy data. In our case study, results (expressed in the ‘overall accuracy’ validation statistic) were optimal when standard errors of regression coefficients estimated from legacy data were multiplied by 10.
- Surprisingly, the validation statistic Area Under Curve (AUC) was maximal if we expressed an infinite trust in the legacy data. We concluded that AUC has limited capabilities for model comparison.
- In the area of interest, about 60% of unripened subsoils have ripened over the past several decades.

Supplementary materials

The supplementary materials, Appendix A., can be downloaded from the journal version of this chapter: Luc Steinbuch, Dick J Brus, Gerard BM Heuvelink, 2018. Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors. *Geoderma*, **v316**, pp.56-69; https://doi.org/10.1016/j.geoderma.2017.12.010

[7] See Chapter 5 of this thesis, however with another dataset.



$f_0(\beta, \sigma^2, \phi) \propto \frac{1}{\sigma^2} f_0(\phi)$
$\Sigma_q = \boldsymbol{v}[\hat{\beta}]_{q,q}$
$f_p(\beta_q|\bar{z}) \propto \int f_p(\phi|\bar{z}) t_\nu(\beta_q; \hat{\beta}_q, \Sigma_q)\,\mathrm{d}\phi$
$f_p(\boldsymbol{\beta}|\bar{z}) \propto \int_{\sigma^2,\phi} f_l(\bar{z}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\boldsymbol{\beta}, \sigma^2, \phi)$
$f_p(\boldsymbol{z}^*|\bar{z})$
$\Sigma_t = \frac{(\bar{z}-\overline{\boldsymbol{X}}\hat{\beta})^T \overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\beta})+2\beta_0}{m-k+2\alpha_0}$
$ln(\sigma^2)$
$f_p(\boldsymbol{z}^*|\bar{z}) = \int_\phi f(\boldsymbol{z}^*|\bar{z}, \phi) f_p(\phi|\bar{z})\,\mathrm{d}\phi$
$f(\boldsymbol{z}^*|\bar{z}, \phi)$
$\nu = m - k$
$\nu_t = m - k + 2\alpha_0$
$p(\Theta|E) = \frac{p(\Theta)\times p}{p(E}$
$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{y}$
$f(\boldsymbol{z}^*|\bar{z}, \phi) \propto \frac{1}{\beta^\alpha}$

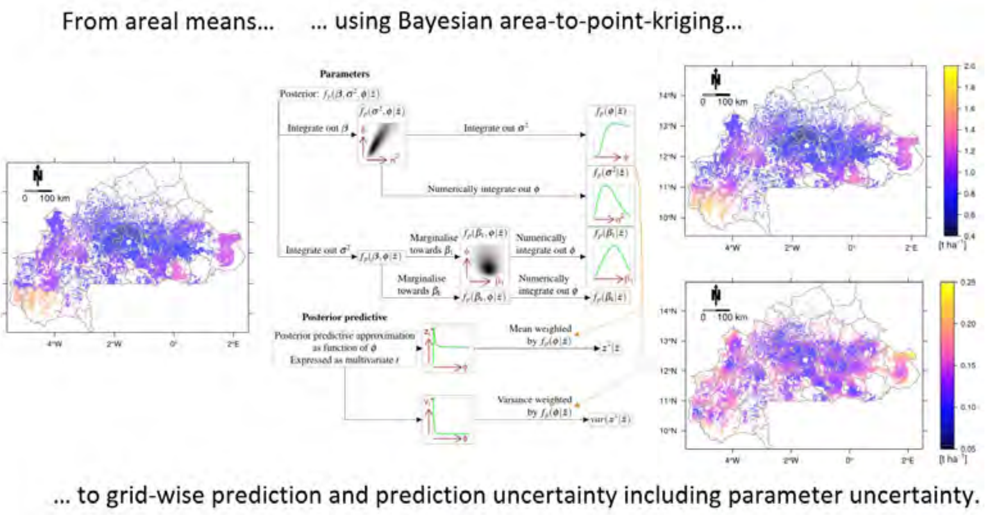
Chapter 4

Model-based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Datasets

Based on:
Luc Steinbuch, Thomas G. Orton, Dick J. Brus , 2020. Model-Based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Data Set. *Mathematical Geosciences*, **v52**, pp.397–423.

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 dataset of areal means is often considerably smaller (< 50 observations) than datasets 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 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.


From areal means...
... using Bayesian area-to-point-kriging...
Parameters
Posterior: $f_p(\beta, \sigma^2, \phi|z)$
Integrate out β
$f_p(\sigma^2, \phi|z)$
Integrate out σ^2
$f_p(\phi|z)$
Numerically integrate out ϕ
$f_p(\sigma^2|z)$
Integrate out σ^2
$f_p(\beta, \phi|z)$
Marginalise towards β_1
$f_p(\beta_1, \phi|z)$
Numerically integrate out ϕ
$f_p(\beta_1|z)$
Marginalise towards β_k
$f_p(\beta_k, \phi|z)$
Numerically integrate out ϕ
$f_p(\beta_k|z)$
Posterior predictive
Posterior predictive approximation as function of ϕ
Expressed as multivariate t
Mean weighted by $f_p(\phi|z)$
$z^*|z$
Variance weighted by $f_p(\phi|z)$
$\mathrm{var}(z^*|z)$
0 100 km
4°W
2°W
0°
2°E
14°N
13°N
12°N
11°N
10°N
[t ha⁻¹]
... to grid-wise prediction and prediction uncertainty including parameter uncertainty.

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

4.1 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, due to limited resources, in many sub-Saharan countries crop yield information is only available sparsely and in aggregated format, while crop yield information at a finer spatial resolution is needed to support increasing crop productivity (see for example Orton et al. (2018)) and thereby improve human welfare as well as ecological sustainability.

Usually, model-based spatial prediction is 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 to create 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 kriging[1] (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; Kitanidis, 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. In the Bayesian statistics paradigm, parameters can be considered stochastic rather than fixed but unknown (Schabenberger and 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[2]. 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

[1] Universal kriging is in this work 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.

[2] With "integrated out" I mean here: the regression coefficients are removed as such from the likelihood, but are implicit in the prediction, and their uncertainty contributes to the prediction error variance.

(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 dataset 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 dataset 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 datasets 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.

4.2 Theory

4.2.1 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

$$\boldsymbol{Z} \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C}(\phi)), \tag{4.1}$$

with MVN indicating a multivariate normal distribution; $\boldsymbol{X}$ 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 $\boldsymbol{C}(\phi)$ the spatial correlation matrix as a parametric function of distance parameter ϕ.[3]

[3] Incidentally, in this chapter I will also use the ‘nugget effect’ which is not in included in Eqn. (4.1). For example Eqn. (5.10), page 119, shows the nugget effect added to an exponential covariance function.

Among several other possibilities, the exponential covariance function

$$(\sigma^2 \boldsymbol{C}(\phi))_{i1,i2} = \sigma^2 exp\left(-\frac{h(s_{i1}, s_{i2})}{\phi}\right) \tag{4.2}$$

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 Euclidian distance between those locations, while $(...)_{i1,i2}$ means the element in row $i1$, column $i2$ of the indicated covariance matrix. Spatial directional dependence is not considered. For notational convenience, $\boldsymbol{C}(\phi)$ is indicated by $\boldsymbol{C}$ from now on.

4.2.2 Area-to-point kriging

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

$$\overline{z}_j = \frac{1}{|A_j|} \int_{s \in Aj} z(s)\,\mathrm{d}s, j \in 1, 2, \ldots m, \tag{4.3}$$

together $\overline{\boldsymbol{z}}$, with $\boldsymbol{z}(s)$ an unobserved realisation of an infinite number of point values $\boldsymbol{Z}$ in area A_j and $|A_j|$ the surface area of area A_j, turning Eq. (4.1) into

$$\overline{\boldsymbol{Z}} \sim MVN(\overline{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2\overline{\boldsymbol{C}}), \tag{4.4}$$

with $\overline{\boldsymbol{Z}}$ the stochastic representation of observations $\overline{\boldsymbol{z}}$, $\overline{\boldsymbol{X}}$ containing covariate values averaged over the corresponding areas, and $\sigma^2\overline{\boldsymbol{C}}$ the matrix with average covariances between and within the areas. Note that $\boldsymbol{\beta}$, σ^2 and ϕ are equivalent in Eqs. (4.4) and (4.1): 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 $\boldsymbol{z}^*$ at n^* point locations $\boldsymbol{s}^*$, together with the corresponding prediction variances $\boldsymbol{v}^*$, it is necessary to define: 1) matrix $\overline{\boldsymbol{C}}^*$, the mean correlation between points in the observation areas and the prediction points and also implicitly a function of ϕ, 2) matrix $\boldsymbol{C}^{**}$, the correlation matrix between the prediction points, again a function of ϕ, 3) the design matrix for the prediction locations $\boldsymbol{X}^*$, and finally 4) the generalised least squares (GLS) estimator for $\boldsymbol{\beta}$ as given by

$$\hat{\boldsymbol{\beta}} = (\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})^{-1}\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{z}}. \tag{4.5}$$

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

$$\begin{bmatrix} \boldsymbol{z}^* \\ \overline{\boldsymbol{z}} \end{bmatrix} \Big| \boldsymbol{\beta}, \sigma^2, \phi \sim MVN_{n^*+m}\left(\begin{bmatrix} \boldsymbol{X}^* \boldsymbol{\beta} \\ \overline{\boldsymbol{X}} \boldsymbol{\beta} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{C}^{**} & \overline{\boldsymbol{C}}^* \\ (\overline{\boldsymbol{C}}^*)^T & \overline{\boldsymbol{C}} \end{bmatrix} \right). \tag{4.6}$$

According to UK theory, the resulting prediction is given by

$$\hat{\boldsymbol{z}}^*_{UK} = \overline{\boldsymbol{C}}^* \, \overline{\boldsymbol{C}}^{-1} \, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}} \hat{\boldsymbol{\beta}}) + \boldsymbol{X}^* \hat{\boldsymbol{\beta}}, \tag{4.7}$$

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

$$\begin{aligned} &var[\hat{\boldsymbol{z}}^*_{UK}] = \\ &\sigma^2 \boldsymbol{C}^{**} - \sigma^2 \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{C}}^*)^T + \sigma^2 (\boldsymbol{X}^* - \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})^{-1} (\boldsymbol{X}^* - \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})^T, \end{aligned} \tag{4.8}$$

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

4.2.3 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

$$p(\Theta|E) = \frac{p(\Theta) \times p(E|\Theta)}{p(E)}, \tag{4.9}$$

where the probability distribution $p(\Theta|E)$ indicates the posterior degree of belief in the parameters given the evidence, $p(\Theta)$ indicates the prior degree belief in the parameters, independent of the evidence, $p(E|\Theta)$ 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

$$p(E) = \int p(\Theta, E) \, \mathrm{d}\Theta. \tag{4.10}$$

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

$$\begin{aligned} p(P|E) &= \int p(P, \Theta|E)\,\mathrm{d}\Theta \\ &= \int p(P|\Theta, E)p(\Theta|E)\,\mathrm{d}\Theta, \end{aligned} \tag{4.11}$$

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 $f_p(\ldots|\ldots)$, the likelihood by $f_l(\ldots|\ldots)$ and a prior distribution – again having propriety or not – by $f_0(\ldots)$. When the function type is ambiguous, its interpretation depends on the context and its arguments.

4.3 Implementation

4.3.1 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. (4.4), the likelihood of data generated by the geostatistical model is based on the multivariate normal distribution

$$f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta}, \sigma^2, \phi) = \frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(\bar{z} - \overline{\boldsymbol{X}}\beta)^T\overline{\boldsymbol{C}}^{-1}(\bar{z} - \overline{\boldsymbol{X}}\beta)\right\}, \tag{4.12}$$

where |...| indicates the determinant.

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

$$f_0(\boldsymbol{\beta}, \sigma^2, \phi) \propto \frac{1}{\sigma^2} f_0(\phi). \tag{4.13}$$

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)

$$\begin{aligned} &f_p(\boldsymbol{\beta}, \sigma^2, \phi | \overline{\boldsymbol{z}}) \\ &\propto \frac{1}{(2\pi)^{\frac{m}{2}} (\sigma^2)^{\frac{m}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})\right\} \frac{1}{\sigma^2} f_0(\phi). \end{aligned} \tag{4.14}$$

Based on the above assumptions, Fig. 4.1 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 Additional material A (see page 92); the summary stating the main equations follows in the coming sections.

4.3.1.1 Marginal posterior distance parameter

Given the joint posterior (Eq. 4.14), $\boldsymbol{\beta}$ and σ^2 are analytically integrated out to arrive at the analytical solution for the marginal posterior for ϕ given by

$$f_p(\phi | \overline{\boldsymbol{z}}) \propto f_0(\phi) \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \, \overline{\boldsymbol{C}}^{-1} \, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]^{\frac{m-k}{2}}}, \tag{4.15}$$

where $\hat{\boldsymbol{\beta}}$ is defined according to Eq. (4.5).

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.

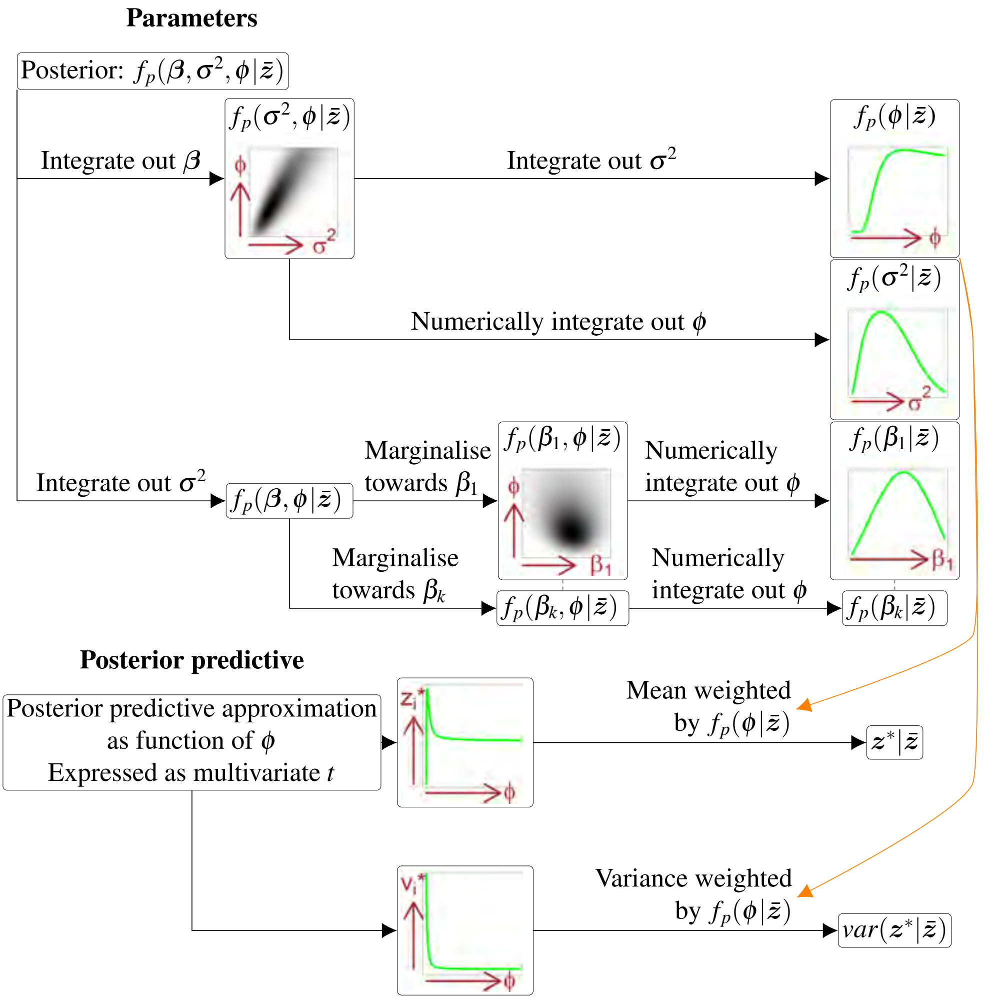

Parameters
Posterior: $f_p(\boldsymbol{\beta}, \sigma^2, \phi | \bar{z})$
Integrate out $\boldsymbol{\beta}$
$f_p(\sigma^2, \phi | \bar{z})$
ϕ
σ2
Integrate out σ^2
$f_p(\phi | \bar{z})$
ϕ
Numerically integrate out ϕ
$f_p(\sigma^2 | \bar{z})$
σ2
Integrate out σ^2
$f_p(\boldsymbol{\beta}, \phi | \bar{z})$
Marginalise towards β_1
$f_p(\beta_1, \phi | \bar{z})$
ϕ
β1
Numerically integrate out ϕ
$f_p(\beta_1 | \bar{z})$
β1
Marginalise towards β_k
$f_p(\beta_k, \phi | \bar{z})$
Numerically integrate out ϕ
$f_p(\beta_k | \bar{z})$
Posterior predictive
Posterior predictive approximation as function of ϕ Expressed as multivariate t
zi*
ϕ
Mean weighted by $f_p(\phi | \bar{z})$
$z^* | \bar{z}$
vi*
ϕ
Variance weighted by $f_p(\phi | \bar{z})$
$var(z^* | \bar{z})$

Figure 4.1: 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.

4.3.1.2 Marginal posterior sill

For the marginal posterior of σ^2, $\boldsymbol{\beta}$ is analytically integrated out from the joint posterior (Eq. 4.14) to arrive at

$$f_p(\sigma^2|\bar{\boldsymbol{z}}) \propto \int f_0(\phi) \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \bar{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \bar{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}}{\left|\bar{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\bar{\boldsymbol{X}}^T \bar{\boldsymbol{C}}^{-1} \bar{\boldsymbol{X}}\right|^{\frac{1}{2}} (\sigma^2)^{\frac{m-k+2}{2}}} \, \mathrm{d}\phi. \tag{4.16}$$

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.

4.3.1.3 Marginal posterior regression coefficients

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

$$f_p(\boldsymbol{\beta}|\overline{z}) \propto \int f_0(\phi) \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left[(\bar{z} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{z} - \overline{\boldsymbol{X}}\boldsymbol{\beta})\right]^{\frac{m}{2}}} \, \mathrm{d}\phi. \tag{4.17}$$

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

$$\boldsymbol{\Sigma}_\beta = \frac{(\bar{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\bar{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})}{m - k} (\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})^{-1}. \tag{4.18}$$

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

$$f_p(\beta_q|\overline{z}) \propto \int f_p(\phi|\overline{z}) t_\nu(\beta_q; \hat{\beta}_q, \Sigma_q) \, \mathrm{d}\phi \tag{4.19}$$

with $f_p(\phi|\overline{z})$ as indicated in Eq. (4.15) and where

$$t_\nu(\beta_q; \hat{\beta}_q, \Sigma_q) = \frac{\Gamma[(\nu + 1)/2]}{\Gamma(\nu/2)(\nu\pi)^{1/2} \left|\Sigma_q\right|^{1/2}} \left[1 + \frac{1}{\nu}(\beta_q - \hat{\beta}_q)^T \Sigma_q^{-1} (\beta_q - \hat{\beta}_q)\right]^{-(\nu+1)/2} \tag{4.20}$$

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

Similarly to Sect. 4.3.1.2, 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. (4.19); finally, it normalises to get the marginal distributions for each individual β_q.

4.3.1.4 Posterior predictive distribution

The conditional distribution for the variable of interest, given the data and any particular value of the distance parameter, $f(\boldsymbol{z}^*|\overline{\boldsymbol{z}}, \phi)$, is a t-distribution with degrees of freedom $\nu = m - k$, with location parameter $\hat{z}^*|\phi$ – an implicit function of ϕ – already given in Eq. (4.7), and with scale parameter as provided in Eq. A85 in Additional material A (see page 92). The variance of this conditional distribution, also a function of ϕ, is given by

$$\begin{aligned} var[\hat{\boldsymbol{z}}^*|\phi] = & \frac{m-k}{m-k-2} \frac{(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})}{m-k} \\ & \times \left\{ \boldsymbol{C}^{**} - \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{C}}^*)^T + (\boldsymbol{X}^* - \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})^{-1} (\boldsymbol{X}^* - \overline{\boldsymbol{C}}^* \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}})^T \right\}. \end{aligned} \tag{4.21}$$

Note that Eq. (4.21) is an increased universal kriging variance (see for comparison Eq. (4.8)) 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,

$$f_p(\boldsymbol{z}^*|\overline{\boldsymbol{z}}) = \int_\phi f(\boldsymbol{z}^*|\overline{\boldsymbol{z}}, \phi) f_p(\phi|\overline{\boldsymbol{z}}) \, \mathrm{d}\phi, \tag{4.22}$$

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 $f_p(\phi|\overline{\boldsymbol{z}})$ and the spacings of the ϕ parameter grid).

4.3.2 Methodological details

In this chapter, a number of increasingly Bayesian approaches to ATPK are applied and compared (Table 4.1. 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 Additional material 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 $f_0(\boldsymbol{\beta})$ 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 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. 4.1) are written in the statistical programming language R, and are available at Steinbuch et al. (2019).

Table 4.1: Geostatistical approaches from maximum likelihood to full Bayesian. Corresponding to their universal kriging counterparts, $\hat{\boldsymbol{z}}^*_{RK}$ and $\hat{\boldsymbol{v}}^*_{RK}$ 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	$f_l \quad := f_l(\boldsymbol{z}\|\boldsymbol{\beta}, \sigma^2, \phi)$	$\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$, $\hat{\phi}$	$f(\boldsymbol{z}^*\|\boldsymbol{z}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\phi})$	N	$\hat{\boldsymbol{z}}^*_{RK}$	$\hat{\boldsymbol{v}}^*_{RK}$
REML - restricted max- imum likelihood	$\int f_l f_0(\boldsymbol{\beta})\,\mathrm{d}\boldsymbol{\beta}$	$\hat{\sigma}^2$, $\hat{\phi}$	$\int f(\boldsymbol{z}^*\|\boldsymbol{z}, \boldsymbol{\beta}, \hat{\sigma}^2, \hat{\phi})$ $\times f(\boldsymbol{\beta}\|\boldsymbol{z}, \hat{\sigma}^2, \hat{\phi})\,\mathrm{d}\boldsymbol{\beta}$	N	$\hat{\boldsymbol{z}}^*_{UK}$	$\hat{\boldsymbol{v}}^*_{UK}$
MML - maximum marginal likelihood	$\int f_l f_0(\boldsymbol{\beta}, \sigma^2)$ $\times\,\mathrm{d}(\boldsymbol{\beta}, \sigma^2)$	$\hat{\phi}$	$\int f(\boldsymbol{z}^*\|\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2, \hat{\phi})$ $\times\, f(\boldsymbol{\beta}, \sigma^2\|\boldsymbol{z}, \hat{\phi})\,\mathrm{d}(\boldsymbol{\beta}, \sigma^2)$	t	$\hat{\boldsymbol{z}}^*_{UK}$	$\hat{\boldsymbol{v}}^*_{UK}\times \frac{m-k}{m-k-2}$
Full Bayesian	$\int f_l f_0(\boldsymbol{\beta}, \sigma^2\phi)$ $\times\,\mathrm{d}(\boldsymbol{\beta}, \sigma^2, \phi)$	-	$\int f(\boldsymbol{z}^*\|\boldsymbol{z}, \boldsymbol{\beta}, \sigma^2, \phi)$ $\times\, f_p(\boldsymbol{\beta}, \sigma^2, \phi\|\boldsymbol{z})\,\mathrm{d}(\boldsymbol{\beta}, \sigma^2, \phi)$	[a]	[b]	[c]

[a] Mixture of t

[b] Weighted average of $\hat{\boldsymbol{z}}^*_{UK}$, see Additional material A

[c] Function of $\hat{\boldsymbol{z}}^*_{UK}$, $\hat{\boldsymbol{v}}^*_{UK} \times \frac{m-k}{m-k-2}$ and $f_p(\phi|\overline{z})$, see Additional material A

4.3.2.1 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[4]; 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.

4.3.2.2 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. Then, the predictive distribution was defined conditionally on this value of ϕ. Mathematically, this equals integrating out $\boldsymbol{\beta}$ 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 4.1 and Eqs. (4.7) and (4.21).

4.3.2.3 Estimating average covariances

The average correlation matrices, $\overline{\boldsymbol{C}}$ and $\overline{\boldsymbol{C}}^*$, 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. (4.2) – and distance parameter ϕ. Then, all correlations per area-area combination are averaged to arrive at $\overline{\boldsymbol{C}}$, and per area-prediction point combination to arrive at $\overline{\boldsymbol{C}}^*$. The discretisation points were on a regular grid in the simulation study, and selected by simple random sampling in the two case studies.

[4] Online resources B..F can be found in the journal version of this chapter: Luc Steinbuch, Thomas G. Orton, Dick J. Brus , 2020. Model-Based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Data Set. *Mathematical Geosciences*, **v52**, pp.397–423.; https://doi.org/10.1007/s11004-019-09840-6

4.3.2.4 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 $\boldsymbol{z}^*$ and prediction uncertainties $\boldsymbol{v}^*$ were assessed in relation to the original signal $\boldsymbol{z}$. As an indication of the quality of the prediction, the root mean squared error (RMSE) defined as

$$RMSE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\{z(s_i) - z^*(s_i)\}^2} \tag{4.23}$$

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

$$StSE(s_i) = \frac{\{z(s_i) - z^*(s_i)\}^2}{v^*(s_i)} \tag{4.24}$$

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’.

4.4 Simulation study

The following shows a single one-dimensional simulation where REML and full Bayesian (defined with $f_0(\phi) \sim$ *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.

4.4.1 Single simulation

4.4.1.1 Simulated dataset

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; $\sigma^2_{sim} = 5$, $\phi_{sim} = 60$) was generated and added to the trend of a linear function of the coordinate ($\beta_{1sim} = 0$ for the intercept, $\beta_{2sim} = 0.02$ for the slope on coordinate). For the Bayesian approach, $f_0(\phi) \sim uniform$ was used, bounded by $\phi_{l,u} = \{10, 300\}$; these bounds also defined bounds for the REML parameter search. The priors for $\boldsymbol{\beta}$ and σ^2 are provided in Eq. (4.13). 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 $\bar{\boldsymbol{X}}$.

4.4.1.2 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. 4.2. 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. 4.3 show that, based on this – small – simulated dataset, 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.

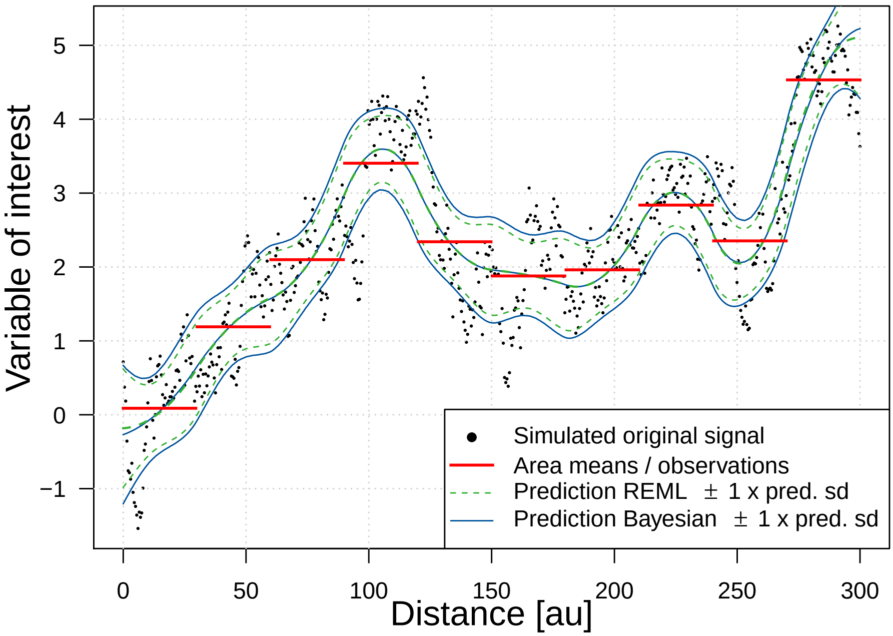


Figure 4.2: One simulation and resulting predictions, selected for illustration purposes. The predictions of REML and Bayesian with uniform prior for ϕ coincide largely, but Bayesian shows a larger prediction uncertainty. Distance is measured in abstract units (au).

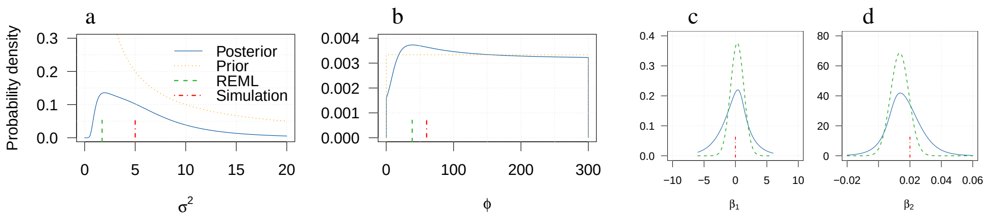


Figure 4.3: Marginal posterior distributions of spatial parameters σ^2 (a) and ϕ (b), and of the two trend parameters $\beta_{1,2}$ (c,d), resulting from the single run with Bayesian-uniform approach. REML parameter estimations and the data simulation settings (i.e. the true values) are also shown.

The pairwise joint posteriors (ϕ with each other parameter) are shown in Fig. 4.4: 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).

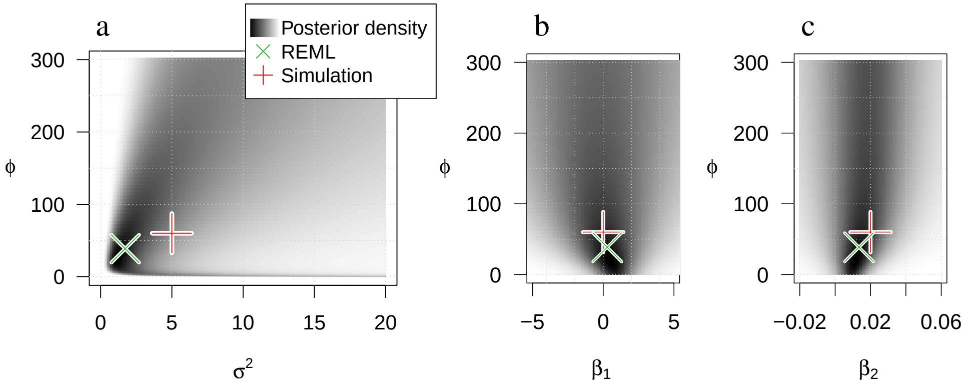


Figure 4.4: The joint posterior distributions of the spatial parameters σ^2 and ϕ (a), and of the two β elements with ϕ (b,c). REML parameter estimation and the data simulation settings are also shown. The grey shade is an indication of the posterior probability density (the darker, the larger).

Figure 4.5 shows the variogram models obtained with REML and the Bayesian-uniform approach.

Table 4.2 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 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.

Table 4.2: Results of single simulation run with validation on original data points. 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.

	mean(StSE)	RMSE	ME	max(MMP)
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

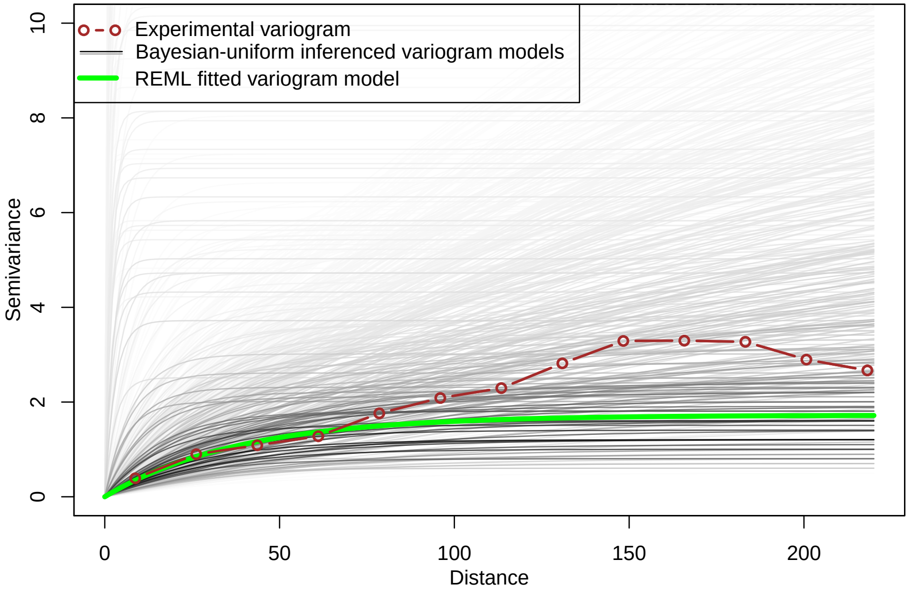


Figure 4.5: Variogram model as estimated by REML and Bayesian-uniform models (for several value combinations of σ^2 and ϕ) shaded according to their probability densities. The empirical residual variogram is added for reference.

4.4.2 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 datasets 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. 4.4.1.1. 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 $\phi = 300au$). 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 to the area sizes and total extent), which might make it difficult to identify parameters.

Table 4.3 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. More validation statistics and assessments about 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.

4.4.2.1 General results

Referring to Online Resources C and E, the maximal difference with respect to the mass preserving property ($max(MPP)$) ranges between 0.09 and 0.28 in the case of the two-dimensional simulations. In the one-dimensional case, $max(MPP)$ 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 4.3 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 *mean(StSE)* approached one, while the corresponding *sd(StSE)* decreased. Even with $m = 20$, the differences between approaches and the deviation from

Table 4.3: 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. Colours are used to highlight larger and smaller values for the readers' convenience.

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; $\nu_{sim} = 2$	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; $\nu_{sim} = .25$	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: $\phi_{sim} = 5$	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: $\phi_{sim} = 600$	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 $_{0.24}$

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 $m = 25$.

4.4.2.2 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. 4.3.2.1), having their own bounds, are not repeated here. Also recall that the extent of the simulated dataset 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.

4.4.2.3 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 known for inference and prediction. The only difference between the sessions was the means: the simulated error signals for sessions 1 to 5 were identical. Compared with a trend on the coordinate, both the GRF trend and a constant mean gave only minor differences in *mean(StSE)*; this also held for the two-dimensional simulations.

4.4.2.4 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 $\nu_{sim} = 0.5$. 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 *mean(StSE)* to be far too conservative, with the Bayesian approaches slightly more conservative, and with average *mean(StSE)* values becoming smaller with increasing m. In the two-dimensional simulations, the values stayed considerably closer to one. A small ν_{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.

4.4.2.5 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 $f_0(\phi)$ now mismatched 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 to the total extent under consideration, REML performed almost perfectly while other approaches tended to be slightly or fairly conservative, but improved with increasing m.

4.5 Case studies

4.5.1 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 dataset, 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. 4.5.2) 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. 4.5.2), 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 km 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.

4.5.2 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 the average yields only are known (averaged over the years 2000–2013, and provided by AGRHYMET), as shown in Fig. 4.6 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. 4.5.1 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. 4.6; 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 4.7 (page 88) 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.

4.6 Discussion

4.6.1 Setting uniform prior

Both in the simulations (Table 4.3 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 datasets 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 coincided with one of the bounds (for example see the ‘$\hat{\phi}, mode(\phi) \approx min, max$’ 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.

4.6.2 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

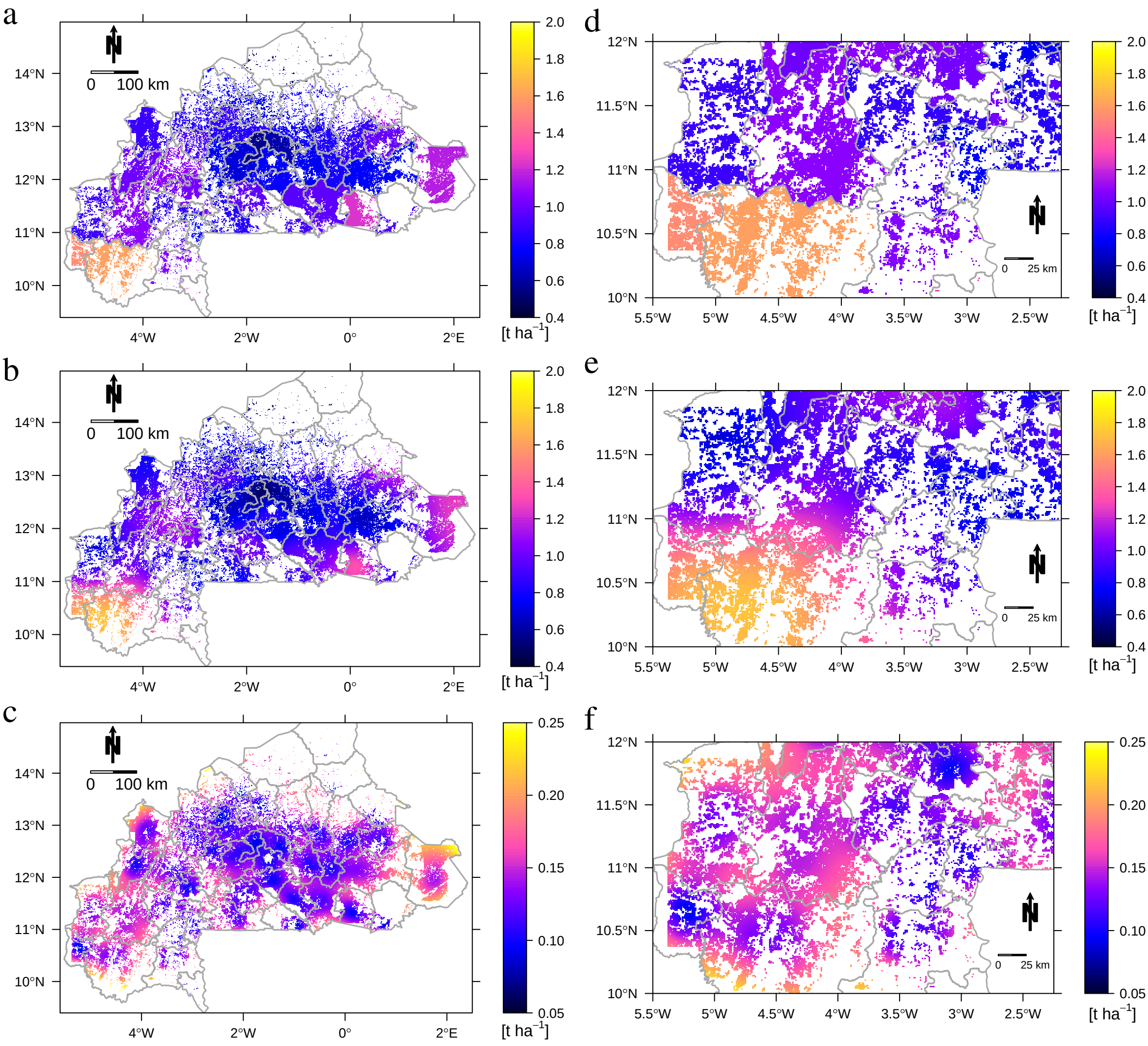


Figure 4.6: Millet case study. a) millet yields per area, averaged over the years 2000 to 2013; source: AGRHYMET (Traore et al., 2014). b) Predicted yields and c) prediction uncertainty, both from the MML approach. Subfigs. d), e) and f) show corresponding map details from the south-western part of Burkina Faso.

of the distance parameter too strongly towards zero (see also Online Resources C and D). In the case of small ν_{sim} or $\tau^2_{sim} > 0$, 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 analo-

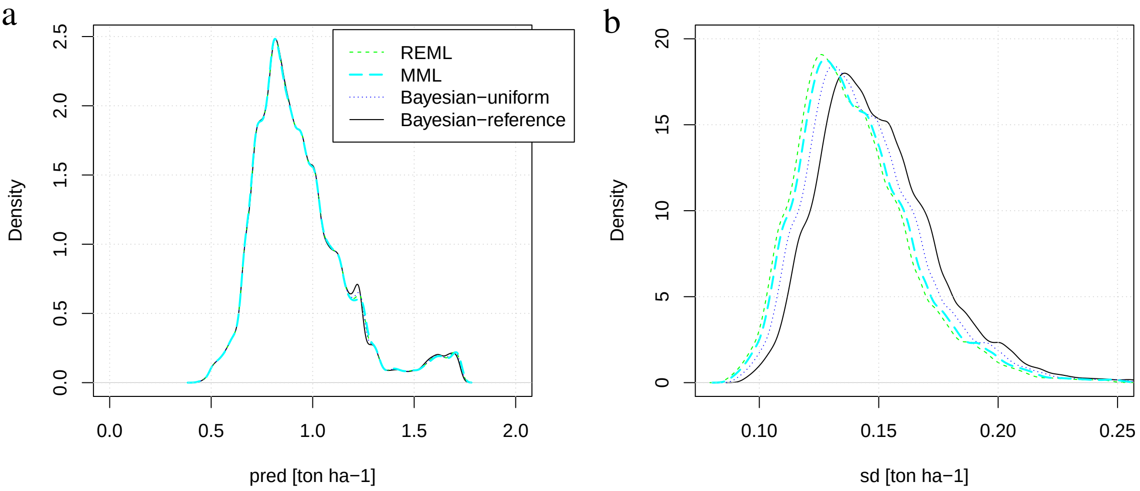


Figure 4.7: Densities of predicted millet (a) and associated sd of prediction errors (b) in Burkina Faso, with restricted maximum likelihood (REML), maximum marginal likelihood (MML), Bayesian with a uniform prior for the distance parameter and Bayesian with the reference prior.

gous 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.

4.6.3 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 my 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.

4.6.4 Number of observations

In simulation sessions 6, 7 and 8 (very smooth or very rough simulated signals), *mean(StSE)* did not converge to one with an increasing number of observations m as might be expected, and actually diverged away from one in most cases. Therefore, more data do not alleviate a poor choice of model. Furthermore, in the simulation setup, 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 datasets (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, perhaps except if strong and honest prior information is available and can be incorporated.

4.6.5 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 100au. For the one-dimensional simulations with ten observations, neighbouring units' centroids are separated by 30au. Therefore, despite there being an almost equal amount of data, there is much more short-range information in the one-dimensional data in the used setup. This might explain the extremely large *mean(StSE)* values in session 9 of the two-dimensional study (up to 49.2) compared to 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).

4.6.6 The algorithm

Although the authors did not compare the used approach with more conventional MCMC methods, it proved an effective and efficient way of performing Bayesian (and also MML) spatial data analysis in the presented area-to-point context. As indicated in Sect. 4.3.2.3, 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 implementation assumes the Markovian property on the spatial Gaussian random field (meaning that any point or area in the region is only influenced 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.

4.6.7 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 datasets 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 datasets with no prior information available, the most robust and in many cases best approach, although 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.

4.7 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 datasets 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 can be used.

Supplementary materials

Additional material A (Derivation of marginal posteriors and posterior predictive distributions): see page 92

Online Resource B: Reference Prior
Online Resource C: Results One-dimensional Simulation Ensemble
Online Resource D: Example Two-dimensional Simulation, and $mean(\mathit{StSE})$ of Two-dimensional Simulation Ensemble
Online Resource E: Results Two-dimensional Simulation Ensemble
Online Resource F: Case Study Covariates Sorghum, and Results Sorghum

Online resources B..F can be found in the journal version of this chapter: Luc Steinbuch, Thomas G. Orton, Dick J. Brus , 2020. Model-Based Geostatistics from a Bayesian Perspective: Investigating Area-to-Point Kriging with Small Data Set. *Mathematical Geosciences*, **v52**, pp.397–423. DOI: 10.1007/s11004-019-09840-6 .

The used R-scripts are provided by Steinbuch et al. (2019); the spatial real world dataset is available upon request.

4.A Additional material A

4.A.1 Introduction

In this additional material, we will – additional to the main manuscript – explain the mathematics of the provided marginal posteriors and posterior predictive distribution, and connect the mathematics to the implementation in the BAK (Bayesian areal kriging) R code.

We assume that the point-support variable is multivariate Gaussian with mean $\boldsymbol{X\beta}$ (design matrix times regression coefficients vector) and covariance matrix entries $\sigma^2\boldsymbol{C}(\boldsymbol{h};\phi)$ where $\boldsymbol{C}(\boldsymbol{h};\phi)$ is the correlation function of separation distance $\boldsymbol{h}$ with the single distance parameter ϕ. Note that since the nugget provides no contribution to the likelihood function, it is not included in this work.

The full likelihood function for these parameters, with areal-support rather than point-support data, is given by:

$$f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta},\sigma^2,\phi) = \frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(\bar{z}-\overline{\boldsymbol{X}}\boldsymbol{\beta})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\boldsymbol{\beta})\right\}, \tag{4.25}$$

with: $\bar{z}$ the m areal averages, $\overline{\boldsymbol{X}}$ the design matrix based on area average covariates), $\overline{\boldsymbol{C}}$ the mean correlation matrix, with entries given by the means of the point-support correlations – or the approximations provided by Gaussian quadrature or some other method.

We assume a prior that represents a priori independence between the trend, variance and correlation range parameters, which can therefore be decomposed as

$$f_0(\boldsymbol{\beta},\sigma^2,\phi) \propto f_0(\boldsymbol{\beta})f_0(\sigma^2)f_0(\phi) \tag{4.26}$$

The assumed prior is unlimited uniform, and thus improper, in the regression coefficient vector $\boldsymbol{\beta}$ (therefore $f_0(\boldsymbol{\beta})$ can be disregarded), is a user-specified proper distribution for ϕ, and is proportional to an inverse-gamma for σ^2 with shape and scale parameters α_0 and β_0 respectively:

$$f_0(\boldsymbol{\beta},\sigma^2,\phi) \propto f_0(\phi)\frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right). \tag{4.27}$$

Note that the inverse-gamma scale parameter, β_0 here and in general β, is not to be confused with regression coefficients vector $\boldsymbol{\beta}$, which is set in boldface, or with its elements β_q. For the inverse-gamma prior for the variance, we allow hyper-parameter values $\alpha_0 \geq -1, \beta_0 \geq 0$ (in contrast to the usual bounds for inverse-gamma parameters of $\alpha_0 > 0, \beta_0 > 0$), which permits improper priors for σ^2; in particular, with $\beta_0 = 0$ and $\alpha_0 = 0$ we obtain a prior that is proportional to the inverse of the variance, a common (improper) uninformative prior and is equivalent to a uniform prior for $log(\sigma^2)$ – it is this prior that we adopt in the main manuscript, and these hyper-parameters therefore do not appear in the text there. Also, with $\beta_0 = 0$ and $\alpha_0 = -1$ we obtain a uniform improper prior for the variance – not such a common prior for the variance , but worth noting as a special case.

In the numerical implementation, all values for $\boldsymbol{\beta}$, σ^2 and ϕ are limited and discrete.

According to Bayes' theorem, the multivariate posterior is the proportional product of prior and likelihood:

$$f_p(\boldsymbol{\beta}, \sigma^2, \phi|\overline{z}) \propto f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\boldsymbol{\beta}, \sigma^2, \phi) \tag{4.28}$$

However, we are mainly interested in the following marginal posteriors:

$f_p(\phi|\overline{z})$, so we have to integrate out $\boldsymbol{\beta}$ and σ^2;

$f_p(\sigma^2|\overline{z})$, so we have to integrate out $\boldsymbol{\beta}$ and ϕ;

$f_p(\beta_q|\overline{z})$ for each β_q in $\boldsymbol{\beta}$, so we have to integrate out σ^2 and ϕ, and subsequently find the marginal $f_p(\beta_q|\overline{z})$'s.

We are also interested in the posterior predictive distribution $f_p(\boldsymbol{z}^*|\overline{z})$, with $\boldsymbol{z}^*$ the predictions on the point locations of interest, and the associated prediction variance $\boldsymbol{v}^*$.

In the remainder of this document, the following topics are discussed:

- In section 4.A.2, *Marginal posterior for ϕ and σ^2 (via Integral a)* :

  To be able to calculate $f_p(\phi|\overline{z})$ and $f_p(\sigma^2|\overline{z})$, first we analytically integrate out $\boldsymbol{\beta}$ from the full joint posterior distribution, eqn. (4.28), to get the joint posterior for ϕ and σ^2. The result involves eqn. (4.41), which we name *Integral a*, and which is equivalent to the REML function.

- Section 4.A.3, *Marginal posterior for ϕ (via Integral b)*:

  To calculate $f_p(\phi|\overline{z})$, we analytically integrate out σ^2 from *Integral a* multiplied by the prior for σ^2. The final result is eqn. (4.53) (*Integral b*). Numerically, we create a one-dimensional, not necessarily regular, grid for ϕ in parameter space, calculate the marginal posterior for each ϕ, and normalise it to a distribution that integrates to one.

- Section 4.A.4, *Marginal posterior for σ^2*:

  To calculate $f_p(\sigma^2|\overline{z})$, the marginal posterior distribution for σ^2, we use *Integral a*. We express it as an inverse-gamma distribution conditional on ϕ, see eqn. (4.59). Numerically, we extend the vector with ϕ to a 2D grid in the parameterspace of ϕ and σ^2 and calculate the marginal posterior for σ^2 for every gridpoint; then we calculate a weighted sum over ϕ to get the marginal distribution for σ^2.

- Section 4.A.5, *Marginal posteriors for β_q (via Integral c)*:

  To calculate $f_p(\beta_q|\overline{z})$'s, the marginal posterior distribution for the β_q, we analytically integrate σ^2 out of the posterior (eqn. (4.28)) to arrive at eqn. (4.88): *Integral c*. Numerically, we extend the vector with ϕ to a grid in parameterspace of ϕ and $\boldsymbol{\beta}$ and calculate the marginal posterior for $\boldsymbol{\beta}$ for every gridpoint; then we calculate a weighted sum over ϕ to get the marginal distribution for the β_q's.

- Section 4.A.6, *Posterior predictive distribution* $f_p(\boldsymbol{z}^*|\overline{z})$:

  We use an analytically derived multivariate t-distribution to provide a prediction, eqn. (4.106), and the associated uncertainty, eqn. (4.110), for every spatial prediction point, for every ϕ in the grid vector. We use this information to numerically calculate the final prediction and uncertainty, weighted by $f_p(\phi|\overline{z})$.

- Finally, section 4.A.7, *Connect to R code* contains a pseudo-code overview to explain the functionality of the main function `postclabetaz0`, which is closely related to, but – especially in sequence – not the same as the above provided overview.

4.A.2 Marginal posterior for ϕ and σ^2 (via *Integral a*)

The joint posterior for ϕ and σ^2 is given by integrating out $\boldsymbol{\beta}$ from the full joint posterior

$$\begin{aligned} f_p(\sigma^2, \phi|\overline{z}) &\propto \int f_l(\overline{z}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\boldsymbol{\beta}, \sigma^2, \phi)\,\mathrm{d}\boldsymbol{\beta} \\ &= f_{ILa}(\overline{z}|\phi, \sigma^2) f_0(\sigma^2) f_0(\phi) \end{aligned} \tag{4.29}$$

where we define *Integrated Likelihood a*, or shortly *Integral a*, as:

$$f_{ILa}(\overline{z}|\phi, \sigma^2) = \int_{\boldsymbol{\beta}} f_l(\overline{z}|\boldsymbol{\beta}, \phi, \sigma^2)\,\mathrm{d}\boldsymbol{\beta}. \tag{4.30}$$

Eqn. (4.30) applied on eqn. (4.25) gives

$$f_{ILa}(\overline{z}|\phi, \sigma^2) = \int_{\boldsymbol{\beta}} \frac{1}{(2\pi)^{\frac{m}{2}} (\sigma^2)^{\frac{m}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp \left\{ -\frac{1}{2\sigma^2} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta}) \right\} \mathrm{d}\boldsymbol{\beta}. \tag{4.31}$$

Next, we apply the following relationship (Harville, 1974) (just preceding his Eq. 2):

$$(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta}) = (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}) (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}). \tag{4.32}$$

Combining eqns. (4.31) and (4.32), and moving some constant terms out of the integral, gives

$$f_{ILa}(\overline{z}|\phi,\sigma^2) =$$
$$\frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}}\int_{\boldsymbol{\beta}} exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\}\mathrm{d}\boldsymbol{\beta}, \tag{4.33}$$

where $\hat{\boldsymbol{\beta}}$ is the generalised least squares estimator for $\boldsymbol{\beta}$, and implicitly a function of ϕ [5]:

$$\hat{\boldsymbol{\beta}} = (\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})^{-1}\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{z}}. \tag{4.34}$$

Then, we split up the exponent (using the general rule $e^{(a+b)} = e^a e^b$), and move additional constant terms out of the integral:

$$f_{ILa}(\overline{z}|\phi,\sigma^2) =$$
$$\frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}\int_{\boldsymbol{\beta}} exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\}\mathrm{d}\boldsymbol{\beta}. \tag{4.35}$$

Define the mean and covariance parameters for a multivariate normal distribution, respectively:

$$\boldsymbol{\mu}_{mvn} = \hat{\boldsymbol{\beta}} \tag{4.36}$$

and

$$\boldsymbol{\Sigma}_{mvn} = \left[\overline{\boldsymbol{X}}^T(\sigma^2\overline{\boldsymbol{C}})^{-1}\overline{\boldsymbol{X}}\right]^{-1}. \tag{4.37}$$

Combining eqns. (4.35), (4.36) and (4.37), and multiplying by a constant term before the integral and its reciprocal inside the integral brings us this equation:

$$\begin{aligned} &f_{ILa}(\overline{z}|\phi,\sigma^2) = \\ &\frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}(2\pi)^{\frac{k}{2}}\left|\boldsymbol{\Sigma}_{mvn}\right|^{\frac{1}{2}} \\ &\times\int_{\boldsymbol{\beta}}\frac{1}{(2\pi)^{\frac{k}{2}}\left|\boldsymbol{\Sigma}_{mvn}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2}\left[(\boldsymbol{\beta}-\boldsymbol{\mu}_{mvn})^T\boldsymbol{\Sigma}_{mvn}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{mvn})\right]\right\}\mathrm{d}\boldsymbol{\beta}, \end{aligned} \tag{4.38}$$

[5] Note that most commonly, the covariance matrix instead of the correlation matrix is used in Eqn. (4.34), however σ^2 cancels out:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\overline{\boldsymbol{X}}^T(\sigma^2\overline{\boldsymbol{C}})^{-1}\overline{\boldsymbol{X}})^{-1}\overline{\boldsymbol{X}}^T(\sigma^2\overline{\boldsymbol{C}})^{-1}\overline{\boldsymbol{z}} \\ &= \left(\overline{\boldsymbol{X}}^T(\sigma^2)^{-1}\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right)^{-1}\overline{\boldsymbol{X}}^T(\sigma^2)^{-1}\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{z}} \\ &= (\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})^{-1}(\sigma^2)(\sigma^2)^{-1}\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{z}} \\ &= (\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})^{-1}\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{z}}, \end{aligned}$$

thus $\hat{\boldsymbol{\beta}}$ does not depend on σ^2.

where the integral part equals the k-variate normal distribution:

$$f_{MVN}(\boldsymbol{\beta};\boldsymbol{\mu}_{mvn},\boldsymbol{\Sigma}_{mvn}) = \frac{1}{(2\pi)^{\frac{k}{2}}\,|\boldsymbol{\Sigma}_{mvn}|^{\frac{1}{2}}}exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{mvn}})^T\boldsymbol{\Sigma}_{mvn}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{mvn}})\right\}. \tag{4.39}$$

By definition of a well-defined probability distribution, eqn. (4.39) integrates, over all possible values for β, to one[6]. Taking the remaining part of eqn. (4.38):

$$f_{ILa}(\overline{z}|\phi,\sigma^2) = \frac{1}{(2\pi)^{\frac{m}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}}exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\beta})^T\overline{\boldsymbol{C}}^{-1}(\overline{z}-\overline{\boldsymbol{X}}\hat{\beta})\right]\right\}(2\pi)^{\frac{k}{2}}\,|\boldsymbol{\Sigma}_{mvn}|^{\frac{1}{2}}, \tag{4.40}$$

and undoing the parametrisation in $\boldsymbol{\Sigma}_{mvn}$, and cancelling out wherever possible, provides us with *Integral a*, also known as the REML function (Laird and Ware, 1982) (which is more commonly presented after taking the logarithm):

$$\begin{aligned} f_{ILa}(\overline{z}|\phi,\sigma^2) &= \frac{\left((\sigma^2)^k\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{-1}\right)^{\frac{1}{2}}}{(2\pi)^{\frac{m-k}{2}}(\sigma^2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}}exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\beta})^T\overline{\boldsymbol{C}}^{-1}(\overline{z}-\overline{\boldsymbol{X}}\hat{\beta})\right]\right\} \qquad (4.41)\\ &= \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\beta})^T\overline{\boldsymbol{C}}^{-1}(\overline{z}-\overline{\boldsymbol{X}}\hat{\beta})\right]\right\}}{(2\pi)^{\frac{m-k}{2}}(\sigma^2)^{\frac{m-k}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}}. \end{aligned}$$

In the last equation, note the power in $(\sigma^2)^k$, where k is also the size of $\overline{\boldsymbol{X}}^T\overline{\boldsymbol{X}}$, based on the following equality:

$$|\boldsymbol{\Sigma}_{mvn}| = \left|\overline{\boldsymbol{X}}^T(\sigma^2\overline{\boldsymbol{C}})^{-1}\overline{\boldsymbol{X}}\right|^{-1} = (\sigma^2)^k\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{-1}, \tag{4.42}$$

which follows from the general rule that $det(b\boldsymbol{A}) = b^k det(\boldsymbol{A})$, with b a scalar, $\boldsymbol{A}$ a square matrix and k the dimensions of $\boldsymbol{A}$.

4.A.3 Marginal posterior for ϕ (via *Integral b*)

To arrive at the marginal posterior for ϕ, we have to integrate out β and σ^2 from the likelihood × prior in eqn. (4.28):

[6] We reserve the term "proper" for any probability distribution that integrates to a finite value rather than exactly to one.

$$
\begin{aligned}
f_p(\phi|\overline{\boldsymbol{z}}) &\propto \int_{\boldsymbol{\beta},\sigma^2} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta},\sigma^2,\phi) f_0(\boldsymbol{\beta},\sigma^2,\phi) d\boldsymbol{\beta} d\sigma^2 \\
&= f_{ILb}(\overline{\boldsymbol{z}}|\phi) f_0(\phi)
\end{aligned}
\tag{4.43}
$$

where we define *Integrated Likelihood b* as

$$
f_{ILb}(\overline{\boldsymbol{z}}|\phi) = \int_{\boldsymbol{\beta},\sigma^2} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta},\sigma^2,\phi) \frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right) d\boldsymbol{\beta} d\sigma^2 \tag{4.44}
$$

This integral, with $\alpha_0 = \beta_0 = 0$, with an alternative form of the prior for ϕ, is used in Berger et al. (2001). Applying eqn.(4.41) (*Integral a*) gives:

$$
f_{ILb}(\overline{\boldsymbol{z}}|\phi) = \int_{\sigma^2} \frac{1}{(2\pi)^{\frac{m-k}{2}} (\sigma^2)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \, \overline{\boldsymbol{C}}^{-1} \, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right\} \frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right) \mathrm{d}\sigma^2 \tag{4.45}
$$

Combining, and subsequently moving the elements that do not depend on σ^2 out of the integral gives:

$$
f_{ILb}(\overline{\boldsymbol{z}}|\phi) = \frac{1}{(2\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \int \sigma^2 \frac{1}{(\sigma^2)^{\frac{m-k}{2}+\alpha_0+1}} exp\left\{-\frac{1}{2\sigma^2}\left\{(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \, \overline{\boldsymbol{C}}^{-1} \, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right\}\right\} \mathrm{d}\sigma^2. \tag{4.46}
$$

Next, we define the two parameters for an inverse-gamma distribution:

$$
\alpha = \frac{1}{2}(m-k) + \alpha_0 \tag{4.47}
$$

$$
\beta = \frac{1}{2}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \, \overline{\boldsymbol{C}}^{-1} \, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + \beta_0, \tag{4.48}
$$

and noting the inverse-gamma distribution:

$$
f(\sigma^2; \alpha, \beta) = \frac{(\beta)^\alpha}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} exp\left\{-\frac{\beta}{\sigma^2}\right\}, \tag{4.49}
$$

we can write eqn. (4.46) as:

$$
f_{ILb}(\overline{\boldsymbol{z}}|\phi) = \frac{1}{(2\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \frac{\Gamma(\alpha)}{(\beta)^\alpha} \int_{\sigma^2} \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} exp\left\{-\frac{\beta}{\sigma^2}\right\} \mathrm{d}\sigma^2. \tag{4.50}
$$

Noting that the integral part equals eqn. (4.49), and when integrated over all possible values of σ^2 integrates to one (assuming that $\alpha > 0$ and $\beta > 0$), we arrive at

$$f_{ILb}(\overline{z}|\phi) = \frac{1}{(2\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \frac{\Gamma(\alpha)}{(\beta)^{\alpha}}. \tag{4.51}$$

Undoing the parametrisation by α and β

$$f_{ILb}(\overline{z}|\phi) = \frac{\Gamma(\frac{m-k+2\alpha_0}{2})}{(2\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[\frac{1}{2}(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + \beta_0\right]^{\frac{m-k+2\alpha_0}{2}}}, \tag{4.52}$$

brings us the final formulation of the integrated likelihood (*Integral b*):

$$\begin{aligned}
f_{ILb}(\overline{z}|\phi) &= \frac{\Gamma(\frac{m-k+2\alpha_0}{2})}{(2\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[\frac{1}{2}(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + \beta_0\right]^{\frac{m-k+2\alpha_0}{2}}} \\
&= \frac{\Gamma(\frac{m-k+2\alpha_0}{2})}{(2)^{\frac{m-k}{2}} (\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} (\frac{1}{2})^{\frac{m-k+2\alpha_0}{2}} \left[(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m-k+2\alpha_0}{2}}} \\
&= \frac{2^{\alpha_0}\, \Gamma\left(\frac{m-k+2\alpha_0}{2}\right)}{(\pi)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m-k+2\alpha_0}{2}}}.
\end{aligned} \tag{4.53}$$

The marginal posterior for ϕ is calculated by multiplying *Integral b* by the prior $f_0(\phi)$, as shown in eq. (4.43); in this final step, we also remove all constants, to arrive at:

$$f_p(\phi|\overline{z}) \propto f_0(\phi) \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m-k+2\alpha_0}{2}}} \tag{4.54}$$

Setting $\alpha_0 = \beta_0 = 0$ leads to the form presented in the main manuscript:

$$f_p(\phi|\overline{z}) \propto f_0(\phi) \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \left[(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \,\overline{\boldsymbol{C}}^{-1}\, (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]^{\frac{m-k}{2}}} \tag{4.55}$$

Numerically, BAK creates an one-dimensional, regularly spaced grid for ϕ in parameter space, calculates the marginal posterior for each ϕ, and normalises the marginal posterior to a proper distribution.

4.A.4 Marginal posterior for σ^2

The posterior distribution for σ^2 is given by marginalising the full joint posterior, eq. (4.28):

$$f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_{\boldsymbol{\beta},\phi} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta},\sigma^2,\phi) f_0(\boldsymbol{\beta},\sigma^2,\phi)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\phi. \tag{4.56}$$

Splitting the prior into its constituent parts (via eq. (4.26)), and taking terms outside of the integrals where appropriate, we have:

$$f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_{\boldsymbol{\beta},\phi} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta},\sigma^2,\phi)\,\mathrm{d}\boldsymbol{\beta} f_0(\sigma^2) f_0(\phi)\,\mathrm{d}\phi. \tag{4.57}$$

The integral with respect to $\boldsymbol{\beta}$ in the above equation is our *Integral a*, already presented in eqn. (4.41). Combining *Integral a* with the prior for σ^2 gives:

$$f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi \frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right) \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}}{(2\pi)^{\frac{m-k}{2}} (\sigma^2)^{\frac{m-k}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi)\,\mathrm{d}\phi. \tag{4.58}$$

Combining and moving, and removing constants gives:

$$f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]\right\}}{(\sigma^2)^{\frac{m-k+2\alpha_0+2}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi)\,\mathrm{d}\phi. \tag{4.59}$$

We can set $\alpha_0 = \beta_0 = 0$ and express the marginal posterior for σ^2 as an integral:

$$f_p(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi f_0(\phi) \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1} (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T \overline{\boldsymbol{C}}^{-1} \overline{\boldsymbol{X}}\right|^{\frac{1}{2}} (\sigma^2)^{\frac{m-k+2}{2}}}\,\mathrm{d}\phi. \tag{4.60}$$

We note an alternative form for expressing the posterior distribution, eq. (4.59), in terms of an an inverse-gamma distribution for σ^2 and the posterior distribution for ϕ. Redefine the inverse-gamma parameters for this context:

$$\alpha = \frac{m - k + 2\alpha_0}{2} \tag{4.61}$$

and

$$\beta = \frac{(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\, \overline{\boldsymbol{C}}^{-1}\, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0}{2}. \tag{4.62}$$

Reminder: the inverse-gamma distribution is, using the above parameters, defined as:

$$f_{IG}(\sigma^2;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \sigma^{2^{-\alpha-1}} exp\left(-\frac{\beta}{\sigma^2}\right). \tag{4.63}$$

Then, we can write:

$$
\begin{aligned}
f(\sigma^2|\overline{\boldsymbol{z}}) &\propto \int_\phi \frac{exp\left\{-\frac{\beta}{\sigma^2}\right\}}{(\sigma^2)^{\alpha+1}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi)\,\mathrm{d}\phi \\
&= \int_\phi \frac{\Gamma(\alpha)}{\beta^\alpha}\frac{\beta^\alpha}{\Gamma(\alpha)} \frac{exp\left\{-\frac{\beta}{\sigma^2}\right\}}{(\sigma^2)^{\alpha+1}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi)\,\mathrm{d}\phi \qquad (4.64) \\
&= \int_\phi \frac{\Gamma(\alpha)}{\beta^\alpha} \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi) f_{IG}(\sigma^2;\alpha,\beta)\,\mathrm{d}\phi.
\end{aligned}
$$

Explicitly substituting the expressions for α and β into eq. (4.64) gives:

$$
f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi \frac{\Gamma(\frac{m-k+2\alpha_0}{2})}{\left[\frac{(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0}{2}\right]^{\frac{m-k+2\alpha_0}{2}}} \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} f_0(\phi) f_{IG}(\sigma^2;\alpha,\beta)\,\mathrm{d}\phi. \qquad (4.65)
$$

and dropping constant terms leads to:

$$
f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}\left[(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m-k+2\alpha_0}{2}}} f_0(\phi) f_{IG}(\sigma^2;\alpha,\beta)\,\mathrm{d}\phi. \qquad (4.66)
$$

Comparing this expression with eq. (4.54), we can see that:

$$
f(\sigma^2|\overline{\boldsymbol{z}}) \propto \int_\phi f_p(\phi|\overline{\boldsymbol{z}}) f(\sigma^2|\overline{\boldsymbol{z}},\phi)\,\mathrm{d}\phi. \qquad (4.67)
$$

where

$$
f(\sigma^2|\overline{\boldsymbol{z}},\phi) = f_{IG}(\sigma^2;\frac{m-k+2\alpha_0}{2},\frac{(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0}{2}) \qquad (4.68)
$$

Numerically, BAK extends the vector with ϕ to a 2D grid in parameterspace of ϕ and σ^2, and calculates the joint posterior (the integrand) for σ^2 and ϕ for every gridpoint; then it performs a trapezoidal integration over ϕ, and normalises, to arrive at the marginal distribution for σ^2.

4.A.5 Marginal posteriors for β_q (via *Integral c*)

To arrive at $f_p(\beta_q|\overline{z})$, we first derive $f_p(\beta|\overline{z}, \phi)$ by again marginalising the full posterior distribution, eq. (4.28):

$$f_p(\boldsymbol{\beta}|\overline{\boldsymbol{z}}) \propto \int_{\sigma^2,\phi} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\boldsymbol{\beta}, \sigma^2, \phi)\, \mathrm{d}\sigma^2\, \mathrm{d}\phi. \tag{4.69}$$

Again splitting the prior into its constituent parts (via eq. (4.26)), and taking terms outside of the integrals where appropriate, we have:

$$\begin{aligned} f_p(\boldsymbol{\beta}|\overline{\boldsymbol{z}}) &\propto \int_{\sigma^2,\phi} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\sigma^2)\, \mathrm{d}\sigma^2 f_0(\phi)\, \mathrm{d}\phi \\ &= \int_{\phi} f_{ILc}(\overline{z}|\boldsymbol{\beta}, \phi) f_0(\phi)\, \mathrm{d}\phi \end{aligned} \tag{4.70}$$

where we define *Integrated Likelihood c* (ILc) as:

$$f_{ILc}(\overline{z}|\boldsymbol{\beta}, \phi) = \int_{\sigma^2} f_l(\overline{\boldsymbol{z}}|\boldsymbol{\beta}, \sigma^2, \phi) f_0(\sigma^2)\, \mathrm{d}\sigma^2. \tag{4.71}$$

Focusing now on the expression for the *Integrated Likelihood c*, we rearrange, and move the constant part out of the integral to give:

$$\begin{aligned} f_{ILc}(\overline{z}|\boldsymbol{\beta}, \phi) &= \int_{\sigma^2} \frac{1}{(2\pi)^{\frac{m}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} \frac{1}{(\sigma^2)^{(\frac{m}{2})}} exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})\right]\right\} \\ &\times \frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right) \mathrm{d}\sigma^2 \\ &= \frac{1}{(2\pi)^{\frac{m}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}} \int_{\sigma^2} \frac{1}{(\sigma^2)^{(\frac{m}{2}+\alpha_0+1)}} exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta})^T \overline{\boldsymbol{C}}^{-1} (\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\boldsymbol{\beta}) + 2\beta_0\right]\right\} \mathrm{d}\sigma^2. \end{aligned} \tag{4.72}$$

Now we redefine the inverse-gamma parameters for this context:

$$\alpha = \frac{m}{2} + \alpha_0, \tag{4.73}$$

$$\beta = \frac{1}{2}\left[(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\, \overline{\boldsymbol{C}}^{-1}\, (\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right], \tag{4.74}$$

and - for readability - we repeat the inverse-gamma distribution with above parameters:

$$f_{IG}(\sigma^2; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} (\sigma^2)^{-\alpha-1} exp\left\{-\frac{\beta}{\sigma^2}\right\}. \tag{4.75}$$

Thus we can write eqn. (4.72) as eqn. (4.76) in which the integrand is replaced by eqn. (4.75):

$$f_{ILc}(\overline{z}|\boldsymbol{\beta},\phi) = \frac{1}{(2\pi)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}}\frac{\Gamma(\alpha)}{\beta^{\alpha}}\int_{\sigma^2}\frac{\beta^{\alpha}}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}exp\left\{-\frac{\beta}{\sigma^2}\right\}\mathrm{d}\sigma^2. \tag{4.76}$$

Applying that the integral integrates to one, and reversing the parametrisation

$$\begin{aligned}
f_{ILc}(\overline{z}|\boldsymbol{\beta},\phi) &= \frac{\Gamma\left(\frac{m}{2}+\alpha_0\right)}{(2\pi)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[\frac{1}{2}(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})+\beta_0\right]^{\frac{m}{2}+\alpha_0}}\\
&= \frac{\Gamma\left(\frac{m}{2}+\alpha_0\right)}{(\pi)^{\frac{m}{2}}(2)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})+2\beta_0\right]^{\frac{m}{2}+\alpha_0}(\frac{1}{2})^{\frac{m}{2}+\alpha_0}} \qquad (4.77)\\
&= \frac{2^{\alpha_0}\;\Gamma\left(\frac{m}{2}+\alpha_0\right)}{(\pi)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})^T\,\overline{\boldsymbol{C}}^{-1}\,(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\boldsymbol{\beta})+2\beta_0\right]^{\frac{m}{2}+\alpha_0}}
\end{aligned}$$

Next, returning to the joint marginal posterior for $\boldsymbol{\beta}$, eq. (4.70), applying the relationship in eqn. (4.32) where $\hat{\boldsymbol{\beta}}$ is again the GLS estimate and thus a function of ϕ (and $\boldsymbol{C}$), and further removing all constants gives:

$$\begin{aligned}
f_p(\boldsymbol{\beta}|\overline{\boldsymbol{z}}) &\propto \int_{\phi} f_{ILc}(\overline{z}|\boldsymbol{\beta},\phi)f_0(\phi)\,\mathrm{d}\phi\\
&\propto \int_{\phi}\frac{2^{\alpha_0}\;\Gamma\left(\frac{m}{2}+\alpha_0\right)}{(\pi)^{\frac{m}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m}{2}+\alpha_0}}f_0(\phi)\,\mathrm{d}\phi\\
&\propto \int_{\phi}\frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m+2\alpha_0}{2}}}f_0(\phi)\,\mathrm{d}\phi.
\end{aligned} \tag{4.78}$$

We rearrange the sum in eqn. (4.78) [equivalent to $a + b \propto 1 + b/a, a \neq 0$], extend the power with $-k + k$, and multiply the main fraction inside the denominator by $\frac{m-k+2\alpha_0}{m-k+2\alpha_0}$

respectively:

$$
\begin{aligned}
f_p(\boldsymbol{\beta}|\overline{\boldsymbol{z}}) &\propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m+2\alpha_0}{2}} \left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0}\right]^{\frac{m+2\alpha_0}{2}}} f_0(\phi)\,\mathrm{d}\phi \\
&\propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m+2\alpha_0}{2}} \left[1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0}\right]^{\frac{m-k+k+2\alpha_0}{2}}} f_0(\phi)\,\mathrm{d}\phi \\
&\propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right]^{\frac{m+2\alpha_0}{2}}} \\
&\qquad \times \frac{1}{\left[1 + \frac{1}{m-k+2\alpha_0}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T \frac{(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})}{\left((\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right)/(m-k+2\alpha_0)}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]^{\frac{m-k+k+2\alpha_0}{2}}} f_0(\phi)\,\mathrm{d}\phi.
\end{aligned}
\tag{4.79}
$$

Note that $(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\,\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})/(m-k)$ (i.e. the denominator with $\alpha_0 = \beta_0 = 0$) is the REML estimate of σ^2 given $\overline{\boldsymbol{C}}^{-1}$. Also note that the form is now comparable to that of the multivariate t distribution (MVt) (Roth, 2013; Gelman et al., 2013), with degrees of freedom[7] ν

$$\nu = m - k + 2\alpha_0, \tag{4.80}$$

location vector

$$\mu = \hat{\boldsymbol{\beta}}, \tag{4.81}$$

and shape matrix $\boldsymbol{\Sigma}$, where

$$\boldsymbol{\Sigma}^{-1} = \frac{(\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}})}{\left((\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right)/(m-k+2\alpha_0)}. \tag{4.82}$$

[7] Don't confuse the degrees of freedom ν of a t distribution with the smoothness parameter of the Matérn covariance function, in this chapter represented by ν_{sim} or ν_M.

The resulting MVt is defined as:

$$t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma}) = \frac{\Gamma[(\nu+k)/2]}{\Gamma(\nu/2)\nu^{k/2}\pi^{k/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\left[1+\frac{1}{\nu}(\beta-\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]^{-(\nu+k)/2} \tag{4.83}$$

and therefore (4.79) can be written as:

$$f_p(\boldsymbol{\beta}|\overline{z}) \propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m+2\alpha_0}{2}}}\frac{\Gamma(\nu/2)\nu^{k/2}\pi^{k/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}{\Gamma[(\nu+k)/2]}t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma})f_0(\phi)\,\mathrm{d}\phi \tag{4.84}$$

We note that the determinant of $\boldsymbol{\Sigma}$ is given by:

$$\left|\boldsymbol{\Sigma}\right| = \frac{\left[\left((\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right)/(m-k+2\alpha_0)\right]^k}{\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|}, \tag{4.85}$$

and therefore dropping constant terms and rearranging leads to:

$$\begin{aligned}
f_p(\boldsymbol{\beta}|\overline{z}) &\propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m+2\alpha_0}{2}}}\left|\boldsymbol{\Sigma}\right|^{1/2}t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma})f_0(\phi)\,\mathrm{d}\phi \\
&\propto \int_\phi \frac{\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{k/2}}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m+2\alpha_0}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}}\ t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma})f_0(\phi)\,\mathrm{d}\phi \\
&\propto \int_\phi \frac{1}{\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})+2\beta_0\right]^{\frac{m-k+2\alpha_0}{2}}}\ f_0(\phi)\ \ t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma})\,\mathrm{d}\phi
\end{aligned} \tag{4.86}$$

Comparing this expression with eq. (4.54), we can write:

$$f_p(\boldsymbol{\beta}|\overline{z}) \propto \int_\phi f_p(\phi|\overline{z})\ \ t_\nu(\boldsymbol{\beta};\hat{\boldsymbol{\beta}},\boldsymbol{\Sigma})\,\mathrm{d}\phi. \tag{4.87}$$

According to Roth (2013) the marginal distributions of the multivariate t distribution are also t (his Eqn. (4.3)):

$$t_\nu(\beta_q;\hat{\beta}_q,\Sigma_q) = \frac{\Gamma[(\nu+k_q)/2]}{\Gamma(\nu/2)\nu^{k_q/2}\pi^{k_q/2}\left|\Sigma_q\right|^{1/2}}\left[1+\frac{1}{\nu}(\beta_q-\hat{\beta}_q)^T\Sigma_q^{-1}(\beta_q-\hat{\beta}_q)\right]^{-(\nu+k_q)/2}. \tag{4.88}$$

with k_q the dimension of β_q, in our case one, $\hat{\beta}_q$ the qth element of $\hat{\boldsymbol{\beta}}$ as the location parameter and $\Sigma_q = \boldsymbol{v}[\hat{\boldsymbol{\beta}}]_{q,q}$ the qth diagonal of matrix $\boldsymbol{v}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\Sigma}^{-1}$, as the scale parameter. Therefore in our case:

$$t_\nu(\beta_q;\hat{\beta}_q,\Sigma_q) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)(\nu\pi)^{1/2}\left|\Sigma_q\right|^{1/2}}\left[1+\frac{1}{\nu}(\beta_q-\hat{\beta}_q)^T\Sigma_q^{-1}(\beta_q-\hat{\beta}_q)\right]^{-(\nu+1)/2}, \tag{4.89}$$

and the final marginal posterior for β_q is given by:

$$f_p(\beta_q|\overline{z}) \propto \int_\phi f_p(\phi|\overline{z})\ \ t_\nu(\beta_q;\hat{\beta}_q,\Sigma_q)\,\mathrm{d}\phi \tag{4.90}$$

Similarly to Section 4.A.4, BAK creates 2-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. (4.89); finally it normalises to get the marginal distributions for the individual β_q's.

4.A.6 Posterior predictive distribution $f_p(\boldsymbol{z}^*|\overline{z})$

We consider our spatial phenomenon of interest a Gaussian Random field, therefore we can define our n^* point predictions ($\boldsymbol{z}^*$) together with m area observations ($\overline{z}$) as one $n^* + m$-dimensional multivariate normal distribution:

$$\begin{bmatrix}\boldsymbol{z}^*\\ \overline{z}\end{bmatrix}\Bigg|\boldsymbol{\beta},\sigma^2,\phi \sim MVN_{n^*+m}\left(\begin{bmatrix}\boldsymbol{X}^*\boldsymbol{\beta}\\ \overline{\boldsymbol{X}}\boldsymbol{\beta}\end{bmatrix},\sigma^2\begin{bmatrix}\boldsymbol{C}^{**} & \overline{\boldsymbol{C}}^*\\ (\overline{\boldsymbol{C}}^*)^T & \overline{\boldsymbol{C}}\end{bmatrix}\right), \tag{4.91}$$

with $\boldsymbol{C}^{**}$ the correlation matrix between prediction points, dimension $n^* \times n^*$, and $\overline{\boldsymbol{C}}^*$ the mean correlation between observation area and prediction points, dimension $n^* \times m$.

The posterior predictive distribution is defined by:

$$f_p(\boldsymbol{z}^*|\overline{z}) = \int_\phi f(\boldsymbol{z}^*|\overline{z},\phi)f_p(\phi|\overline{z})\,\mathrm{d}\phi. \tag{4.92}$$

Taking out the first part under the integral

$$\begin{aligned} f(\boldsymbol{z}^*|\overline{z},\phi) &= \int\limits_{\boldsymbol{\beta},\sigma^2} f(\boldsymbol{z}^*,\boldsymbol{\beta},\sigma^2|\overline{z},\phi)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2 \\ &= \int\limits_{\boldsymbol{\beta},\sigma^2} f(\boldsymbol{z}^*|\overline{z},\beta,\sigma^2\phi) f(\boldsymbol{\beta},\sigma^2|\overline{z},\phi)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2 \\ &= \int\limits_{\boldsymbol{\beta},\sigma^2} f(\boldsymbol{z}^*|\overline{z},\boldsymbol{\beta},\sigma^2\phi) f(\boldsymbol{\beta}|\overline{z},\sigma^2,\phi) f(\sigma^2|\overline{z},\phi)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2 \\ &= \int\limits_{\sigma^2} \left(\int\limits_{\boldsymbol{\beta}} f(\boldsymbol{z}^*|\overline{z},\boldsymbol{\beta},\sigma^2\phi) f(\boldsymbol{\beta}|\overline{z},\sigma^2,\phi)\,\mathrm{d}\boldsymbol{\beta} \right) f(\sigma^2|\overline{z},\phi)\,\mathrm{d}\sigma^2 \end{aligned} \tag{4.93}$$

The integral with respect to $\boldsymbol{\beta}$ results in a Gaussian pdf (Kitanidis, 1986, chapter 4) for $\boldsymbol{z}^*$, known as universal kriging (here formulated for the area-to-point context) with mean:

$$\hat{\boldsymbol{z}}^*|\sigma^2,\phi = \overline{\boldsymbol{C}}^*\ \boldsymbol{C}^{-1}\ (\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + \boldsymbol{X}^*\hat{\boldsymbol{\beta}}, \tag{4.94}$$

and corresponding variance-covariance matrix:

$$var[\hat{\boldsymbol{z}}^*|\sigma^2,\phi] = \sigma^2\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi], \tag{4.95}$$

where

$$\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi] = \boldsymbol{C}^{**} - \overline{\boldsymbol{C}}^*\boldsymbol{C}^{-1}(\overline{\boldsymbol{C}}^*)^T + (\boldsymbol{X}^* - \overline{\boldsymbol{C}}^*\boldsymbol{C}^{-1}\overline{\boldsymbol{X}})(\overline{\boldsymbol{X}}^T\boldsymbol{C}^{-1}\overline{\boldsymbol{X}})^{-1}(\boldsymbol{X}^* - \overline{\boldsymbol{C}}^*\boldsymbol{C}^{-1}\overline{\boldsymbol{X}})^T. \tag{4.96}$$

Note that we write $\hat{\boldsymbol{z}}^*|\sigma^2,\phi$ and $var[\hat{\boldsymbol{z}}^*|\sigma^2,\phi]$ here to make explicit the dependence of the distribution that these values parameterise on the parameters σ^2 and ϕ. Also note that we have cancelled out σ^2 wherever possible, in particular so that $\hat{\boldsymbol{z}}^*|\sigma^2,\phi$ does not actually depend on σ^2, thus we write $\hat{\boldsymbol{z}}^*|\sigma^2,\phi = \hat{\boldsymbol{z}}^*|\phi$, and that the diagonal elements of matrix $var[\hat{\boldsymbol{z}}^*]$ are the prediction variances, also known as kriging variances, as often presented in other literature.

Now we have:

$$f(\boldsymbol{z}^*|\overline{z},\phi) \propto \int\limits_{\sigma^2} \frac{1}{\left|\sigma^2\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}}} exp\left\{ -\frac{1}{2\sigma^2}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)^T\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi) \right\} f(\sigma^2|\overline{z},\phi)\,\mathrm{d}\sigma^2 \tag{4.97}$$

For $f(\sigma^2|\bar{z},\phi)$, we refer back to the REML equation in Equation (4.41) preceeded by the prior for σ^2, expressed as an inverse-gamma distribution with parameters α_0 and β_0:

$$f(\sigma^2|\bar{z},\phi) \propto \frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right) \times \frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}}{(2\pi)^{\frac{m-k}{2}}(\sigma^2)^{\frac{m-k}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \tag{4.98}$$

Combining the previous two equations, and taking out the constant $(2\pi)^{\frac{m-k}{2}}$:

$$\begin{aligned} f(\boldsymbol{z}^*|\bar{z},\phi) \propto \int\limits_{\sigma^2} &\frac{1}{\left|\sigma^2\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}}} exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)^T\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)\right\} \\ &\times\frac{1}{(\sigma^2)^{\alpha_0+1}} exp\left(\frac{-\beta_0}{\sigma^2}\right)\frac{exp\left\{-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}}{(\sigma^2)^{\frac{m-k}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}}\,\mathrm{d}\sigma^2 \end{aligned} \tag{4.99}$$

Separating (and subsequently combining) exponents and denominators:

$$\begin{aligned} f(\boldsymbol{z}^*|\bar{z},\phi) \propto \int\limits_{\sigma^2} &\frac{1}{\left|\sigma^2\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}}(\sigma^2)^{\alpha_0+1}(\sigma^2)^{\frac{m-k}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \\ &\times exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)^T\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)-\frac{-\beta_0}{\sigma^2}-\frac{1}{2\sigma^2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]\right\}\mathrm{d}\sigma^2 \end{aligned} \tag{4.100}$$

Next, we separate elements without σ^2 from the integral, and reorganise. Note that the dimension of $\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]$ is $n^*\times n^*$.

$$\begin{aligned} f(\boldsymbol{z}^*|\bar{z},\phi) \propto &\frac{1}{\left|\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}}\left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}}\int\limits_{\sigma^2}\frac{1}{(\sigma^2)^{\frac{n^*+m-k}{2}+\alpha_0+1}} \\ &\times exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)^T\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^*-\hat{\boldsymbol{z}}^*|\phi)+\frac{1}{2}\left[(\bar{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{z}-\overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right]+\beta_0\right]\right\}\mathrm{d}\sigma^2 \end{aligned} \tag{4.101}$$

Now, the integral part can be rewritten as proportional to an inverse-gamma distribution with shape

$$\alpha = \frac{n^*+m-k}{2}+\alpha_0, \tag{4.102}$$

and scale

$$\beta = \frac{1}{2}\left[(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)^T \boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)\right] + \frac{1}{2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right] + \beta_0. \quad (4.103)$$

Integrating out σ^2, keeping the equation correct, and then dropping constant term gives:

$$\begin{aligned} f(\boldsymbol{z}^*|\overline{z}, \phi) &\propto \frac{1}{\left|\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}}} \frac{\Gamma(\alpha)}{\beta^\alpha} \\ &\propto \frac{1}{\left|\boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{C}}\right|^{\frac{1}{2}} \left|\overline{\boldsymbol{X}}^T\overline{\boldsymbol{C}}^{-1}\overline{\boldsymbol{X}}\right|^{\frac{1}{2}} \beta^\alpha}. \end{aligned} \quad (4.104)$$

Dropping more terms that do not depend on $\boldsymbol{z}^*$:

$$f(\boldsymbol{z}^*|\overline{z}, \phi) \propto \frac{1}{\beta^\alpha}. \quad (4.105)$$

Substituting back α and β, and rearranging:

$$\begin{aligned} f(\boldsymbol{z}^*|\overline{z}, \phi) &\propto \frac{1}{\left[\frac{1}{2}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)^T \boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi) + \frac{1}{2}\left[(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right] + \beta_0\right]^{\frac{n^*+m-k}{2}+\alpha_0}} \\ &\propto \frac{1}{\left[1 + \frac{(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)^T \boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)}{(\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0}\right]^{\frac{n^*+m-k}{2}+\alpha_0}} \\ &\propto \frac{1}{\left[1 + \frac{1}{m-k+2\alpha_0}\frac{(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)^T \boldsymbol{A}[\hat{\boldsymbol{z}}^*|\phi]^{-1}(\boldsymbol{z}^* - \hat{\boldsymbol{z}}^*|\phi)}{\left((\bar{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\overline{\boldsymbol{C}}^{-1}(\overline{\boldsymbol{z}} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0\right)/(m-k+2\alpha_0)}\right]^{\frac{n^*+m-k+2\alpha_0}{2}}} \end{aligned} \quad (4.106)$$

This final expression is proportional to the multivariate t distribution (Roth, 2013) with – in this context – degrees of freedom

$$\nu_t = m - k + 2\alpha_0, \quad (4.107)$$

mean

$$\mu = \hat{z^*}|\phi, \tag{4.108}$$

and shape matrix parameter

$$\Sigma_t = \frac{(\bar{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T \overline{\boldsymbol{C}}^{-1}(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}}) + 2\beta_0}{m - k + 2\alpha_0} \boldsymbol{A}[\hat{z^*}|\phi]. \tag{4.109}$$

Its variance is given by $\frac{\nu_t}{\nu_t - 2}\Sigma_t$ (4.110). Both mean and variance are functions of ϕ. Note that Σ_t with $\alpha_0 = \beta_0 = 0$ equals the UK variance with REML estimated $\hat{\sigma^2}$. Equation (4.110) shows that the variance slightly increases now the uncertainty in σ^2 is considered, as one would expect.

We will approximate Equation (4.92) by a finite sum using trapezoidal integration, that is:

$$f_p(z^*|\overline{z}) \approx \sum_{i=1}^{G} \frac{1}{2}\left[f(z^*|\overline{z}, \phi_i) f_p(\phi_i|\overline{z}) + f(z^*|\overline{z}, \phi_{i-1}) f_p(\phi_{i-1}|\overline{z})\right] \delta_{\phi_i}, \tag{4.111}$$

where $\delta_{\phi_i} = \phi_i - \phi_{i-1}$ is the ith spacing between consecutive values in the 1D ϕ parameter grid, $\phi_i; i = 0, ..., G$. This can be rewritten as:

$$f_p(z^*|\overline{z}) \approx \sum_{i=0}^{G} w_i \;\; f(z^*|\overline{z}, \phi_i), \tag{4.112}$$

where we define the weights:

$$w_i = \begin{cases} \frac{1}{2} f_p(\phi_0|\overline{z})(\phi_1 - \phi_0) & \text{if } i = 0 \\ \frac{1}{2} f_p(\phi_G|\overline{z})(\phi_G - \phi_{G-1}) & \text{if } i = G \\ \frac{1}{2} f_p(\phi_i|\overline{z})(\phi_{i+1} - \phi_{i-1}) & \text{otherwise} \end{cases} \tag{4.113}$$

This defines a mixture distribution for z^*, with weights given by w_i. Its mean is

$$\hat{z^*} = \sum_{i=0}^{G} w_i E[f(z^*|\overline{z}, \phi_i)] \tag{4.114}$$

and its variance is

$$var[\hat{z^*}] = \sum_{i=0}^{G} w_i \left\{E[f(z^*|\overline{z}, \phi_i)]^2 + var[f(z^*|\overline{z}, \phi_i)] - \hat{z^*}\right\} \tag{4.115}$$

Numerically, BAK first creates for each prediction point s^* a vector of predictions and a vector of corresponding prediction variances, both as 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 $f_p(\phi|\overline{z})$ and the spacings of the ϕ parameter grid, eq. (4.113)).

4.A.7 Connect to R code

Function `postc1abetaz0` as pseudo-code

- Create parametergrids for
  - ϕ (`aVec`)
  - σ^2 (`c1Vec`)
  - β (`betaMat`)
- Create discretized log prior for ϕ (`lnf0a`)
- For each ϕ in `aVec`:
  - Calculate, using function `intLikb`, the log likelihood `lnfla`. This function also returns other variables to be used in the following calculations, such as: `betahat`, `sigma2hat`, `Ac` and `resinvAres`
  - Calculate log posterior σ^2 (as grid: `lnfpc1a`) for each σ^2 in `c1Vec`
  - Calculate the marginal log posterior for each β_q (as grid: `lnfpbetaa`), using function `margbetab`, for each β in `betaMat`
  - Calculate posterior prediction (`z0IFa`) and the associated uncertainty (`v0IFa`) using function `fz0IFa`
- For σ^2, using the 2 dimensional variable `lnfpc1a`, backtransform the (log) posterior probabilities, then marginalize and normalize
- For ϕ: Backtransform and normalize vector `lnfla`
- For β_q: Backtransform, marginalize and normalize the $1 \times p$ dimensional variable `margbetab`
- Marginalize predictions $z^*(s)$ over ϕ and calculate corresponding prediction uncertainties $v^*(s)$, using `z0IFa` and `v0IFa`.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}$$

$$f(\boldsymbol{\theta}|\boldsymbol{z}) = \int f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}$$

$$f(\boldsymbol{y}^*|\boldsymbol{z}) = \iint f(\boldsymbol{y}^*,\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y}$$

$$var(\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}^*) = \sigma^2\left(\boldsymbol{1} - \boldsymbol{C}^*(\phi)^T(\boldsymbol{C}(\phi))^{-1}\boldsymbol{C}^*(\phi)\right)$$

$$p(y^*|y,\phi) = MVt_{\nu_0+n}(\mu^*, \varsigma_c^2\boldsymbol{E})$$

$$f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}\ (\varsigma^2)^{-\nu/2}$$

$$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2)$$

$$\int f(\boldsymbol{y}^*|\boldsymbol{\beta})f(\boldsymbol{\beta}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\beta}$$

$$\int f(\boldsymbol{z}^*|\overline{z},\boldsymbol{\beta},\sigma^2\phi)f(\boldsymbol{\beta},\sigma^2|\overline{z},\phi)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\sigma^2$$

$$(2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}exp\left(-\frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})\right)\ \ \eta^2 = \frac{\tau^2}{\sigma^2}$$

$$\boldsymbol{\Sigma}_\beta = (\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$$

$$f(\sigma^2) = \chi^2_{ScI}(\nu_0,\varsigma_0^2)$$

$$\hat{\boldsymbol{y}}^* = X^*\boldsymbol{\beta} +$$

$$\sim MVN($$

$$f(\phi|\boldsymbol{y}) \propto$$

$$\propto \int_\phi$$

$$f_{ILc}(\overline{z}|\boldsymbol{\beta},\phi) = \int_{\sigma^2}$$

$$|\boldsymbol{C}|^{\frac{1}{2}}\left[(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-1}(\overline{z} - \overline{\boldsymbol{X}}\hat{\boldsymbol{\beta}})\right.$$

$$f(\sigma^2|\overline{z}) \propto$$

$$-\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + \boldsymbol{\beta}^T\boldsymbol{D}^{-1}\boldsymbol{\beta} - 2\right.$$

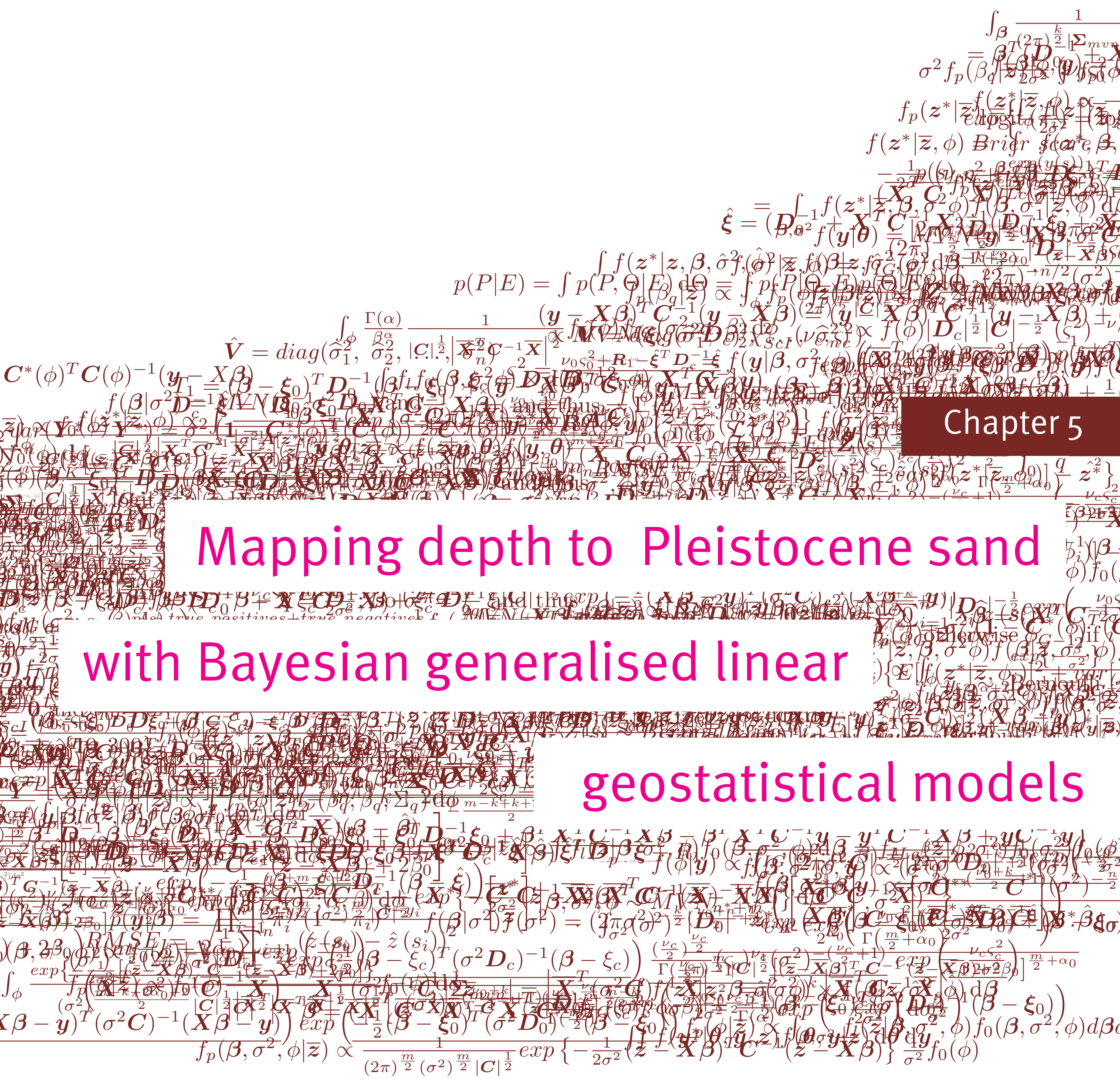
$$f_p(\boldsymbol{\beta}|\overline{z}) \propto \int_0$$

$$f_{ILb}(\overline{z}|\phi) =$$

$$f(\sigma^2|\overline{z}) \propto \int_\phi$$

$$f(\sigma^2|\overline{z}) \propto$$

$$exp\left(-\frac{1}{2}(\boldsymbol{X}\right.$$

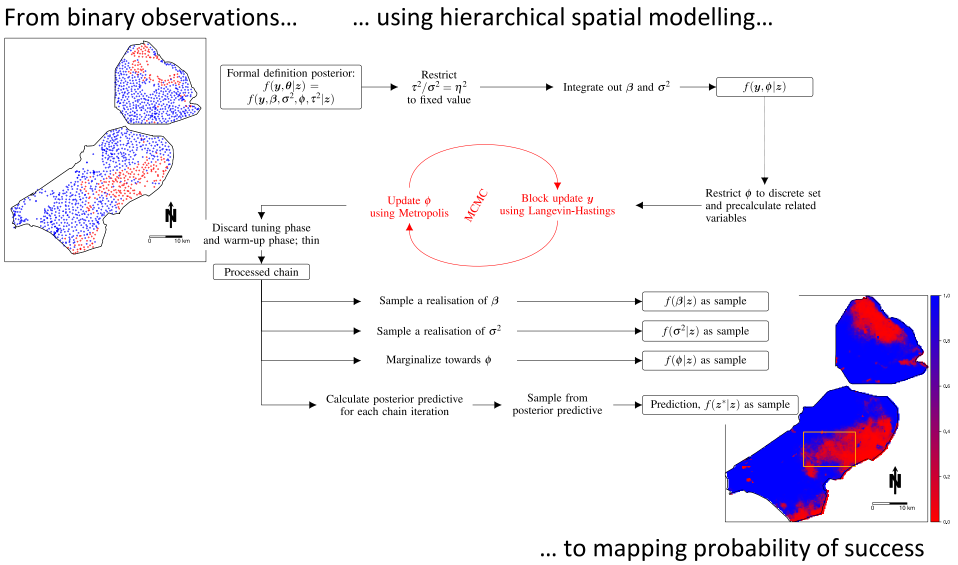


Chapter 5

Mapping depth to Pleistocene sand with Bayesian generalised linear geostatistical models

Based on:
Luc Steinbuch, Dick J Brus, Gerard BM Heuvelink. Mapping depth to Pleistocene sand with Bayesian generalised linear geostatistical models.
To be submitted to the European Journal of Soil Science

Abstract Some spatial soil applications involve binomial variables. If relevant environmental covariates are available, using a Bayesian generalised linear model (BGLM) might be a solution for mapping such discrete soil properties. The geostatistical extension, a Bayesian generalised linear geostatistical model (BGLGM) adds spatial dependence and is thus potentially better equipped. The objective of this work is to evaluate whether it pays off to extend from BGLM to BGLGM for mapping binary soil properties, evaluated in terms of prediction accuracy and modelling complexity. As motivating example, we mapped the presence/absence of the Pleistocene sand layer within 120 cm from the land surface in the Dutch province of Flevoland, using the BGLGM implementation in the R-package `geoRglm`. We found that BGLGM yields considerably better statistical validation metrics compared to BGLM, especially with – as in our case – a large (n = 1000) observation sample and few relevant covariates available. Also, the inferred posterior BGLGM parameters enable the quantification of spatial relationships. However, calibrating and applying a BGLGM is quite demanding with respect to the minimal required sample size, tuning the algorithm, and computational costs. We replaced manual tuning by an automated tuning algorithm (which eases implementing applications) and found a sample composition that delivers meaningful results within 50 hrs calculation time. With the gained insights and shared scripts spatial soil practitioners and researchers can – for their specific cases – evaluate if using BGLGM is feasible and if the extra gain is worth the extra effort.


From binary observations…
… using hierarchical spatial modelling…
Formal definition posterior: $f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{z}) = f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi,\tau^2|\boldsymbol{z})$
Restrict $\tau^2/\sigma^2 = \eta^2$ to fixed value
Integrate out β and σ^2
$f(\boldsymbol{y},\phi|\boldsymbol{z})$
Restrict ϕ to discrete set and precalculate related variables
Block update $\boldsymbol{y}$ using Langevin-Hastings
MCMC
Update ϕ using Metropolis
Discard tuning phase and warm-up phase; thin
Processed chain
Sample a realisation of β
$f(\beta|\boldsymbol{z})$ as sample
Sample a realisation of σ^2
$f(\sigma^2|\boldsymbol{z})$ as sample
Marginalize towards ϕ
$f(\phi|\boldsymbol{z})$ as sample
Calculate posterior predictive for each chain iteration
Sample from posterior predictive
Prediction, $f(z^*|\boldsymbol{z})$ as sample
N
0
10 km
1,0
0,8
0,6
0,4
0,2
0,0
… to mapping probability of success

Abbreviations

BGLM: Bayesian generalised linear model; BGLGM: Bayesian generalised linear geostatistical model; GLM: Generalised linear model; GLGM: Generalised linear geostatistical model; LH: Langevin-Hastings (algorithm); MCMC: Markov chain Monte Carlo; MVN: Multivariate normal (distribution); PI: Proportional–integral (controller)

5.1 Introduction

In many soil mapping applications, spatial variables are continuous and can, perhaps after a transformation (Müller, 2007), conveniently be modelled using Gaussian spatial models. However, some spatial soil applications involve discrete count variables or categorical variables (Kempen et al., 2012; Malone et al., 2017). In other cases, the end-user is only interested in information relative to threshold values (Lark and Ferguson, 2004). By definition, discrete and categorical variables cannot be transformed to normality. With such variables, non-linear methods such as indicator kriging (Journel, 1983) deliver reasonable results, but are suboptimal in case of a trend (Papritz, 2009) and lack model-based consistency, which implies, among others, that they are unsuited to facilitate kriging with change of support (Emery and Ortiz, 2004).

If environmental covariates are available, using a generalised linear model (GLM) might be a solution for mapping discrete soil properties. GLM assumes that the actual observations are a realisation of a discrete random process, such as Bernoulli, binomial or Poisson. Given this assumption, GLMs use, for each location, a transformation of a virtual continuous variable – the linear predictor – to the distribution parameter. For instance, the logit transformation links the linear predictor to the probability of success for a Bernoulli process. The linear predictor is the linear combination of covariate values (including an intercept) scaled by regression parameters (Myers et al., 2002, Section 4.2). For mapping purposes, GLM relies on the availability of covariates available for the whole area of interest, related to the target variable.

However, when applied to spatial data GLM ignores any spatial dependencies. The geostatistical extension, a generalised linear geostatistical model (GLGM) adds spatial dependence to the model and linear predictor, by means of a spatially correlated Gaussian field (Diggle and Ribeiro, 2007). Therefore, compared to a GLM, a GLGM is more generic and flexible, and thus potentially better equipped, provided that its spatial dependence parameters can be estimated well. However, GLGMs also require more skills of the modeller and are computationally more expensive.

Another possible extension is application of Bayesian statistics. Bayesian statistics demands explicit incorporation of pre-observation knowledge – even if this means an explicit definition of our ignorance (Lindley, 2004). It also considers all model parameters to be stochastic quantities, thus parameter uncertainty is taken into account (McElreath, 2016). Both Bayesian properties are applied in soil mapping by for example Steinbuch et al. (2018) and Poggio et al. (2016). Both properties together enable the construction of hierarchical models in a convenient and statistically sound way (Gelman et al., 2013), which can be very useful in spatial statistics (Banerjee et al., 2004). In our context, both GLMs and GLGMs have Bayesian extensions, abbreviated to BGLM and BGLGM respectively. BGLGM is a spatial implementation of a Bayesian hierarchical model (Diggle and Ribeiro, 2007).

The objective of this work is to evaluate whether it pays off to extend from BGLM to BGLGM for mapping binary soil properties, when evaluated in prediction accuracy, modelling effort and computational costs. We pay due attention to software implementation issues and provide scripts in the supplementary information, because especially a BGLGM cannot easily be programmed from scratch, thus we need to rely on existing software libraries and functions.

As motivating example, we map the depth of the Pleistocene cover sand layer relative to the surface in the Dutch province of Flevoland. This depth is defined as a Bernoulli variable obtained by determining whether the depth is above or below 1.2m.

The provincial government and the regional water board are interested in the soil composition until the Pleistocene sand layer (Brouwer et al., 2018). This is because unripened clay and peat might be present above the Pleistocene substrate, whose properties change under the influence of air. These changing properties cause a subsiding ground level (on some spots with over $2cm\ y^{-1}$), while the Pleistocene sand layer is stable. For example, in the municipality of Zeewolde subsidence as a result of ripening soils causes problems with road maintenance. Also, spatially variable subsidence will cause changes in surface water flows, and both the water board and farmers have to anticipate. A map of the depth of the Pleistocene sand layer might help to locate potential problems with land subsidence. The map might also be helpful in the foundation of construction work (such as buildings, roads and railroads).

5.2 Theory

We model a spatial process for all locations $s \in A$, where $A \subset \mathbb{R}^2$ is the study area. As first modelling step, we state that at every s, $z(s)$ is a binary realisation (zero or one) of the Bernoulli distribution

$$Z(s) \sim Bernoulli\,(p(s))\,, \tag{5.1}$$

where parameter $p(s) \in < 0, 1 >$ indicates the probability of success. We further assume that Z is an independent process, i.e. $Z(s)$ and $Z(s')$ are statistically independent for all $s, s' \in A, s \neq s'$.

This spatial process $Z = \{Z(s), s \in A\}$ is observed at n locations s_i, $i = 1, 2, \ldots, n$, providing observation vector $\boldsymbol{z} = [z(s_1) \ldots z(s_n)]^T$, being a realisation of $\boldsymbol{Z} = [Z(s_1) \ldots Z(s_n)]^T$. We model $p(s)$, and thus $Z(s)$, initially with a generalised linear model (GLM), as explained in the following section. Next we extend this GLM to a Bayesian generalised linear model (BGLM) and a Bayesian generalised linear geostatistical model (BGLGM), the two models to be compared in this research.

5.2.1 BGLM

5.2.1.1 Generalised Linear Model (GLM)

We base the explanation of GLM on textbooks such as Myers et al. (2002). The logit transformation projects $p(s)$ to the mathematically more convenient $y(s)$:

$$y(s) = \log\left(\frac{p(s)}{1-p(s)}\right) = logit(p(s)) \tag{5.2}$$

while the inverse logit function derives $p(s)$ from $y(s)$:

$$p(s) = \frac{exp(y(s))}{exp(y(s))+1} = \frac{1}{1+exp(-y(s))} = logit^{-1}(y(s)). \tag{5.3}$$

The unobservable variable $y(s), s \in A$, represents the "signal" over the study area. The vector $\boldsymbol{y} = [y(s_1)\dots y(s_n)]^T$ indicates the signal on the observation locations. In a GLM, $\boldsymbol{y}$ is modelled as a linear predictor:

$$\boldsymbol{y} = \boldsymbol{X\beta}. \tag{5.4}$$

Here, $\boldsymbol{X}$ indicates the $k \times n$ design matrix, containing leading ones and covariate values at all observation locations. The vector of k regression coefficients, including the intercept, is given as $\boldsymbol{\beta}$. Later we discuss how we can calibrate this model using the vector of observations $\boldsymbol{z}$. Note that in the GLM defined above (but not in the other models considered next), all randomness is captured in Eqn. (5.1).

5.2.1.2 Bayesian extension: BGLM

To include parameter uncertainty, we extend the GLM by considering $\boldsymbol{\beta}$ a stochastic parameter vector (but note that in this work we neither assume s nor $\boldsymbol{X}$ to be uncertain). To work conveniently with stochastic parameters we shape our models using Bayesian statistics – we assume that the reader is familiar with concepts such as Bayes' law, prior (conjugate or otherwise), conditional distributions, likelihood, posterior, the posterior predictive distribution and Markov Chain Monte Carlo (MCMC) simulation; see for example textbooks Banerjee et al. (2004), Gelman et al. (2013) or McElreath (2016). In the following, we introduce Bayesian GLM (BGLM).

The posterior density of $\boldsymbol{\beta}$, conditional on the observations is given by Bayes' rule:

$$f(\boldsymbol{\beta}|\boldsymbol{z}) \propto f(\boldsymbol{z}|\boldsymbol{\beta})f(\boldsymbol{\beta}), \tag{5.5}$$

with $f(\boldsymbol{\beta}|\boldsymbol{z})$ the posterior probability distribution of $\boldsymbol{\beta}$ conditional on $\boldsymbol{z}$, containing all our knowledge after taking the observations. The prior $f(\boldsymbol{\beta})$ contains our pre-observation knowledge; it might be a mathematical abstraction of our assumed ignorance. The likelihood of the observations conditional on model parameter $f(\boldsymbol{z}|\boldsymbol{\beta})$ follows directly from the Bernoulli probability mass function and can, while incorporating Eqns. (5.3) and (5.4), be expressed as:

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{\beta}) &= \prod_{i=1}^{n} p_i^{z_i} \{1 - p_i\}^{1-z_i} \\ &= \prod_{i=1}^{n} logit^{-1}(\boldsymbol{X}\boldsymbol{\beta})^{z_i} \left\{1 - logit^{-1}(\boldsymbol{X}\boldsymbol{\beta})\right\}^{1-z_i} . \end{aligned} \tag{5.6}$$

In Section 5.3.1 we elaborate on the choice of $f(\boldsymbol{\beta})$ and on how we implemented an algorithm to infer $f(\boldsymbol{\beta}|\boldsymbol{z})$.

5.2.1.3 Prediction with BGLM

To derive a prediction map of the probabilities of success, we first define n^* prediction locations $\boldsymbol{s}_i, i = n + 1, \dots, n + n^*$. Typically these are the nodes of a grid covering the study area, for which all covariates must be available. The prediction probabilities at the prediction locations result from backtransforming the signal at these locations using Eqn. (5.3):

$$\boldsymbol{p}^* = logit^{-1}(\boldsymbol{y}^*). \tag{5.7}$$

We therefore require the vector $y^* = [y(s_{n+1}) \dots y(s_{n+n^*})]^T$ of signals at prediction locations.

In contrast to GLM, in a BGLM y^* is stochastic. The posterior probability distribution, referred to as the ‘posterior predictive’, equals:

$$\begin{aligned} f(\boldsymbol{y}^*|\boldsymbol{z}) &= \int f(\boldsymbol{y}^*, \boldsymbol{\beta}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{\beta} \\ &= \int f(\boldsymbol{y}^*|\boldsymbol{\beta}, \boldsymbol{z}) f(\boldsymbol{\beta}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{\beta} \\ &= \int f(\boldsymbol{y}^*|\boldsymbol{\beta}) f(\boldsymbol{\beta}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{\beta}. \end{aligned} \tag{5.8}$$

Note that the last identity holds because $\boldsymbol{y}^*$is completely characterised by the covariates and $\boldsymbol{\beta}$: given $\boldsymbol{\beta}$, $\boldsymbol{y}^*$ and $\boldsymbol{z}$ are independent. In Section 5.3.1 we show the implementation of the BGLM prediction.

5.2.2 BGLGM

5.2.2.1 Generalised Linear Geostatistical Model (GLGM)

In a generalised linear geostatistical model (GLGM) (Diggle and Ribeiro, 2007; Webster and Oliver, 2007), Eqn. (5.4) is extended by adding a spatially correlated random effect:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U}. \tag{5.9}$$

Here $\boldsymbol{U}$ is a vector of random variables $U(s_i), i = 1, \ldots n$, taken from a stochastic spatial process U, which is modelled as a zero mean, stationary Gaussian random field defined over A. Note that we use again $\boldsymbol{Y}$, and its realisation $\boldsymbol{y}$, for the signal, although they are modelled differently in a GLM and GLGM.

The spatial structure of U is in this research characterised by an exponential correlation function:

$$c(\phi, h, \eta^2) = \begin{cases} 1 & h = 0 \\ \frac{1}{1+\eta^2} exp(-\frac{h}{\phi}) & h > 0 \end{cases} \tag{5.10}$$

where h indicates the Euclidean distance between locations s and $s + h \in A$, and ϕ is in geostatistical context called the 'range parameter' or distance parameter. The 'nugget to partial sill ratio' η^2 is given by

$$\eta^2 = \frac{\tau^2}{\sigma^2} \tag{5.11}$$

where parameter τ^2 captures short-distance variation, i.e. the 'nugget', while σ^2 ('partial sill' in geostatistics) is the variance of U, minus the variance of U already captured by τ^2. Note that when $\tau^2 = 0$, $\eta^2 = 0$ and the correlation function Eqn. (5.10) on very short distances approximates one.

In this research, we made simplifying choices regarding the spatial covariance function $\sigma^2 c(\phi, h, \eta^2)$, such as not to consider spatial anisotropy (c is a function of Euclidean distance only, not of direction), and not to consider other covariance functions.

Each element of the $n \times n$ correlation matrix $\boldsymbol{C}(\phi, \boldsymbol{h}, \eta^2)$ is given by the correlation function Eqn. (5.10) corresponding to the distance between two observation locations. With $\boldsymbol{C}(\phi, \boldsymbol{h}, \eta^2)$, we can extend Eqn. (5.9) with spatial parameters

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X\beta} + \boldsymbol{U} \\ &\sim MVN(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{C}(\phi, \boldsymbol{h}, \eta^2)\ ). \end{aligned} \tag{5.12}$$

For notational convenience, in what follows we often merge spatial and regression parameters into one parameter vector $\boldsymbol{\theta} = \langle \boldsymbol{\beta}, \sigma^2, \phi, \eta^2 \rangle$, and we will often indicate $\boldsymbol{C}(\phi, \boldsymbol{h}, \eta^2)$ by $\boldsymbol{C}(\phi)$ or $\boldsymbol{C}$.

5.2.2.2 Bayesian extension: BGLGM

In the Bayesian extension of the GLGM we consider $\boldsymbol{\theta}$ stochastic. $\boldsymbol{Y}$ is stochastic as well, due to $\boldsymbol{\beta}$ and the spatial signal $\boldsymbol{U}$. We therefore need their joint distribution. The joint posterior density is, according to Bayes' rule and the chain rule of conditional probability, proportional to:

$$\begin{aligned} f(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{z}) &\propto f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) f(\boldsymbol{y}, \boldsymbol{\theta}) \\ &= f(\boldsymbol{z}|\boldsymbol{y}) f(\boldsymbol{y}|\boldsymbol{\theta}) f(\boldsymbol{\theta}), \end{aligned} \tag{5.13}$$

with $f(\boldsymbol{y}|\boldsymbol{\theta})$ the proportional density of the signal conditional on the parameters. Note that in this context $f(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{z}|\boldsymbol{y})$ because conditional on $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{\theta}$ are independent. Eqn. (5.13) is often referred to as a 'hierarchical model', referring to the three related levels of probability distributions (Banerjee et al., 2004).

Eqn. (5.13) shows that three terms need to be inferred. The likelihood of the observations, conditional on the signal, follows from Eqns. (5.1) and (5.3) and is much like Eqn. (5.6):

$$\begin{aligned} f(\boldsymbol{z}|\boldsymbol{y}) &= \prod_{i=1..n} p_i^{z_i} \{1 - p_i\}^{1-z_i} \\ &= \prod_{i=1..n} logit^{-1}(y_i)^{z_i} \left\{1 - logit^{-1}(y_i)\right\}^{1-z_i} \quad . \end{aligned} \tag{5.14}$$

The likelihood of the signal, conditional on the parameters is given by the MVN probability density distribution:

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{C}) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} |\boldsymbol{C}|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T (\sigma^2 \boldsymbol{C})^{-1} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})\right), \end{aligned} \tag{5.15}$$

with $|\boldsymbol{C}|$ indicating the determinant of $\boldsymbol{C}$.

The prior $f(\boldsymbol{\theta})$ in Eqn. (5.13) is addressed in Section 5.3.2. In the next section we first extend above theory to prediction on new locations. Note that from Eqn. (5.13) we can infer the posterior of $\boldsymbol{\theta}$ by integrating out the signal:

$$f(\boldsymbol{\theta}|\boldsymbol{z}) = \int f(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{z})\, \mathrm{d}\boldsymbol{y}. \tag{5.16}$$

5.2.2.3 Prediction with BGLGM

To predict the signal $\boldsymbol{y}^*$ conditional on $\boldsymbol{z}$ within a BGLGM context, we formulate the posterior predictive distribution as an integral over both the parameters $\boldsymbol{\theta}$ and the signal at the observation locations $\boldsymbol{y}$:

$$\begin{aligned} f(\boldsymbol{y}^*|\boldsymbol{z}) &= \iint f(\boldsymbol{y}^*, \boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} \\ &= \iint f(\boldsymbol{y}^*|\boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{z}) f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y} \\ &= \iint f(\boldsymbol{y}^*|\boldsymbol{\theta}, \boldsymbol{y}) f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y}. \end{aligned} \tag{5.17}$$

Note that $\boldsymbol{y}^*$ and $\boldsymbol{z}$ are conditionally independent given $\boldsymbol{\theta}$ and $\boldsymbol{y}$. In the final expression of Eqn. (5.17), we can consider $f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{z})$ (the joint posterior density of parameters and signal as provided by Eqn. (5.13)) as weights to derive $f(\boldsymbol{y}^*|\boldsymbol{z})$ from $f(\boldsymbol{y}^*|\boldsymbol{\theta}, \boldsymbol{y})$. The other component of the integrand, $f(\boldsymbol{y}^*|\boldsymbol{\theta}, \boldsymbol{y})$, the distribution of a Gaussian response at prediction locations given observations and model parameters, boils down to simple kriging. This is because $\boldsymbol{y}^*$ and $\boldsymbol{y}$ are part of the joint multivariate normal distribution:

$$\begin{bmatrix} \boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \tau^2, \phi \\ \boldsymbol{y}^*|\boldsymbol{\beta}, \sigma^2, \tau^2, \phi \end{bmatrix} \sim MVN\left(\begin{bmatrix} X \\ X^* \end{bmatrix} \boldsymbol{\beta},\ \sigma^2 \begin{bmatrix} \boldsymbol{C}(\phi) & \boldsymbol{C}^*(\phi) \\ \boldsymbol{C}^*(\phi)^T & \boldsymbol{C}^{**}(\phi) \end{bmatrix} \right) \tag{5.18}$$

where $\boldsymbol{C}^*(\phi)$ indicates the correlations between the signals at observation and prediction locations and $\boldsymbol{C}^{**}(\phi)$ the correlation matrix of $\boldsymbol{y}^*$, both of which can be calculated using the geographic distance between the relevant locations and Eqn. (5.10). The simple kriging predictor, for a given realisation of model parameters and signal, equals a Gaussian distribution (Searle, 1997, page 47) with mean :

$$\hat{\boldsymbol{y}}^* = X^*\boldsymbol{\beta} + \boldsymbol{C}^*(\phi)^T \boldsymbol{C}(\phi)^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) \tag{5.19}$$

and with prediction error variance:

$$var(\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}^*) = \sigma^2 \left(\mathbf{1} - \boldsymbol{C}^*(\phi)^T (\boldsymbol{C}(\phi))^{-1} \boldsymbol{C}^*(\phi) \right). \tag{5.20}$$

In theory, one could calculate both the prediction probability density as provided above and the corresponding weight factor $f(\boldsymbol{\theta}, \boldsymbol{y}|\boldsymbol{z})$ in Eqn. (5.17) for every possible combination of parameters and signal at the prediction locations, and numerically integrate the outcomes to infer the posterior predictive distribution $f(\boldsymbol{y}^*|\boldsymbol{z})$. However, even if the space spanned by the model parameters and signal $\boldsymbol{y}$ would be discretised by a coarse grid this would be prohibitive because of the high dimension of the parameter-signal space. In Section 5.3.2 we discuss how to relieve the burden of computational costs and show how the actual inference of posterior parameters and posterior predictive was done in this research.

5.3 Implementation

5.3.1 BGLM implementation

We chose to apply a uniform prior $f(\boldsymbol{\beta})$, indicating our pre-observation ignorance about $\boldsymbol{\beta}$. Because $\boldsymbol{\beta}$ is a centrality parameter, this prior might be considered 'flat' and can from a mathematical point of view be left out from Eqn. (5.5). This means that the posterior $\boldsymbol{\beta}$ for BGLM is proportional to the likelihood (expressed as a function of $\boldsymbol{\beta}$) as derived for GLM. According to Kutner et al. (2005, Section 14.5) and Myers et al. (2002, Sections 4.3 and 4.4.1) this likelihood proportionally approaches a MVN distribution in case of large sample sizes, where the mean of this MVN distribution can be estimated by a maximum likelihood estimator using the iteratively reweighted least squares algorithm, and the corresponding variances and covariances are estimated via the Fisher information matrix. For these calculations, we applied the algorithm as implemented in the base function `glm` in the statistical programming language `R` (R Core Team, 2017). Above mathematical and numerical steps are illustrated in Figure 5.1.

We obtain a prediction of $\boldsymbol{p}^*$ by substituting the maximum likelihood estimate of $\boldsymbol{\beta}$ in Eq. (5.4), with $\boldsymbol{X}$ replaced by $\boldsymbol{X}^*$ and the result of that in Eq. (5.7).

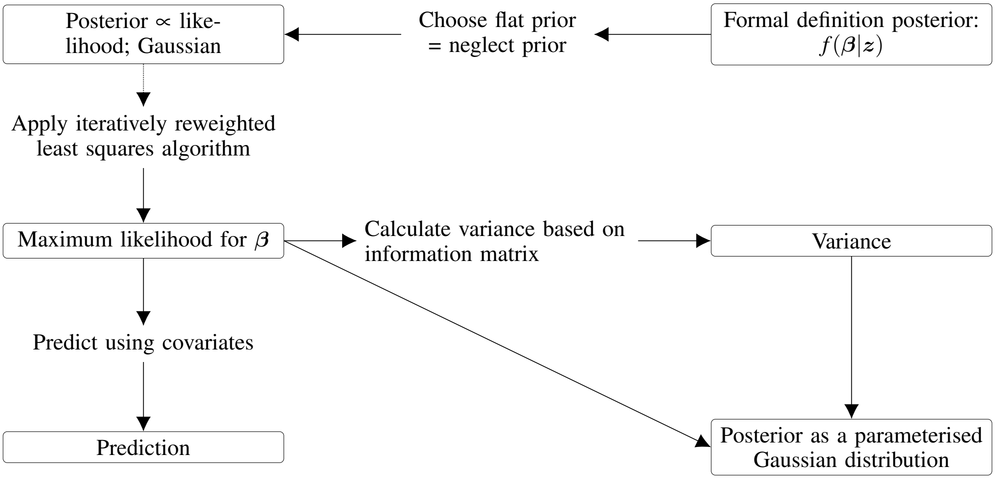

Posterior ∝ likelihood; Gaussian
Choose flat prior = neglect prior
Formal definition posterior: $f(\boldsymbol{\beta}|\boldsymbol{z})$
Apply iteratively reweighted least squares algorithm
Maximum likelihood for $\boldsymbol{\beta}$
Calculate variance based on information matrix
Variance
Predict using covariates
Prediction
Posterior as a parameterised Gaussian distribution

Figure 5.1: Workflow to infer the posterior of $\boldsymbol{\beta}$ and predict the probability on success using a Bayesian generalised linear model (BGLM).

5.3.2 BGLGM implementation

For BGLGM inference and prediction we shaped our approach around several functions from the `geoRglm` R package (Christensen and Ribeiro Jr, 2002, 2015). These functions numerically approximate the posterior distribution of signal and model parameters by simulating a large sample from the posterior. However, sampling from a posterior can be computationally expensive, especially if – as in our case – the posterior contains a signal of size n (the number of observations), making it necessary to sample from a high-dimensional combined parameter- and signal space. To lower computational costs, several extensions and simplifications were implemented by Christensen et al. (2006), to be discussed in the following sections together with several choices we made for this research. The corresponding workflow, showing analytical and numerical processing steps between the boxed representations, is illustrated in Figure 5.2. Our embedding of `geoRglm` follows in Section 5.3.2.3.

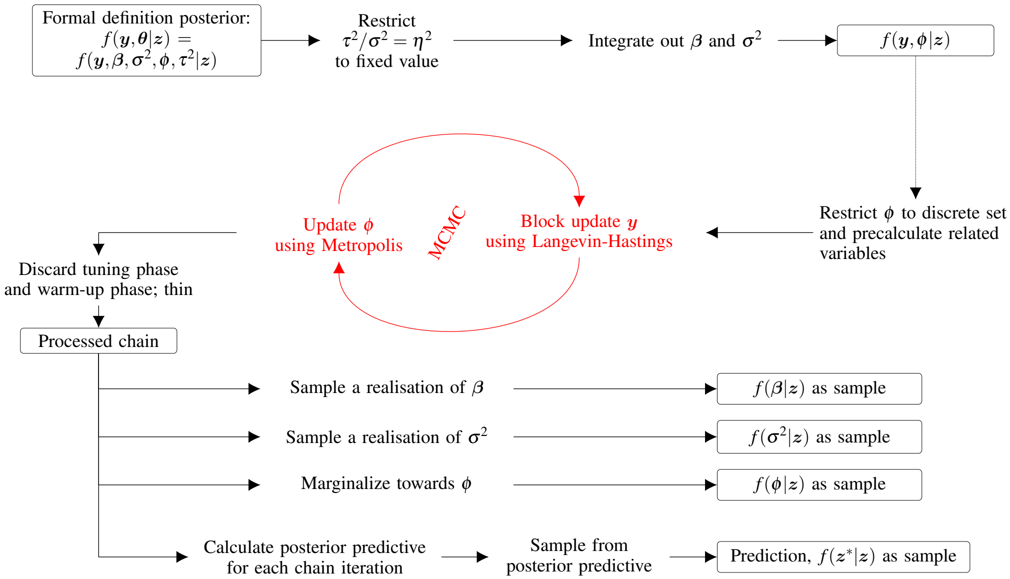

Formal definition posterior:
$f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{z}) =$
$f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi,\tau^2|\boldsymbol{z})$
Restrict
$\tau^2/\sigma^2 = \eta^2$
to fixed value
Integrate out $\boldsymbol{\beta}$ and σ^2
$f(\boldsymbol{y},\phi|\boldsymbol{z})$
Restrict ϕ to discrete set and precalculate related variables
Block update $\boldsymbol{y}$ using Langevin-Hastings
MCMC
Update ϕ using Metropolis
Discard tuning phase and warm-up phase; thin
Processed chain
Sample a realisation of $\boldsymbol{\beta}$
$f(\boldsymbol{\beta}|\boldsymbol{z})$ as sample
Sample a realisation of σ^2
$f(\sigma^2|\boldsymbol{z})$ as sample
Marginalize towards ϕ
$f(\phi|\boldsymbol{z})$ as sample
Calculate posterior predictive for each chain iteration
Sample from posterior predictive
Prediction, $f(\boldsymbol{z}^*|\boldsymbol{z})$ as sample

Figure 5.2: Workflow to sample parameter posteriors and the posterior predictive using a Bayesian generalised linear geostatistical model (BGLGM), as implemented in the `geoRglm` package (Christensen and Ribeiro Jr, 2002, 2015). The Markov Chain Monte Carlo part (MCMC, indicated in red) represents many thousands of iterations. The mentioned 'tuning phase' is our addition.

5.3.2.1 Choice of BGLGM priors; restrict nugget and range parameter

We assume that the marginal prior ϕ, the marginal prior τ^2 and the combined prior of $\boldsymbol{\beta}$ and σ^2 are independent:

$$f(\boldsymbol{\theta}) = f(\phi)f(\tau^2)f(\boldsymbol{\beta},\sigma^2). \tag{5.21}$$

We write the joint prior for $\boldsymbol{\beta}$ and σ^2 as a product of conditional and marginal densities:

$$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2). \tag{5.22}$$

Next, following ideas of Diggle and Ribeiro (2007), we assign as combined prior a multivariate normal (MVN) and inverse scaled Chi-squared (χ^2_{ScI}) distribution, respectively:

$$\begin{aligned} f(\beta|\sigma^2) &= MVN(\boldsymbol{\xi}_0, \sigma^2\boldsymbol{D}_0) \text{ and} \\ f(\sigma^2) &= \chi^2_{ScI}(\nu_0, \varsigma^2_0). \end{aligned} \tag{5.23}$$

From this we derive the distribution of $\boldsymbol{\beta}$ and σ^2 conditional on $\boldsymbol{y}$ and ϕ – $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \phi)$ – which also has a $MVN\chi^2_{ScI}$ distribution; the proof is provided in the Additional material A. Because $f(\boldsymbol{\beta}, \sigma^2)$ and $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \phi)$ have the same type of distribution, they are 'conjugate' and in this context, we consider $f(\boldsymbol{\beta}, \sigma^2)$ a 'conjugate prior'. The 'hyperpriors' $\boldsymbol{\xi}_0$, $\boldsymbol{D}_0$, ν_0 and ς^2_0 define the distribution of the prior; subscript 0 indicates that a variable is a prior- or hyperprior parameter.

For pragmatic reasons, η^2 is restricted to a fixed plug-in value. This means that we should apply external information about η^2 or try out different values and compare results. Furthermore, there are quite some computational costs involved for each possible value of ϕ. Therefore, the prior of ϕ is restricted to a discrete set of equally distanced values, thus forcing every possible ϕ in any calculation to be one of those values. For these values for ϕ, related internal variables are pre-calculated and stored.

Within above indicated constraints, we chose in our research a low-informative prior $f(\boldsymbol{\beta})$ for the regression coefficients, by stating that $\boldsymbol{D}_0^{-1} = \mathbf{0}$. This is a limit situation that causes $f(\beta|\sigma^2)$ to become improper. For $f(\sigma^2)$ and $f(\phi)$, we followed the reasoning of Berger et al. (2001) for Gaussian spatial models, which provides a reference prior composed of $f(\sigma^2) = 1/\sigma^2$ (which equals $\nu_0 = 0$, in this case a limit situation for the χ^2_{ScI} distribution whereby $f(\sigma^2)$ becomes improper) and a proper prior for ϕ, depending on the observation locations and related covariate values. An example of this reference prior $f(\phi)$ follows in the case study, among others in Figure 5.4 . We fixed the prior for τ^2 (and thus η^2) to an arbitrary zero, because we have no other information about its possible value nor its distribution, and exploring multiple values was outside the scope of this research.

5.3.2.2 Integrate out regression coefficients and variance; sample from posterior range parameter and signal

Computational costs can be substantially reduced by integrating out the regression coefficients $\boldsymbol{\beta}$ and variance of the spatial signal σ^2 (see the Additional material A), rather than be part of numerically approximating the posterior. This 'integrating out' means that these parameters as such disappear from the equation while their influence remains, expressed in the relations between the remaining variables.

The remaining $\phi, \boldsymbol{y}$ parameter- and signal space is sampled using Markov chain Monte Carlo (MCMC) (Brooks et al., 2011). In each iteration the MCMC algorithm alternates between a proposal for ϕ and a proposal for $\boldsymbol{y}$. For the one-dimensional update in ϕ, a Metropolis algorithm is applied, meaning that the proposal distribution is symmetric. For the n-dimensional block update in signal space, the Langevin-Hastings algorithm (LH) is applied (Christensen et al., 2006). LH is a special case of Metropolis-Hastings, with faster conversion: while with Metropolis the number of iterations to reach convergence is proportional to the dimension of the combined parameter-signal space (i.e., it is $O(n)$, Chivers and Sleightholme (2015)), LH convergences proportionally to the cubic root of the dimensionality ($O(n^{1/3})$) (Roberts and Rosenthal, 1998). It does so by using information captured in the spatial correlation matrix $\boldsymbol{C}$ and in the observations $\boldsymbol{z}$, to form a proposal probability gradient field (Christensen et al., 2006).

5.3.2.3 Tuning the proposal distributions, chain phases

The proposal distributions for the MCMC iteration steps of ϕ and $\boldsymbol{y}$ are respectively: 1) a normal distribution, scaled and rounded to the earlier defined discrete values for ϕ; 2) a multivariate normal distribution scaled according to the LH algorithm (Christensen et al., 2006). Each proposal distribution is scaled by a single variance parameter, that is γ_ϕ and γ_Y, respectively. These scale settings should force the acceptance rates to be as close as possible to the ideal rates: 0.44 for a single-parameter Metropolis algorithm (Rosenthal, 2011) and 0.57 for LH (Christensen et al., 2006).

Manually tuning γ_ϕ and γ_Y proved to be a very cumbersome process, because a) a test run to assess a tuning setting can take up half an hour of computation time and we need to assess several settings of two scaling parameters combined, and b) sometimes at the very beginning of a chain a low value for γ_Y is needed to achieve an acceptance rate for $\boldsymbol{y}$ larger than zero (in other words: to start exploring the parameter-signal space), while soon afterwards this value has to be increased. To overcome both issues we used a self-tuning algorithm that searches values for γ_ϕ and γ_Y, forcing acceptance rates near the optimal values as given above. Because the chain continues while trying out different settings during the first several thousand iterations, this approach supports the start up.

For the self-tuning algorithm we applied a proportional–integral (PI) feedback controller for each scale setting. Such controllers are used in many industrial processes, and are known for their robustness with respect to process modelling uncertainties (Aström and Murray, 2008) and easy mathematical description. Thus, in our research the complete MCMC chain consisted of three phases: 1) the tuning phase, where γ_ϕ and γ_Y

have to be tuned and exploration of the parameter-signal space has to start; 2) the warm-up phase, where the chain arrives in an appropriate subspace of the parameter-signal space (note that, according to Gelman et al. (2013), the expression “warm-up” is a better analogy than the often used expression “burn-in” for this phase); and 3), the production phase, where the chain explores the parameter- and signal space and thus samples the posterior. During phases 2 and 3, γ_ϕ and γ_Y are constant. The chain part from phase 3 is thinned, and ultimately used for inference and prediction.

5.3.2.4 Inferencing posterior and posterior predictive

Using Eqn. (5.16), we arrive at the marginal posteriors by integrating out the other parameters and the signal:

$$\begin{aligned} f(\sigma^2|\boldsymbol{z}) &= \iiint f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\phi\ , \\ f(\boldsymbol{\beta}|\boldsymbol{z}) &= \iiint f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\sigma^2\,\mathrm{d}\phi \text{ and} \\ f(\phi|\boldsymbol{z}) &= \iiint f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\sigma^2\,\mathrm{d}\boldsymbol{\beta}; \end{aligned} \tag{5.24}$$

where in our approach the integrations over $\boldsymbol{\beta}$ and σ^2 are done analytically while the integrations over ϕ and $\boldsymbol{y}$ are numerically approximated. To understand the analytical part, we first re-formulate the marginal posteriors for σ^2 and $\boldsymbol{\beta}$, respectively:

$$\begin{aligned} f(\sigma^2|\boldsymbol{z}) &= \iiint f(\boldsymbol{y},\boldsymbol{\beta},\sigma^2,\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\phi \\ &= \iiint f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\phi,\boldsymbol{z})f(\boldsymbol{y},\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\phi \\ &= \iiint f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\phi)f(\boldsymbol{y},\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\phi. \end{aligned} \tag{5.25}$$

In a similar way we derive for $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}|\boldsymbol{z}) = \iiint f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\phi)f(\boldsymbol{y},\phi|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}\,\mathrm{d}\sigma^2\,\mathrm{d}\phi. \tag{5.26}$$

In Eqns. (5.25) and (5.26) the joint conditional distribution $f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\phi)$ is an important building block, which we will discuss first in the remainder of this section. Then, we will discuss the implementations of Eqns. (5.25) and (5.26) as well as the implementation of $f(\phi|\boldsymbol{z})$ and conclude with a similar reasoning regarding prediction. Note that the upcoming Eqns. (5.27), (5.28), (5.30) and (5.32) are explained in the Additional material A, while only the results are provided here.

Joint conditional distribution

The joint conditional distribution $f(\boldsymbol{\beta}, \sigma^2|\phi, \boldsymbol{y})$ can, because of the earlier choice for the conjugate priors (see Section 5.3.2.1 and the Additional material A) be written as the product of:

$$\begin{aligned} f(\sigma^2|\phi, \boldsymbol{y}) &\propto \chi^2_{ScI}(\nu_c, \varsigma_c^2) \text{ and} \\ f(\beta|\sigma^2, \phi, \boldsymbol{y}) &\propto MVN(\boldsymbol{\xi}_c, \sigma^2 \boldsymbol{D}_c) \end{aligned} \tag{5.27}$$

with (Diggle and Ribeiro, 2007, section 7.2.1):

$$\begin{aligned} \nu_c &= \begin{cases} n-k & \text{if } \boldsymbol{D}_0^{-1} = \boldsymbol{0} \text{ and } \nu_0 = 0 \\ n+\nu_0 & \text{if } \boldsymbol{D}_0^{-1} \neq \boldsymbol{0} \text{ and } \nu_0 > 0 \end{cases} \\ \boldsymbol{D}_c &= (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{X})^{-1} \\ \boldsymbol{\xi}_c &= \boldsymbol{D}_c(\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{X}^T \boldsymbol{C}^{-1} \boldsymbol{y}) \\ \varsigma_c^2 &= \frac{\nu_0 \varsigma_0^2 + \boldsymbol{\xi}_0^T \boldsymbol{D}_0^{-1} \boldsymbol{\xi}_0 + \boldsymbol{y}^T \boldsymbol{C}^{-1} \boldsymbol{y} - \boldsymbol{\xi}_c^T \boldsymbol{D}_c^{-1} \boldsymbol{\xi}_c}{\nu_c}, \end{aligned} \tag{5.28}$$

where k represents the number of covariates, including an intercept. Note that subscript c is used for parameters and hyperparameters conditional on ϕ and $\boldsymbol{y}$. We reserve the expression “posterior” for parameter distributions conditional on the observations $\boldsymbol{z}$.

Marginal posterior $f(\sigma^2|\boldsymbol{z})$

Elaborating Eqn. (5.25):

$$\begin{aligned} f(\sigma^2|\boldsymbol{z}) &= \iiint f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \phi) f(\boldsymbol{y}, \phi|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\phi \\ &= \iiint f(\sigma^2|\phi, \boldsymbol{y}) f(\beta|\sigma^2, \phi, \boldsymbol{y}) f(\boldsymbol{y}, \phi|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\phi \\ &= \iint f(\sigma^2|\phi, \boldsymbol{y}) f(\boldsymbol{y}, \phi|\boldsymbol{z}) \left\{ \int f(\beta|\sigma^2, \phi, \boldsymbol{y}) \, \mathrm{d}\beta \right\} \mathrm{d}\boldsymbol{y} \, \mathrm{d}\phi \\ &= \iint f(\sigma^2|\phi, \boldsymbol{y}) f(\boldsymbol{y}, \phi|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\phi. \end{aligned} \tag{5.29}$$

For each thinned MCMC iteration we calculate $f(\sigma^2|\phi, \boldsymbol{y})$ (based on the corresponding $\boldsymbol{y}$ and ϕ) and take one sample, exploiting the fact that standard algorithms exist to sample from the well-known χ^2_{ScI} distribution. All those samples together are the empirical marginal posterior distribution $f(\sigma^2|\boldsymbol{z})$, or in other words: $\boldsymbol{y}$ and ϕ are numerically integrated out from $\iint f(\sigma^2|\boldsymbol{y}, \phi) f(\boldsymbol{y}, \phi|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}\phi$.

Marginal posterior $f(\beta|\boldsymbol{z})$

The marginal posterior $f(\beta|\phi, \boldsymbol{y})$ of Eqn. (5.26) can be defined as a scaled multivariate Student’s t distribution (see Additional material A):

$$f(\beta|\phi, \boldsymbol{y}) \propto MVt_{\nu_c+2}(\boldsymbol{\xi}_c, \varsigma_c^2 \boldsymbol{D}_c) \tag{5.30}$$

with $\nu_c + 2$ degrees of freedom, location vector $\boldsymbol{\xi}_c$ and scale matrix $\varsigma_c^2 \boldsymbol{D}_c$. Almost identical to the construction of the empirical marginal distribution for σ^2, we take one sample for $\boldsymbol{\beta}$ for each thinned MCMC iteration to compose the empirical marginal posterior distribution $f(\boldsymbol{\beta}|\boldsymbol{z})$.

Marginal posterior $f(\phi|\boldsymbol{z})$ and MCMC posterior evaluations

Elaborating on Eqn. (5.24) gives:

$$\begin{aligned} f(\phi|\boldsymbol{z}) &= \iiint f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \phi|\boldsymbol{z})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\beta} \\ &= \iiint f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \phi) f(\boldsymbol{y}, \phi|\boldsymbol{z})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\beta} \\ &= \iiint f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \phi) f(\phi|\boldsymbol{y}) f(\boldsymbol{y}|\boldsymbol{z})\, \mathrm{d}\boldsymbol{y}\, \mathrm{d}\sigma^2\, \mathrm{d}\boldsymbol{\beta} \end{aligned} \tag{5.31}$$

In this case, the empirical marginal posterior of $f(\phi|\boldsymbol{z})$ is provided by the ϕ values in the thinned MCMC iterations. To construct this MCMC chain, the conditional distribution $f(\phi|\boldsymbol{y})$ is evaluated at every MCMC iteration using the following equation (see Additional material A):

$$f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}\,(\varsigma^2)^{-\nu/2}. \tag{5.32}$$

Likewise, a multivariate t distribution (derived from Eqn. (5.15)) multiplied by Eqn. (5.14) is evaluated at every iteration for the signal part $f(\boldsymbol{y}|\phi, \boldsymbol{z})$ of the MCMC iteration (Diggle and Ribeiro, 2007, Section 7.5.4). Note that in both evaluated distributions ($f(\phi|\boldsymbol{y})$ and $f(\boldsymbol{y}|\phi, \boldsymbol{z})$), $\boldsymbol{\beta}$ and σ^2 have been integrated out.

Posterior predictive

For the posterior predictive, we start with expanding $\boldsymbol{\theta}$ in the final expression of Eqn. (5.17):

$$f(\boldsymbol{y}^*|\boldsymbol{z}) = \iiiint f(\boldsymbol{y}^*|\boldsymbol{\beta}, \sigma^2, \phi, \boldsymbol{y}) f(\boldsymbol{\beta}, \sigma^2, \phi, \boldsymbol{y}|\boldsymbol{z}) \, \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\sigma^2 \, \mathrm{d}\phi \, \mathrm{d}\boldsymbol{y}, \tag{5.33}$$

where again integration over $\boldsymbol{\beta}$ and σ^2 is done mathematically and integration over ϕ and $\boldsymbol{y}$ numerically. In the implementation, prediction is done by taking – on every thinned MCMC iteration – one sample from the following multivariate t-distribution (Diggle and Ribeiro, 2007, Eqn. (7.13)):

$$p(y^*|y, \phi) = MVt_{\nu_0+n}(\mu^*, \varsigma_c^2 \boldsymbol{E}) \tag{5.34}$$

with

$$\begin{aligned} \boldsymbol{\mu}^* &= (\boldsymbol{X}^* - \boldsymbol{C}^{*T}\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{D}_c\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \left(\boldsymbol{C}^{*T}\boldsymbol{C}^{-1} + (\boldsymbol{X}_0\boldsymbol{C}^*\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{D}_c\boldsymbol{X}^T C^{-1)}\right)\boldsymbol{y} \text{ and} \\ \boldsymbol{E} &= \boldsymbol{C}^{**} - \boldsymbol{C}^{*T}\boldsymbol{C}^{-1}\boldsymbol{C}^{*T} + (\boldsymbol{X}^* - \boldsymbol{C}^{*T}\boldsymbol{C}^{-1}\boldsymbol{X})(\boldsymbol{D}_0^{-1} + D_c^{-1})^{-1}(\boldsymbol{X}^* - \boldsymbol{C}^{*T}\boldsymbol{C}^{-1}\boldsymbol{X})^T. \end{aligned} \tag{5.35}$$

All those samples together constitute the posterior predictive $f(\boldsymbol{y}^*|z)$, which are back-transformed by the inverse logit function to the predicted probability on success $\boldsymbol{p}^*$ and then summarised, for example by calculating the median per prediction location.

5.4 Case study: Depth of Pleistocene sand layer in Flevoland

The Dutch province of Flevoland consists of two land reclamations (or polders): the Noordoostpolder (which translates literally to “North-East polder”) reclaimed in 1943, and the Flevopolder, reclaimed in two parts, 1957 and 1968 respectively. The total land area is ca. 1,400 km^2. The Noordoostpolder encapsulates two former islands: Urk (founded on boulder clay deposited in the Saalien ice age) and Schokland (founded on Holocene peat). During the last ice age (Weichselien) cover sand, an aolian deposit with a median grain size of 105-210 μm (Koster, 2009) was deposited in this area. During the Holocene, the cover sand was covered by marine deposits (silt and clay) and with peat. The depth below the land surface of this Pleistocene cover sand layer now ranges from 0 to over 16m.

5.4.1 Soil data and covariates

In 2018, the soil in the province of Flevoland was sampled at 1507 points (Brouwer et al., 2018) selected by spatial coverage sampling (Walvoort et al., 2010) over the area of interest. At each location, the depth of the Pleistocene cover sand layer below the land surface was determined. At some locations the cover sand was not encountered within augering depth, which was at least 1.20m. At these locations we have so-called right-censored observations of the cover sand depth. We transformed the continuous depth data into indicators, having value 1 if the depth of the cover sand exceeds 1.20m, and 0 else.

We used two covariates in raster format, the first being elevation as derived form the digital elevation model (DEM) of the Netherlands, more specifically the LiDAR based Dutch AHN2 (PDOK, 2020). We used the AHN2 version with 5m resolution and void filling, and re-sampled it to a 25m square grid. We chose this covariate because a DEM is often used to derive covariates for mapping soil properties. For the calibration locations, the DEM ranges from -5.4m to 8.6m, with a mean of -3.6m, relative to the reference sea level.

The second covariate is the estimated thickness of the Holocene layer according to the 3-dimensional Dutch Digital Geological Model (Gunnink et al., 2013; Dinoloket, 2020). This covariate is a legacy map of the depth to the Pleistocene cover sand. Depending on the quality of the legacy map, we expect this covariate to be a good candidate for modelling and mapping our Bernoulli variable. For the calibration locations, the thickness of the Holocene layer ranges from 0m to 8.6m with mean 2.9m.

After cleaning the data and defining 500 observation locations as ‘validation locations’, exactly 1,000 calibration locations remain, of which 782 have value ‘1’, i.e. depth to Pleistocene sand >1.20m. The removed observations contain two observations at exactly the same location as other observations and five locations without full covariate coverage. The calibration- and statistical validation locations, and the covariates are shown in Figure 5.3. Note that urban areas, swamps, waterbodies and areas where

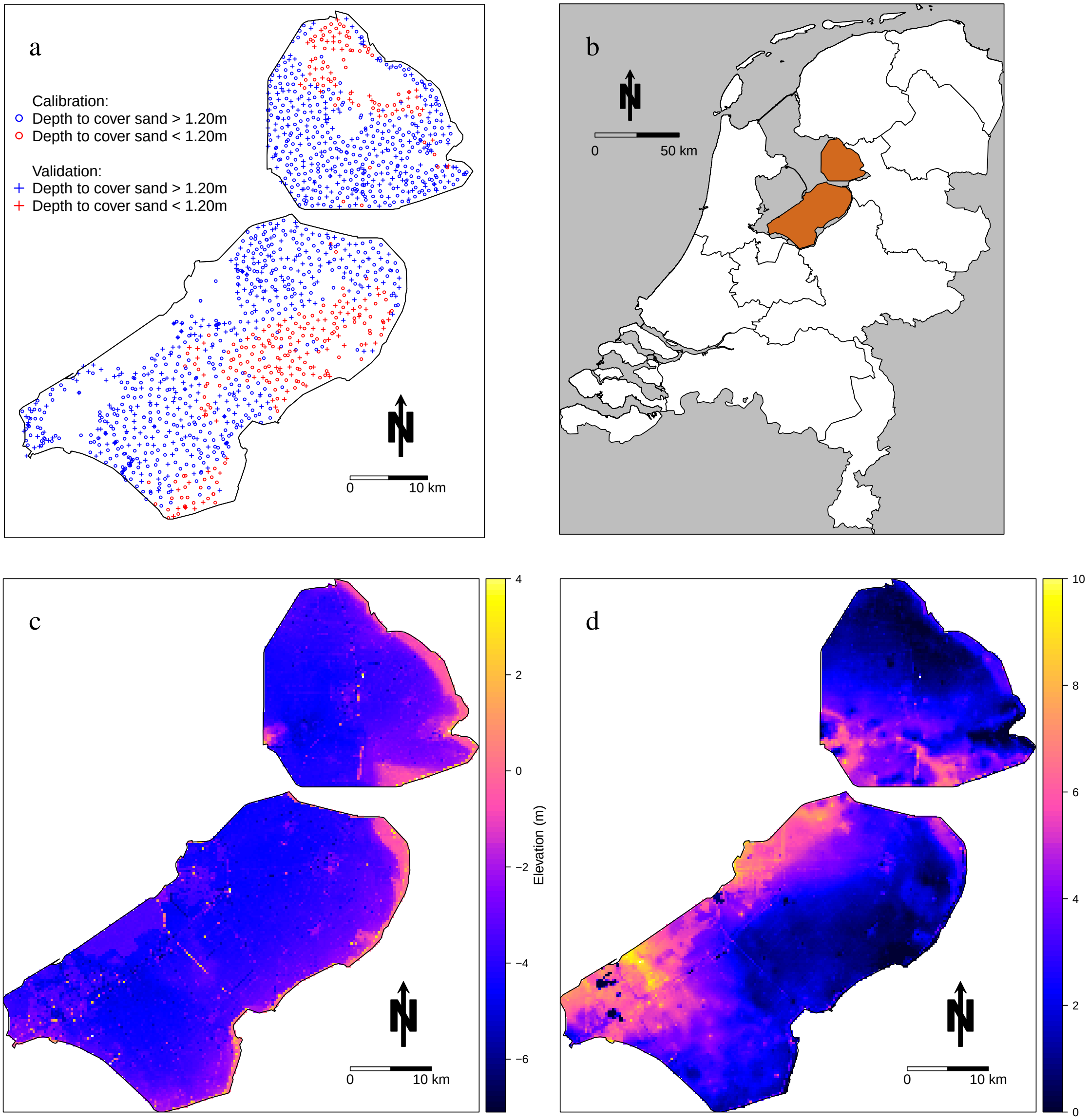

a
Calibration:
○ Depth to cover sand > 1.20m
○ Depth to cover sand < 1.20m
Validation:
+ Depth to cover sand > 1.20m
+ Depth to cover sand < 1.20m
N
0
10 km
b
N
0
50 km
c
Elevation (m)
4
2
0
−2
−4
−6
0
10 km
d
Thickness Holocene (m)
10
8
6
4
2
0
0
10 km

Figure 5.3: Case study observations and covariates. The point observation data, fully randomly divided into a calibration- and statistical validation set, is presented in subfigure a, while b shows Flevoland, our area of interest, as one of twelve Dutch provinces. The covariates elevation and thickness of the Holocene layer are presented in panels c and d, respectively

the Pleistocene sand layer is known to be at the surface were not or less intensively sampled since they are outside the area of interest (Brouwer et al., 2018).

5.4.2 Research approach

We compared four regression models: no covariates (indicated by ‘empty regression model’), elevation as covariate, thickness Holocene as covariate, and both elevation and thickness Holocene as covariates (‘full regression model’). All four regression models have an intercept. For each regression model, we used a BGLM and a BGLGM approach. For reasons explained later, each BGLGM model was fitted twice. Thus we calibrated 12 models (four BGLM models and eight BGLGM models), and assessed their performance by statistical validation for comparison. We selected the models without covariates and the models with both covariates to show the posterior distributions of model parameters and in prediction over the whole area of interest.

In case of BGLGM we aggregated (compressed) the data by constructing pairs of points based on minimal separation distance, to reduce computing time. The sum of the two Bernoulli variables was treated as a binomial variable of two trials, located at a virtual observation point halfway the two points of a pair. The binomial distribution is a straightforward generalisation of the Bernoulli distribution shown in the theory sections of this work, allowing for several “draws” or observations per location instead of one; the number of observations at a given point is denoted by m or m_i, not to be confused with the unit of distance. In our case, $m = 2$ for all virtual points. The covariate values at the individual points were averaged and assigned to the virtual observation point.

We ran the warm-up phase and production phase with 5 million iterations each, and thinned the production phase to 50 iterations for posterior inference and statistical validation. This strong thinning was needed for computational feasibility.

5.4.2.1 Chain convergence assessment for BGLGM

As we are not aware of any easy to use MCMC convergence diagnostics for high-dimensional posteriors, we applied the between-chain approach (Gelman et al., 2011): for each BGLGM under investigation, we started two chains, with different random starting values for ϕ and $\boldsymbol{Y}$, and different pseudo random number seeds, forcing the proposals of each iteration to be different between the two chains. We assessed the convergence by visually comparing the posterior densities of ϕ, $\boldsymbol{\beta}$ and σ^2 of the two chains, as well as the statistical validation metrics (i.e. measures of the quality of the spatial predictions). Because the model parameters and the statistical validation metrics depend on the signal, we assume that this approach, although indirectly, provides sufficient information to assess the convergence of inference of both ϕ and the signal.

5.4.2.2 Parameter inference and statistical validation

We show the resulting marginal posterior density plots of all available parameters of the selected calibrated models.

As already indicated, we randomly divided the original observations into a calibration set (originally $n = 1000$, for BGLGM aggregated to 500) and a validation set ($n_v = 500$). As a statistical validation metric for continuous predictions $\boldsymbol{p}_v$ – probabilities between (and including) zero and one – in combination with binary validation observations z_v – zero or one – we used the Brier score (Spiegelhalter, 2019)

$$Brier\ score = \frac{1}{n_v} \sum_{i=1}^{n_v} (p_{vi} - z_{vi})^2, \tag{5.36}$$

which is analogue to the general mean squared error statistic.

Based on an arbitrary probability threshold of 0.5, we also computed confusion matrices for each calibrated model, each matrix showing the numbers of true positives, false positives, true negatives and false negatives. We summarised these confusion matrices by the overall accuracy:

$$overall\ accuracy = \frac{true\ positives + true\ negatives}{n_v}. \tag{5.37}$$

To investigate the improvement by the recent soil survey and our models as compared to the legacy map (used as a covariate in some of our models), we also calculated the validation metrics for this legacy map. We transformed the continuous variable thickness Holocene (which is equal to the depth of the Pleistocene cover sand) according to the legacy map into an indicator applying the depth threshold of 1.20m.

5.4.2.3 Prediction

For prediction, we resampled the covariate rasters to a prediction mask resolution of $250m \times 250m$. In case of BGLM, we used the `predict` function based on a `glm` object. For BGLGM, we found that the calculation time of the prediction (additional to calculating the MCMC itself) was substantial. This prediction calculation time appeared to be roughly proportional to $O(n_*^2 n_t)$ with n_* the number of prediction locations and n_t the number of remaining MCMC samples after thinning. We separated the prediction mask into a batch of 10 small prediction masks with ca. 2,500 pixels each, and predicted each small mask with a separate run, additional to the single run mentioned in Section 5.4.2 to infer the posterior parameters. All runs are provided with equal random generator seed and calibration data. Within each run we thinned the MCMC production phase to five samples; again, this quite rigid approach in separation and thinning was needed for computational feasibility. For each of those five samples, we calculated a spatial prediction as depicted in Figure 5.2. For each prediction location, we selected the median of the five predictions and finally merged the small prediction masks into the complete prediction map.

5.4.3 Results

5.4.3.1 General performance and computational costs

Comparing the calculation approaches, we found that BGLM always and swiftly produces an answer. On the contrary, we had initially quite some problems with getting BGLGM to work at all. A first hurdle was that we obtained MCMC chains with acceptance rates very close to zero or one, meaning that the chain does not, or does very slowly, explore the parameter- and signal space. This was due to difficulties with finding a functioning MCMC proposal setting γ_ϕ and γ_Y, and probably also with the need for different proposal settings during the initial iterations, compared to the subsequent iterations. Those problems were solved for all but a few of our many (not shown) trials using the tuning algorithm as described in Section 5.3.2.3. Given a proper working MCMC algorithm with reasonable acceptance rates, the second hurdle was to produce meaningful results, such as convergence to reasonable posteriors. We found that enough data, and in these data enough (detrended) spatial structure was needed, although in our case we also found that convergence problems due to weak spatial structure could be solved by increasing the MCMC chain length.

The BGLMs produced answers within a few seconds of calculation time, while each single BGLGM run had a calculation time of 30-50 hours, using contemporary computer technology optimised for speed, one processing core and sufficient working memory.

5.4.3.2 Posterior parameters

Posteriors of regression- and spatial parameters are shown for the empty regression model and the full regression model, using the BGLM and the BGLGM models, see Figures 5.4 and 5.5 respectively. The regression coefficient distributions for elevation are around zero for both the BGLM and the BGLGM models. In case of BGLGM, the distributions of $\boldsymbol{\beta}$ are wider compared to those obtained with BGLM. In the empty regression model the posterior of β_0 is even almost bi-modal, with one peak above $\beta_0 = 0$. In case of the BGLGM models, the posteriors of β_1 and β_2 have more probability mass away from zero than those obtained with BGLM. The posterior of spatial parameter σ^2 has quite some probability mass at the higher values in case of the empty regression model. With the full regression model, this probability mass moves closer to zero. The median of the posterior of σ^2 for the two BGLGMs with empty regression model equal 47 and 57, while for the two BGLGMs with full regression model these are 6.8 and 5.4. Posterior ϕ shows the same movement: the medians are 29, 30 for the empty regression models, and 8.3, 7.0 km for the full regression models.

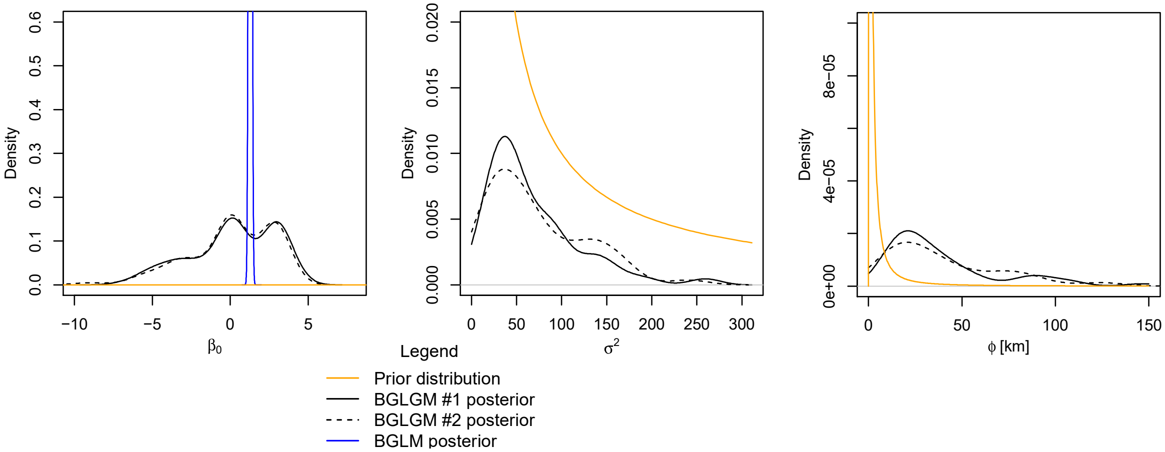


Figure 5.4: Posterior parameter distributions of BGLM and BGLGM models with empty regression model (intercept β_0 only). Note that the BGLM model only has a regression coefficient (no spatial parameters). Note also that prior σ^2 cannot be normalised because it is improper, in other words: its surface under the curve cannot be scaled to one. For each BGLGM, the posteriors of two separate runs (different in random seed) are given, allowing to explore their convergence.

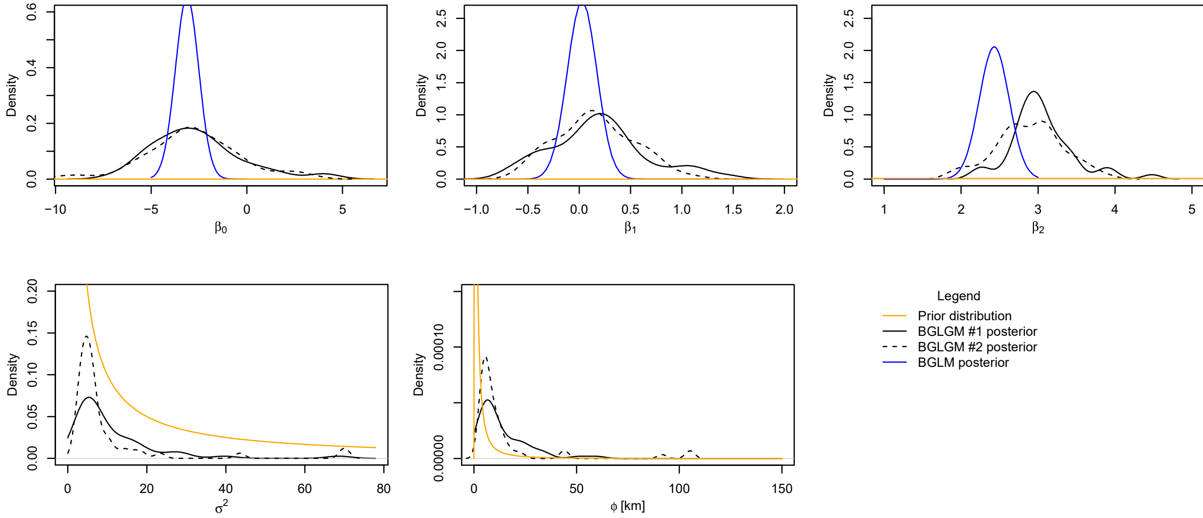


Figure 5.5: Posterior parameter distributions of BGLM and BGLGM models with full regression model. Regression coefficients are: β_0 – intercept; β_1 – elevation; β_2 – thickness Holocene.

5.4.3.3 Statistical validation

The validation metrics are given in Table 5.1, and the confusion matrices in Figure 5.6 (next page). It is clear that all BGLGMs outperform all BGLMs according to both metrics. Note also that BGLGM extracts a bit more information out of the elevation covariate. See, for example, the Brier score: adding elevation to the BGLM regression model does not improve the performance, while with BGLGM it does. Based on the overall accuracy, the full BGLGM model perform worse than both single covariate regression models, and almost equal to the empty model.

There are minor differences between the two different runs of the BGLGM models, especially according to overall accuracy, with a maximum deviation of 0.008 in case of the elevation regression model.

Compared with the legacy map (thickness Holocene), the overall accuracy of every BGLGM map is higher, and only the accuracies of the BGLM maps without covariate thickness Holocene in the regression model are lower. This is confirmed by the Brier scores.

The confusion matrices confirm that, in case of BGLM, adding elevation to the empty regression model or to the regression model with thickness Holocene does not change anything. Because almost 80% of the calibration data have value 1 (see Sect. 5.4.1), a BGLM model based on just taking the mean of the observations results in a spatially constant prediction larger than 0.5, and thus to an overall accuracy of around 0.8 – and also to around 20% false positives.

Table 5.1: Statistical validation of all models according to the Brier score (smaller is better) and overall accuracy (larger is better). For each BGLGM, the figures of two separate runs (different in random seed) are given, allowing to explore their convergence.

	Brier score		Overall accuracy	
	BGLM	BGLGM	BGLM	BGLGM
no covariates	0.161	0.064, 0.063	0.798	0.912, 0.914
elevation	0.161	0.059, 0.058	0.798	0.916, 0.924
thickness Holocene	0.089	0.059, 0.057	0.874	0.920, 0.924
elevation + thickness Holocene	0.089	0.058, 0.055	0.874	0.914, 0.914
Covariate thickness Holocene as direct binary predictor	0.128		0.872	

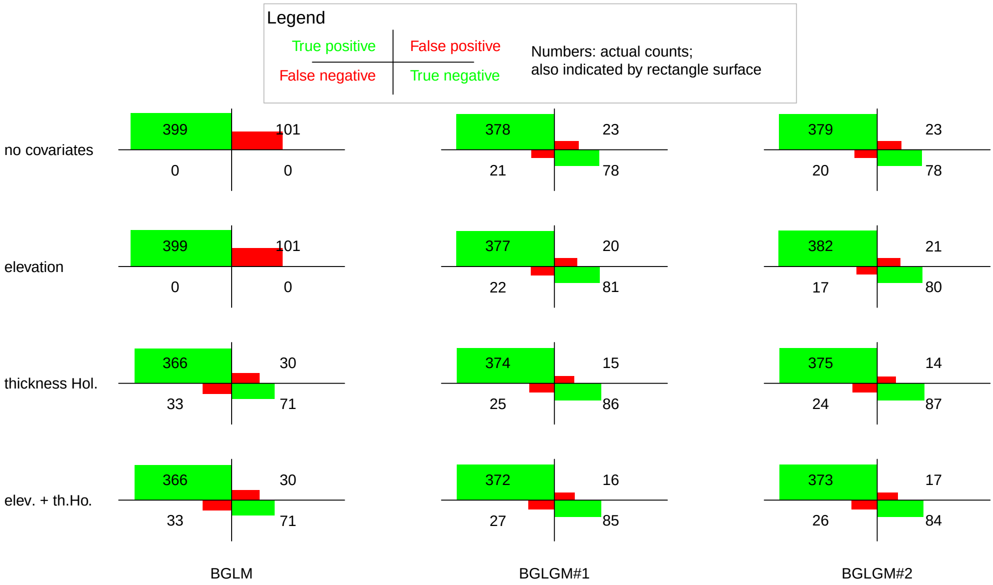


Figure 5.6: Confusion matrices for all eight models, based on the validation data set. For each BGLGM, the matrices for two separate runs are given. “Positive” means: the depth to the Pleistocene cover sand exceeds 1.20m.

5.4.3.4 Predictions

The predictions for two full models are presented in Figure 5.7 where the orange rectangle indicates the detail area of Figure 5.8 (next page) with the same predictions. For BGLGM, we used run #1.

Recall that the northern area is called Noordoostpolder and the southern unit Flevopolder. In both maps, we can see the effect of the covariates elevation and thickness Holocene layer (compare with Figure 5.3), for example in the clearly distinctive spots with low probability in the very south-west of the Flevopolder. This may be an artefact in the urban area of the city of Almere. The prediction by BGLM is in general somewhat smoother; especially in the very south of the Flevopolder. BGLGM allows predicted probabilities much closer to zero than does BGLM, while in large parts of the Noordoostpolder BGLGM predicts probabilities closer to one compared to BGLM. In the north of the Flevopolder, the BGLGM map shows an almost circular pattern with low probabilities that is hardly visible in BGLM.

Relating the prediction to the observations (Figure 5.8), BGLGM follows the observations in the calibration data closer – both spatially and in probability.

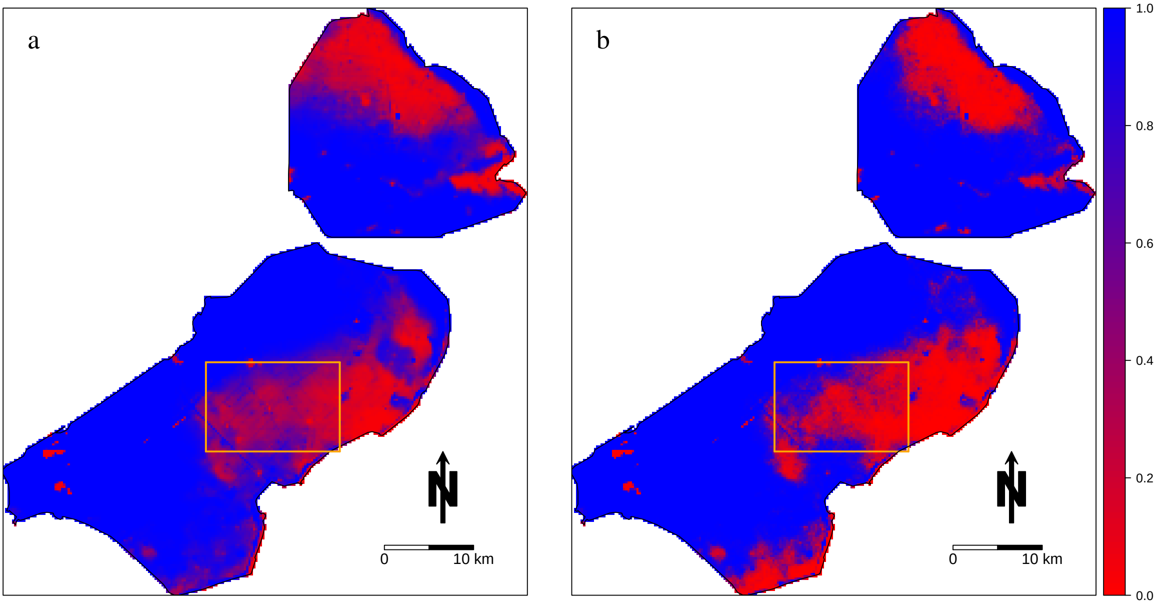


Figure 5.7: Predicted probabilities of depth of Pleistocene sand > 1.20m according to BGLM (subfigure a) and BGLGM (b) using the full regression model, using a 250m resolution prediction mask. The rectangle is the location of the detail area, shown in the next figure.

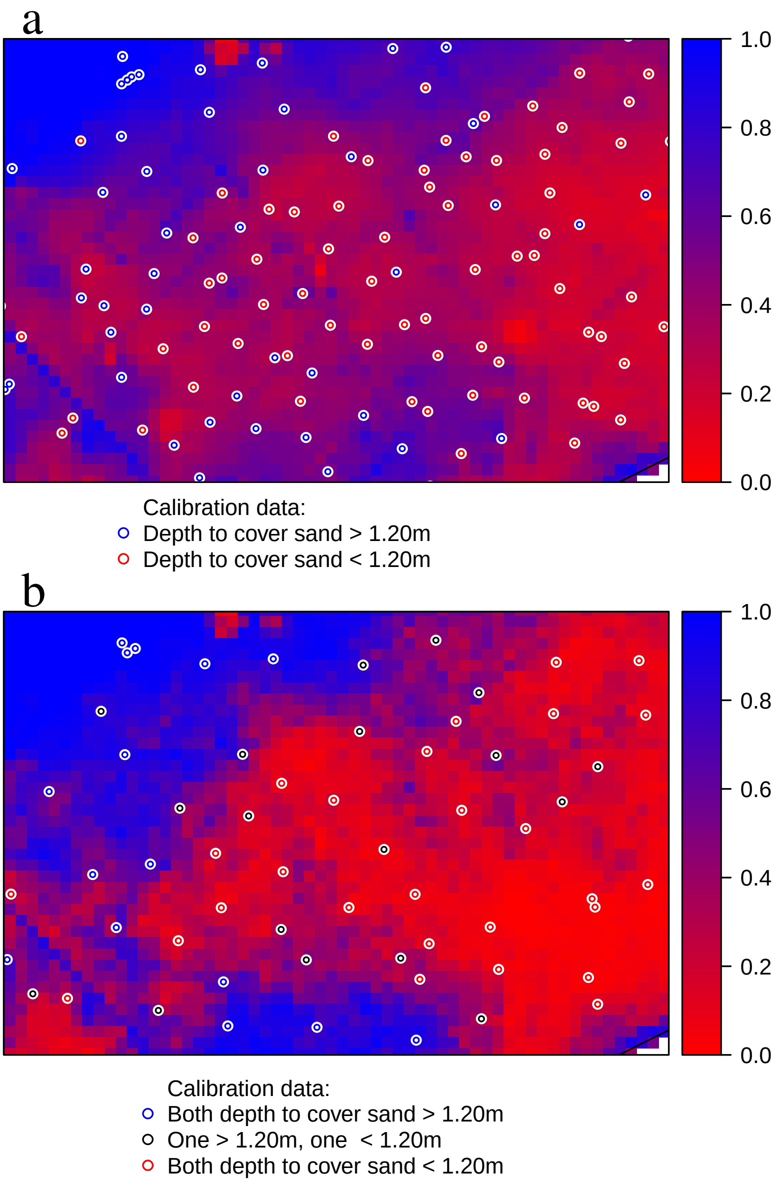

a
1.0
0.8
0.6
0.4
0.2
0.0
Calibration data:
○ Depth to cover sand > 1.20m
○ Depth to cover sand < 1.20m
b
1.0
0.8
0.6
0.4
0.2
0.0
Calibration data:
○ Both depth to cover sand > 1.20m
○ One > 1.20m, one < 1.20m
○ Both depth to cover sand < 1.20m

Figure 5.8: Detail map of predicted probabilities. Subfigure a shows the prediction according to BGLM, b the prediction according to BGLGM. For reference, the original calibration data is added in a. The location of the calibration data for BGLGM, aggregated from the original observations and assuming two observations on every (artificial) location, is added to subfigure b.

5.5 Discussion

5.5.1 Mapping Pleistocene sand layer depth

All maps confirm the general spatial pattern depicted on the covariate map thickness Holocene, showing several areas where the Pleistocene cover sand is close to the surface. The areas where the Pleistocene sand is deeper reflect old fluvial patterns, e.g. the deltas formed by the ancient ancestors of the current rivers Overijsselse Vecht and Gelderse IJssel (Brouwer et al., 2018) and probably also the Eem. The models predict also in the urban areas, the swampy areas, and areas otherwise excluded from the sampling design, lacking both calibration and validation data. One of these excluded parts is in the area just below the very north of the Noordoostpolder. Here, the predicted probability of Pleistocene cover sand deeper than 1.20m correctly approaches zero. The maps indicate (although not conclusively) which areas are most susceptible to soil subsidence (Brouwer et al., 2018), so where a closer look into the soil properties would make sense. The same methodology could also be used to make a map of the probability of Pleistocene cover sand at the surface. The statistical validation data of these maps suggest that we should have more trust in Figure 5.7b. Also visually, it seems that the BGLGM prediction in Figure 5.7b follows the observations as provided in Figure 5.3 more closely. We should be careful with the low probability dots in the southwest of the Flevopolder as shown on both maps, because they are caused by the locally deviating covariates values (very probably artefacts) and not from the observation data.

As a general note: With the given original dataset, other approaches which take all available information into account (i.e.: without simplification to binary observations) are also possible. Those approaches are however outside the scope of this research.

5.5.2 Posterior parameters

5.5.2.1 Posterior regression parameters

The posteriors of the regression coefficients were calculated differently in BGLM and BGLGM. For BGLM, we used straightforward mathematics and a numerical algorithm converging within a limited number of iterations. This was possible because of our choice for a low-informative prior we could borrow well-developed existing concepts based on non-Bayesian statistics (i.e. estimated maximum likelihood and the corresponding estimated variances). In case of BGLGM, the posterior is a sample generated from a MCMC chain. We suppose that extending a BGLM model with variogram parameters and a signal so that a BGLGM model is obtained, allows for wider posteriors of the regression parameters and even for a non-Gaussian posterior – see β_0 in Figure 5.4. The empty regression model resulted in a very narrow BGLM posterior centred at the logit transformed mean of the observations, while the corresponding BGLGM posteriors explored a much wider range of values. We could not explain why the two posterior BGLGM modes consistently are around, but different from, the mode of the posterior BGLM. In the full model (Figure 5.5), the modes of the intercepts β_0 of BGLM and BGLGM are about equal; more freedom for $\boldsymbol{\beta}$ results in larger modes for β_1 and β_2, suggesting that BGLGM gets more information out of the covariates than BGLM. This was already suggested by the statistical validation. Perhaps this is due to the covariance structure or to the hierarchical nature in BGLGM.

5.5.2.2 Posterior spatial parameters

The influence of the regression model on the spatial parameters, as shown in Section 5.4.3.2, indicates that the covariates explain part of the variation (i.e., the probability mass emphasises lower values of σ^2 if covariates are included) and explain also part of the spatial structure at larger distances (i.e., the probability mass emphasises lower values of ϕ after including covariates). It would be interesting – in follow-up research – to explore if the resulting variogram slopes close to the origin are different when comparing the empty regression model with the full regression model. It would also be interesting to investigate the correlation between the spatial parameters σ^2 and ϕ for a given regression model by exploring the joint posterior distribution $f(\sigma^2, \phi|z)$ which is especially in a Bayesian setting like this fairly easily to do.

5.5.3 Issues concerning BGLGM

5.5.3.1 Spatial parameters and signal

The applied reference prior (Berger et al., 2001) is based on a Gaussian spatial dataset, not on Bernoulli observations; within the scope of this work we could not determine if this is mathematically correct. Note that by definition, the reference prior is designed to maximise the expected impact of the data on the posterior (Bernardo, 1979; Berger et al., 2009). This means that the ϕ component of the spatial reference prior uses the geographic positions of the observation locations. We calculated the reference prior using the aggregated locations, but using the original locations for calculation would also have been possible; we are not sure which approach is mathematically preferable. We found a stabilising effect on the chain convergence when using the discretised reference prior for ϕ rather than the often applied (and in our opinion often mis-used) discrete uniform distribution. Because the software only allowed discretisation of prior ϕ in regular steps, which has a limit of 2000 discretisation points, and fine steps are needed for the small values of ϕ, we were not able to explore the posterior probability mass for higher values of ϕ.

With trial datasets lacking enough information (elaborated in Section 5.5.3.2), the posteriors of σ^2 and ϕ tended to get shapes highly dependent on the chosen random number generator seed, sometimes with most probability mass around very high values (in case of σ^2) or close to the upper limit (in case of ϕ). Those posteriors did not resemble the priors, as we – perhaps a bit naively – initially expected based on our knowledge about single parameter Bayesian models, where in case of lack of data (or: a weak likelihood) the prior is shaping the posterior. In similar trials, the signal layer values at the observation locations tended to go to its extremes $-\infty$ and ∞, resulting in predicted probabilities of 0 and 1. We considered all those outcomes as 'not meaningful'. This behaviour of the signal layer values also indicates that one should be careful when comparing hierarchical models based on goodness-of-fit metrics. Also, in those trials it proved more difficult to correctly tune the proposal distributions.

Fixing τ^2 (or actually the nugget-to-sill ratio η^2) to zero was an arbitrary choice; a better approach with the available software would have been to test several values for η^2 as proposed by Diggle and Ribeiro (2007, Section 7.2.1), and compare the associated statistical validation data. This was outside the scope of our research question. The same holds for testing other spatial covariance models, such as Mátern, which would add more choices or an extra parameter.

5.5.3.2 Sample size and information

Although not extensively explored, we have strong indications that enough information in the data is needed to get meaningful outcomes (see Section 5.5.3.1) in case of BGLGM, which is also suggested by Diggle and Ribeiro (2007). Enough information is composed of: 1) enough observations locations, 2) enough observations (or 'binomial draws') per location in cases where this applies, and 3) enough spatial signal in the actual observations. In our case, the algorithm was less sensitive to lack on 3) if we applied a longer MCMC chain but we could not determine if this is generally true, nor if this also solves possible problems with 1) and 2).

For numerical reasons the number of locations n is limited in practice (Ribeiro Jr et al., 2003), basically due to variance-covariance matrix inversions. With binomial data, information can be added by collecting data from more locations (which increases the computational costs) or by collecting data with more observations per location – in other words: increase m_i (which does not affect the computational costs). For given m_i, this suggests that there is a window of opportunity for n: too low n contains not enough information, too high n is computationally not feasible. Solutions specifically for big n are being researched (Zhang et al., 2018). In real world applications, a large n might also be unfeasible because of practical and financial constraints.

In our real world case, we aggregated $n = 1000$ locations with one observation each to $n_a = 500$ artificial locations with two observations each, to be able to offer enough information to the BGLGM algorithm while calculations were still feasible. In Figure 5.8 the calibration data locations in subfigure a or b, compared to the locations in subfigure c show the difference for the detail area. According to Hodge and Vieland (2017), aggregating binomial data (in their non-spatial context called 'compression') causes quantifiable loss of information. In our case, we assumed that this loss is limited to the very short distance spatial information and is negligible – especially because we already fixed the nugget and the covariance model for reasons explained earlier.

A formal definition and quantification of 'information' in spatial binomial data is outside the scope of this research, but would be a useful tool to assess the feasibility of model-based geostatistical approaches such as BGLGM for a given dataset or sampling design. In case of $m_i = 1$ (Bernoulli) we expect the number of observation locations required to arrive at meaningful results to be much larger than in case of the usual Gaussian geostatistical model, because much less information is available at each single location and because we need to calibrate a more complex (hierarchical) model. Possible starting points for such research offer Li and Reynolds (1995); Mays et al. (2002); Nowosad and Stepinski (2019).

5.5.3.3 Prediction and calculation time

As already indicated in Section 5.4.2.3, our numerical calculations showed that the calculation complexity for prediction was roughly $O(n_*^2 n_t)$, perhaps because a set of linear equations has to be solved which has a complexity of at least $O(n_*^{2.376})$ (Bae et al., 2014). Because prediction adds substantially to the total calculation time, we chose – as already explained in Section 5.4.2 – to thin the MCMC chain to 50 iterations in order to predict at the 500 validation locations, while inferring the posterior parameters within the same run. For the prediction itself – see Section 5.4.2.3 – we ran 10 parallel sessions in which we thinned the chain to 5 iterations only. This is another example showing that, compared to BGLM, BGLGM might offer better results but also requires much more modelling effort and much more computational costs. In practice we must accept compromises like prediction based on just five MCMC iterations (after thinning) to make calculations within a reasonable amount of time.

5.5.3.4 Tuning the proposal distribution

To our knowledge this is the first time a proportional-integrative (PI) feedback controller is applied to tune proposal distributions for an MCMC, although the use of separate 'pilot runs' for tuning is not unknown (Griffin and Walker, 2013). Far more attention received adaptive MCMC where the proposal distribution is continuously adjusted during the MCMC chain development, based on the posterior density (Griffin and Walker, 2013; Roberts and Rosenthal, 2009; Garthwaite et al., 2016). Note that the applied Langevin-Hastings algorithm also has adaptive properties. Because we chose to use an existing implementation of BGLGM, an external feedback controller together with a separate tuning phase was *de facto* a necessity, and a PI controller was a natural choice for reasons already mentioned.

5.5.3.5 Chain convergence assessment

Cowles and Carlin (1996) offer an overview of MCMC chain convergence diagnostics, with metrics generally based on between-chain and/or within-chain comparison of variances. For multivariate MCMC chains, those variances are extended to variance-covariance matrices (Brooks and Gelman, 1998). A slightly different approach would be to compare sample distributions using the Kullback Leibler divergence, tested by Dixit and Roy (2017) on a 10-dimensional parameter space. We are not aware of any work considering convergence diagnostics in a space of substantially higher dimension. In the scope of our research, we considered the applied visual comparison of the results of the two runs as sufficient, but for future development of BGLGM applications we encourage research on methods to assess high-dimensional convergence. Note however that given a finite MCMC length and a finite number of MCMC chains there is no guarantee that all local modes are being explored, even if a good convergence is indicated.

5.5.4 Statistical validation

The threshold of 0.5, used for computing the confusion matrices and the overall accuracies, is a fairly arbitrary choice. Additionally, the overall accuracy metric assumes that the penalties of a false negative and a false positive conclusion are equal. As a validation metric without threshold we chose the Brier score because it is very straightforward, and also because in earlier work we had disappointing experiences with the often applied Area Under Curve metric (Steinbuch et al., 2018). Note that in case of an unbalanced statistical validation dataset (meaning: containing many zeros and few ones, or vice versa), or with non-uniform penalties on prediction error, the Brier score as means of model comparison can be counter-intuitive, while other metrics might offer more flexibility (Assel et al., 2017; Jewson, 2004). Statistical validation of probabilistic predictors in the context of soil mapping is discussed by Rossiter et al. (2017) and Beaudette (2020), including alternative metrics. However, because of our focus on model comparison rather than on real-world applications where penalties indicate practical consequences, we assume that the quite consistent message from the two applied metrics – overall accuracy and Brier score – already provides a useful answer to our research questions. We did not calculate any other validation metrics. The only inconsistency is the drop in overall accuracy from the regression model with the thickness Holocene as covariate to the full model in case of BGLGM; this drop is not reflected in an increased Brier score. Based on the overall accuracy, we suspect some overfitting of the data by the elevation covariate model.

5.5.5 Comparison of the models, and alternative approaches

In our research, BGLM performed reasonable, see Table 5.1. However, for this it leaned heavily on the thickness Holocene covariate which is based on the 3D geological model as described in Gunnink et al. (2013).

The covariate elevation, often included as covariate in digital soil mapping, hardly added predictive power to any of the models. We should also note that the DEM is already used in the 3D geological model.

Based on the validation statistics (Table 5.1), BGLGM clearly outperformed BGLM. But as indicated in Section 5.4.3.1, it was quite an effort to get BGLGM functioning properly, and when it worked there were serious computational costs involved, with an associated electrical energy footprint (Taffoni et al., 2019) and a waiting time up to 50 hours until results were available. For BGLGM, we followed the approach and implementation of Diggle and Ribeiro (2007) and Christensen (2002). Comparable approaches in the R-universe such as `geoCount` (Jing and De Oliveira, 2015), `spBayes` (especially function `spMvGLM` (Finley et al., 2015)), `geoBayes` (Evangelou and Roy, 2019), `PrevMap` (Giorgi et al., 2017) and approximations such as `INLA` (Rue et al., 2009) – eventually embedded in `geostatsp` (Brown et al., 2015) – were outside our scope, but might be interesting to consider in future research and practical applications. We also did not look into more pragmatic approaches such as indicator kriging, creating a Gaussian model with the available data or data-mining methods such as random forest (Heung et al., 2016) (in combination with additional covariates), all of which have their merits and drawbacks in case of a binomial spatial variable.

A decision on which method to use should depend on the available or feasible sample size and the available covariates, on the final goal of the map and/or other deliverables such as the posteriors, uncertainty quantification and – more pragmatically – on the available resources. Often, in digital soil mapping many covariates are used, some based on direct observation (such as DEM and its derivatives as generated by radar and vegetation data visible on multi-spectral satellite images) and some based rather on modelling and interpolation (such as climate data). In case few covariates related to the variable of interest are available, we need to have methods at hand that deliver reasonable mapping results based only on the sample itself. In such cases, indicator kriging – although statistically not completely sound – might be a good approach. If we want to add already existing knowledge not captured in covariates, a Bayesian approach might be a good option. If we, from a pedometrics' viewpoint, are interested in a solid statistical description of binomial soil properties over space, an approach such as BGLGM might very well be worth the effort. Furthermore, the addition of an automated tuning phase as described in Section 5.3.2.3 makes both methodological research and practice directed data processing much more feasible than manual tuning would allow.

5.6 Conclusions

Adding a geostatistical component to a Bayesian generalised linear model (BGLM) for mapping binary soil properties yielded considerably better statistical validation metrics in a case study with a large ($n = 1000$) observation sample and few relevant covariates available. However, the resulting Bayesian generalised linear geostatistical model (BGLGM) is quite demanding with respect to sample size, tuning the algorithm, and computational costs. In this study we focused on the implementation of BGLGM as provided in the R-package `geoRglm`, which we embedded in our own scripts. We replaced manual tuning by an automated tuning algorithm and found a sample composition (in size and in number of observations per location) that delivers meaningful results within 50 hours calculation time. With the gained insights spatial soil practitioners and researchers can – for their specific cases – better evaluate if using BGLGM would be possible at all and if the extra gain would be worth the extra effort. The developed automated tuning algorithm (of which the code is available) makes implementation of BGLGM in applications more easy.

5.A Additional material A

5.A.1 Introduction

In this additional material we show: 1) that the proposed $MVN\chi^2_{ScI}$ prior for $\boldsymbol{\beta}$ and σ^2 is indeed conjugate; 2) how to arrive at the marginal conditional distributions $f(\boldsymbol{\beta}|\boldsymbol{y})$ and $f(\sigma^2|\boldsymbol{y})$; 3) one of the possibilities to integrate out $\boldsymbol{\beta}$ and σ^2 to arrive at the marginal distribution $f(\phi|\boldsymbol{y})$. Most of this material is based on Diggle and Ribeiro (2007); Christensen (2004); Diggle et al. (2003); Ribeiro and Diggle (2006); Ribeiro Jr et al. (2003); Diggle and Ribeiro (2002); we also draw inspiration from Steinbuch et al. (2020, Additional material A). For general knowledge about Bayesian statistics we relied – among others – on Gelman et al. (2013) and O'Hagan and Forster (2004). We aim this material to readers with basic knowledge about probability distributions, conditional probabilities, integration and matrix algebra. For the meaning and context of most symbols we refer to the main paper; this document isn't meant to be read stand-alone. In the remaining of this section we provide the used distributions, an integral and assumptions.

We will extensively use the multivariate normal, also known as multivariate Gaussian distribution (MVN), in general terms described as (Gelman et al., 2013, page 578):

$$MVN(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}exp\left(-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{y})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{y})\right) \tag{5.38}$$

where $|\boldsymbol{\Sigma}|$ indicates the determinant of the enclosed matrix, and n the dimension of the square, symmetric and positive-definite matrix $\boldsymbol{\Sigma}$. The parameter before the semicolon – $\boldsymbol{y}$ – indicates the variable whose distribution is given. We will also use the scaled inverse Chi squared distribution (χ^2_{ScI}) (Gelman et al., 2013, page 578):

$$\chi^2_{ScI}(\sigma^2;\nu,\varsigma^2) = \frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}(\sigma^2)^{-(\frac{\nu}{2}+1)}exp\left(-\frac{\nu\varsigma^2}{2\sigma^2}\right), \tag{5.39}$$

and use the multivariate Student's t distribution (MVt) (Gelman et al., 2013, page 580):

$$MVt_\nu(\boldsymbol{y};\mu,\boldsymbol{\Sigma}) = \frac{\Gamma(\,(\nu+n)/2\,)}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\left(1+\frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right)^{-(\nu+n)/2}. \tag{5.40}$$

In the rest of this document, we will often omit the variable + semicolon part of the notation. An important likelihood is given by (Christensen, 2004):

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{\theta}) &= MVN(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{C}+\tau^2\boldsymbol{I}) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}exp\left(-\frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})\right). \end{aligned} \tag{5.41}$$

To solve an integral, we will apply the Gamma function (Nahin, 2020, note that for a given n, $\Gamma(n)$ is a constant):

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} \,\mathrm{d}x, \ \mathrm{n} > 0. \tag{5.42}$$

We assume that $\eta^2 = \frac{\tau^2}{\sigma^2}$ is fixed.

5.A.2 Conjugate prior for $\boldsymbol{\beta}, \beta$ and σ^2

In Bayesian statistics, a conjugate prior is a prior which, in combination with the likelihood, provides a closed-form[1] distribution being the same distribution as the prior. For mathematical convenience, we apply a conjugate prior for $\boldsymbol{\beta}$ and σ^2, while we assume for now that ϕ (and thus $\boldsymbol{C}$) is fixed.

Using the product rule of conditional probability, we define this joint prior as the product of two distributions:

$$f(\boldsymbol{\beta}, \sigma^2) = f(\boldsymbol{\beta}|\sigma^2) f(\sigma^2). \tag{5.43}$$

We apply the multivariate normal (MVN) and inverse scaled Chi-squared (χ^2_{ScI}) distributions respectively:

$$\begin{aligned} f(\boldsymbol{\beta}|\sigma^2) &\sim MVN(\boldsymbol{\xi}_0, \sigma^2 \boldsymbol{D}_0) \\ f(\sigma^2) &\sim \chi^2_{ScI}(\nu_0, \varsigma_0^2), \end{aligned} \tag{5.44}$$

where $\boldsymbol{\xi}_0$ is our prior belief about the k regression coefficient means, and $\boldsymbol{D}_0$ (a valid $k \times k$ matrix) our prior belief about the scaled variance of, and correlation between, the elements of $\boldsymbol{\xi}_0$. Expressed in qualitative terms, ς_0^2 (> 0; called the 'scale factor') can be considered our prior belief about the value of σ^2, and ν_0 (> 0; 'degrees of freedom') our prior belief how much we trust ς_0^2, where a higher value for ν_0 indicates more trust.

5.A.3 Conditional distributions for $\boldsymbol{\beta}$ and σ^2

In the main article, we show the calculation of the marginal posteriors $f(\boldsymbol{\beta}|\boldsymbol{z})$ and $f(\sigma^2|\boldsymbol{z})$ for which the conditional distribution $f(\boldsymbol{\beta}, \sigma^2|\phi, \boldsymbol{y})$ is an intermediate result. With "conditional" we mean here: conditional on ϕ and $\boldsymbol{y}$, not on the observations $\boldsymbol{z}$; we use the expression "posterior" for distributions conditional on the observations. The prior presented in the previous section, combined with the likelihood $f(\boldsymbol{y}|\boldsymbol{\theta})$, delivers the conditional distribution $f(\boldsymbol{\beta}, \sigma^2|\phi, y)$ which is, apart from a proportionality factor, a joint $MVN\chi^2_{ScI}$ distribution. The final result is shown in Eqs. (5.61) and (5.63).

[1] With 'closed-form' we mean the formulation as found on Wikipedia (https://en.wikipedia.org/wiki/Closed-form_expression, accessed August 12, 2020): *In mathematics, a closed-form expression is a mathematical expression expressed using a finite number of standard operations. It may contain constants, variables, certain "well-known" operations (e.g., + − × ÷), and functions (e.g., n-th root, exponent, logarithm, trigonometric functions, and inverse hyperbolic functions), but usually no limit, differentiation, or integration.*

We start with writing the joint prior, given in Eq. (5.44), as mathematical equation:

$$\begin{aligned} f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2) = {} & (2\pi\sigma^2)^{-\frac{k}{2}}|\boldsymbol{D}_0|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)\right) \\ & \cdot \frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}(\varsigma_0)^{\nu_0}(\sigma^2)^{-(\frac{\nu_0}{2}+1)} exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}\right). \end{aligned} \tag{5.45}$$

We rearrange the elements, noting that, expressed in general terms, $e^a e^b = e^{(a+b)}$:

$$\begin{aligned} & f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2) \\ & = (\sigma^2)^{-\frac{k}{2}}|\boldsymbol{D}_0|^{-\frac{1}{2}}(\sigma^2)^{-(\frac{\nu_0}{2}+1)} exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)\right)(2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}(\varsigma_0)^{\nu_0} \\ & = (\sigma^2)^{-(\frac{\nu_0+k}{2}+1)}|\boldsymbol{D}_0|^{-\frac{1}{2}} exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)\right)(2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}(\varsigma_0)^{\nu_0}. \end{aligned} \tag{5.46}$$

We write out the conditional distribution as proportional product of prior Eq. (5.46) and likelihood Eq. (5.41):

$$\begin{aligned} f(\boldsymbol{\beta},\sigma^2|\phi,\boldsymbol{y}) \propto {} & (\sigma^2)^{-(\frac{\nu_0+k}{2}+1)}|\boldsymbol{D}_0|^{-\frac{1}{2}} exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)\right)(2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}(\varsigma_0)^{\nu_0} \\ & \cdot (2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{X\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X\beta}-\boldsymbol{y})\right), \end{aligned} \tag{5.47}$$

and we rearrange Eq. (5.47) to:

$$\begin{aligned} f(\boldsymbol{\beta},\sigma^2|\phi,\boldsymbol{y}) \propto {} & \underline{(\sigma^2)^{-(\frac{\nu_0+k}{2}+1)}(\sigma^2)^{-\frac{n}{2}}} \\ & \cdot exp\underline{\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)-\frac{1}{2}(\boldsymbol{X\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X\beta}-\boldsymbol{y})\right)} \\ & \cdot \underline{|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}(2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_0}{2})^{\frac{\nu_0}{2}}}{\Gamma(\frac{\nu_0}{2})}(\varsigma_0)^{\nu_0}(2\pi)^{-\frac{n}{2}}.} \end{aligned} \tag{5.48}$$

Note that for clarification and reference of how we group, discuss and rearrange elements inside equations, we occasionally add coloured underlining. Those lines have no mathematical meaning.

First, we consider the blue underlined part, which we can shape into the same appearance as the σ^2 power in (5.39) by defining the conditional distribution degrees of freedom ν_c:

$$\nu_c = n + \nu_0 \tag{5.49}$$

and simplifying and substituting:

$$\begin{aligned}(\sigma^2)^{-(\frac{\nu_0+k}{2}+1)}(\sigma^2)^{-\frac{n}{2}} &= (\sigma^2)^{-(\frac{\nu_0+k+n}{2}+1)} \\ &= (\sigma^2)^{-(\frac{\nu_c+k}{2}+1)}\ .\end{aligned} \tag{5.50}$$

Secondly, we consider the green underlined part and simplify it by taking out a factor:

$$\begin{aligned}-\frac{\nu_0\varsigma_0^2}{2\sigma^2} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0) - \frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y}) \\ = -\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + (\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T\boldsymbol{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0) + (\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})^T\boldsymbol{C}^{-1}(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{y})\right)\end{aligned} \tag{5.51}$$

To separate $\boldsymbol{\beta}$, we expand the two quadratic matrix terms in Eq. (5.51):

$$-\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + \boldsymbol{\beta}^T\underline{\boldsymbol{D}_0^{-1}}\boldsymbol{\beta} - \underline{\xi_0^T\boldsymbol{D}_0^{-1}}\boldsymbol{\beta} - \boldsymbol{\beta}^T\underline{\boldsymbol{D}_0^{-1}\xi_0} + \underline{\xi_0^T\boldsymbol{D}_0^{-1}\xi_0} + \boldsymbol{\beta}^T\underline{\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}}\boldsymbol{\beta} - \boldsymbol{\beta}^T\underline{\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y}} - \underline{\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{X}}\boldsymbol{\beta} + \underline{\boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}}\right), \tag{5.52}$$

and group those terms around the different powers and orientations of $\boldsymbol{\beta}$, while making use of the identities $(\boldsymbol{P}\cdot\boldsymbol{Q})^T = \boldsymbol{Q}^T\cdot\boldsymbol{P}^T$ and $(\boldsymbol{P}\cdot\boldsymbol{Q}\cdot\boldsymbol{R})^T = \boldsymbol{R}^T\cdot\boldsymbol{Q}^T\cdot\boldsymbol{P}^T$:

$$\begin{aligned}&-\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + \boldsymbol{\beta}^T\underline{(\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})}\boldsymbol{\beta} - \underline{(\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y})^T}\boldsymbol{\beta} - \boldsymbol{\beta}^T\underline{(\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y})} + \underline{\xi_0^T\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}}\right) \\ &= -\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + \boldsymbol{\beta}^T(\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{\beta} - 2\,(\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y})^T\boldsymbol{\beta} + \xi_0^T\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}\right),\end{aligned} \tag{5.53}$$

while noting that $\boldsymbol{D}_0^{-1}$ and $\boldsymbol{C}^{-1}$ are symmetric matrices.

With the substitution:

$$\begin{aligned}\boldsymbol{D}_c &= (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \\ \xi_c &= \boldsymbol{D}_c(\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y}),\end{aligned} \tag{5.54}$$

where $\boldsymbol{D}_c$ and $\boldsymbol{\xi}_c$ represent the conditional versions of $\boldsymbol{D}_0$ and $\boldsymbol{\xi}_0$, respectively, it follows that Eq. (5.53) is equivalent to:

$$-\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2 + \boldsymbol{\beta}^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta} - 2\,\xi_c^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta} + \xi_0^T\boldsymbol{D}_0^{-1}\xi_0 + \boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}\right); \tag{5.55}$$

note hereby that $\boldsymbol{D}_c$, and thus $\boldsymbol{D}_c^{-1}$, is symmetric.

Because we aim for a result that includes the following matrix quadratic form:

$$(\boldsymbol{\beta}-\xi_c)^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta}-\xi_c) \quad = \boldsymbol{\beta}^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta} - 2\,\xi_c^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta} + \xi_c^T\boldsymbol{D}_c^{-1}\xi_c, \tag{5.56}$$

we rewrite Eq. (5.55) to:

$$-\frac{1}{2\sigma^2}\left(\nu_0\varsigma_0^2+\boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+\boldsymbol{y}\boldsymbol{C}^{-1}\boldsymbol{y}-\boldsymbol{\xi}_c^T\boldsymbol{D}_c^{-1}\boldsymbol{\xi}_c+(\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)\right). \tag{5.57}$$

With the next substitution, which defines the conditional distribution scale factor,

$$\varsigma_c^2=\frac{\nu_0\varsigma_0^2+\boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+\boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y}-\boldsymbol{\xi}_c^T\boldsymbol{D}_c^{-1}\boldsymbol{\xi}_c}{\nu_c}, \tag{5.58}$$

it follows that the exponential, green underlined part in Eq. (5.48), which is equal to Eq. (5.57), can be written as:

$$\begin{aligned}&-\frac{1}{2\sigma^2}\left(\nu_c\varsigma_c^2+(\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)\right)\\ =&-\frac{\nu_c\varsigma_c^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T(\sigma^2\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)\end{aligned} \tag{5.59}$$

Thirdly, as the red underlined part in Eq. (5.48) is constant (remember that since ϕ is considered constant, so is the resulting $\boldsymbol{C}$), it can be left out. To make the following equation fitting, we replace this component by this – also constant – expression:

$$\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}(\varsigma_c)^{\nu_c}(2\pi)^{-\frac{k}{2}}|\boldsymbol{D}_c|^{-\frac{1}{2}}. \tag{5.60}$$

Combining results above for all three components, we obtain the following expression for the conditional distribution for $\boldsymbol{\beta}$ and σ^2:

$$\begin{aligned}f(\boldsymbol{\beta},\sigma^2|\phi,\boldsymbol{y})\propto{}&\underline{(\sigma^2)^{-(\frac{\nu_c+k}{2}+1)}}\,exp\left(\underline{-\frac{\nu_c\varsigma_c^2}{2\sigma^2}-\frac{1}{2}(\boldsymbol{\beta}-\xi_c)^T(\sigma^2\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\xi_c)}\right)\underline{\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}(\varsigma_c)^{\nu_c}(2\pi)^{-\frac{k}{2}}|\boldsymbol{D}_c|^{-\frac{1}{2}}}\\ ={}&\underline{(\sigma^2)^{-\frac{k}{2}}}\underline{(2\pi)^{-\frac{k}{2}}|\boldsymbol{D}_c|^{-\frac{1}{2}}}exp\left(\underline{\frac{1}{2}(\boldsymbol{\beta}-\xi_c)^T(\sigma^2\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\xi_c)}\right)\underline{\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}(\varsigma_c)^{\nu_c}}\underline{(\sigma^2)^{-(\frac{\nu_c}{2}+1)}}\,exp\left(\underline{-\frac{\nu_c\varsigma_c^2}{2\sigma^2}}\right).\end{aligned} \tag{5.61}$$

This shows that the joint conditional distribution is the product of a multivariate normal and an inverse scaled Chi-squared distribution:

$$\begin{aligned}f(\boldsymbol{\beta},\sigma^2|\phi,\boldsymbol{y})&=f(\boldsymbol{\beta}|\sigma^2,\phi,\boldsymbol{y})f(\sigma^2|\phi,\boldsymbol{y})\\ &\propto MVN(\boldsymbol{\xi}_c,\sigma^2\boldsymbol{D}_c)\,\chi^2_{ScI}(\nu_c,\varsigma_c^2).\end{aligned} \tag{5.62}$$

For convenience, we repeat the above used substitutions and thus conditional distribution parameters and hyperparameters, and we add (without proof) the limit situation if $\boldsymbol{D}_0^{-1}=\boldsymbol{0}$ and $\nu_0=0$ as provided by Diggle and Ribeiro (2007):

$$\begin{aligned}
\nu_c &= \begin{cases} n-k & \text{if } \boldsymbol{D}_0^{-1} = \boldsymbol{0} \text{ and } \nu_0 = 0 \\ n+\nu_0 & \text{if } \boldsymbol{D}_0^{-1} \neq \boldsymbol{0} \text{ and } \nu_0 > 0 \end{cases} \\
\boldsymbol{D}_c &= (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1} \\
\boldsymbol{\xi}_c &= \boldsymbol{D}_c(\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y}) \\
\varsigma_c^2 &= \frac{\nu_0\varsigma_0^2 + \boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y} - \boldsymbol{\xi}_c^T\boldsymbol{D}_c^{-1}\boldsymbol{\xi}_c}{\nu_c};
\end{aligned} \tag{5.63}$$

5.A.4 Marginal conditional distributions for σ^2 and $\boldsymbol{\beta}$

The marginal conditional density for σ^2 is already given in Eq. (5.62). To derive the marginal conditional density for $\boldsymbol{\beta}$ conditional on $\boldsymbol{y}$ and ϕ, we need to integrate out σ^2 from the joined conditional distribution as given in Eq. (5.62):

$$\begin{aligned}
f(\boldsymbol{\beta}|\phi,\boldsymbol{y}) &= \int_0^\infty f(\boldsymbol{\beta},\sigma^2|\phi,\boldsymbol{y})\,\mathrm{d}\sigma^2 \\
&= \int_0^\infty MVN(\boldsymbol{\beta};\boldsymbol{\xi}_c,\sigma^2\boldsymbol{D}_c)\chi^2_{ScI}(\sigma^2;\nu_c,\varsigma_c^2)\,\mathrm{d}\sigma^2 \\
&= \int_0^\infty (2\pi)^{-\frac{k}{2}}(\sigma^2)^{-\frac{k}{2}}|\boldsymbol{D}_c|^{-\frac{1}{2}} exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T(\sigma^2\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)\right) \\
&\quad \cdot \frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}(\varsigma_c)^{\nu_c}(\sigma^2)^{-(\frac{\nu_c}{2}+1)} exp\left(-\frac{\nu_c\varsigma_c^2}{2\sigma^2}\right)\mathrm{d}\sigma^2 \\
&= (2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}|\boldsymbol{D}_c|^{-\frac{1}{2}}(\varsigma_c)^{\nu_c}\int_0^\infty (\sigma^2)^{-\frac{k+\nu_c+2}{2}} exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T(\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)+\nu_c\varsigma_c^2\right)\right)\mathrm{d}\sigma^2
\end{aligned} \tag{5.64}$$

Inspired by Gelman et al. (2013, Section 3.2), we substitute:

$$\begin{aligned}
&F = (\boldsymbol{\beta}-\boldsymbol{\xi}_c)^T(\boldsymbol{D}_c)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_c)+\nu_c\varsigma_c^2 \text{ and} \\
&q = \frac{F}{2\sigma^2};\ \text{so } \sigma^2 = \frac{F}{2q} \text{ and thus} \\
&\frac{\mathrm{d}q}{\mathrm{d}\sigma^2} = -\frac{F}{2}(\sigma^2)^{-2};\ \text{so} \\
&\mathrm{d}\sigma^2 = -\left(\frac{F}{2}\right)^{-1}(\sigma^2)^2\,\mathrm{d}q = -\left(\frac{F}{2}\right)^{-1}\left(\frac{F}{2q}\right)^2\mathrm{d}q = -\left(\frac{F}{2}\right)q^{-2}\,\mathrm{d}q
\end{aligned} \tag{5.65}$$

and write:

$$\begin{aligned}
f(\boldsymbol{\beta}|\phi,\boldsymbol{y}) &= (2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}|\boldsymbol{D}_c|^{-\frac{1}{2}}(\varsigma_c)^{\nu_c}\int_\infty^0 (\frac{F}{2q})^{-\frac{k+\nu_c+2}{2}} exp(-q)\cdot -\left(\frac{F}{2}\right)q^{-2}\,\mathrm{d}q \\
&= (2\pi)^{-\frac{k}{2}}\frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})}|\boldsymbol{D}_c|^{-\frac{1}{2}}(\varsigma_c)^{\nu_c}(\frac{F}{2})^{-\frac{k+\nu_c}{2}}\int_0^\infty q^{\frac{k+\nu_c-2}{2}} exp(-q)\,\mathrm{d}q.
\end{aligned} \tag{5.66}$$

Note that 1) with changing the integral over $\mathrm{d}\sigma^2$ to an integral over $\mathrm{d}q$ the limits change, because if $\sigma^2 \rightarrow 0$, $q \rightarrow \infty$ and vice versa; 2) with swapping the integral limits in the second line, the integral is multiplied by -1 and thus we get rid of the minus sign; 3) the final integral in Eq. (5.66) is the Gamma function (Eq. (5.42)), defined for any $k + \nu_c > 0$. Thus

$$\begin{aligned} f(\beta|\phi, \boldsymbol{y}) &= (2\pi)^{-\frac{k}{2}} \frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})} |\boldsymbol{D}_c|^{-\frac{1}{2}} (\varsigma_c)^{\nu_c} F^{-\frac{k+\nu_c+2}{2}}\ \Gamma\left(\frac{k+\nu_c+2}{2}\right) \\ &= (2\pi)^{-\frac{k}{2}} \frac{(\frac{\nu_c}{2})^{\frac{\nu_c}{2}}}{\Gamma(\frac{\nu_c}{2})} \Gamma\left(\frac{k+\nu_c+2}{2}\right) |\boldsymbol{D}_c|^{-\frac{1}{2}} (\varsigma_c)^{\nu_c} \left((\boldsymbol{\beta} - \boldsymbol{\xi}_c)^T \boldsymbol{D}_c^{-1} (\boldsymbol{\beta} - \boldsymbol{\xi}_c) + \nu_c \varsigma_c^2\right)^{-\frac{k+\nu_c+2}{2}} \\ &\propto \left(1 + \frac{(\boldsymbol{\beta} - \boldsymbol{\xi}_c)^T \boldsymbol{D}_c^{-1} (\boldsymbol{\beta} - \boldsymbol{\xi}_c)}{\nu_c \varsigma_c^2}\right)^{-\frac{k+\nu_c+2}{2}} . \end{aligned} \tag{5.67}$$

This latter expression is proportional to the multivariate t distribution (Eq. (5.40)) with degrees of freedom $\nu_c + 2$, location vector $\boldsymbol{\xi}_c$ and scale matrix $\varsigma_c^2 \boldsymbol{D}_c$.

5.A.5 Marginal conditional density for ϕ

To derive the marginal conditional density for ϕ given the signal $\boldsymbol{y}$, we need to integrate out β and σ^2:

$$f(\phi|\boldsymbol{y}) = \int_{\boldsymbol{\beta}, \sigma^2} f(\boldsymbol{\beta}, \sigma^2, \phi|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\beta}\, \mathrm{d}\sigma^2 \tag{5.68}$$

In the following, we will first integrate out β, then σ^2.

5.A.6 Integrating out β

Writing out the conditional density term inside the integral:

$$\begin{aligned} f(\boldsymbol{\beta}, \sigma^2, \phi|\boldsymbol{y}) &\propto f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \phi) f(\boldsymbol{\beta}, \sigma^2, \phi) \\ &= f(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \phi) f(\boldsymbol{\beta}, \sigma^2|\phi) f(\phi) \\ &\propto |2\pi\sigma^2 \boldsymbol{C}|^{-\frac{1}{2}} exp\left(\underline{-\frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T (\sigma^2 \boldsymbol{C})^{-1} (\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})}\right) \\ &\cdot\ |2\pi\sigma^2 \boldsymbol{D}_0|^{-\frac{1}{2}} exp\left(\underline{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\xi}_0)^T (\sigma^2 \boldsymbol{D}_0)^{-1} (\boldsymbol{\beta} - \boldsymbol{\xi}_0)}\right) \\ &\cdot\ (\sigma^2)^{-(\frac{\nu_0}{2}+1)} exp\left(-\frac{\nu_0 \varsigma_0^2}{2\sigma^2}\right) \\ &\cdot\ f(\phi). \end{aligned} \tag{5.69}$$

We first work out the exponential terms containing $\boldsymbol{\beta}$, corresponding to the purple underlined parts, while introducing for convenience the temporary variable T_1 for expressing the term of interest:

$$\begin{aligned}
&-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\\
&=-\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T\boldsymbol{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)+(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right\};\\
&=-\frac{1}{2\sigma^2}T_1;
\end{aligned}\tag{5.70}$$

We use a relationship used by Harville (1974) (preceding his Eq. 2,); see also Searle (1971, with $\boldsymbol{C}^{-1}$ added at both sides of Eq. 104 {page 113}):

$$(\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})=(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}),\tag{5.71}$$

with $\hat{\boldsymbol{\beta}}$ the generalised least squares estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y}.\tag{5.72}$$

Combining (5.71) with (5.70), while for now neglecting the term $-\frac{1}{2\sigma^2}$ gives the following expressions for T_1:

$$\begin{aligned}
T_1&=(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T\boldsymbol{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0)+(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\\
&=\boldsymbol{\beta}^T(\boldsymbol{D}_0^{-1}+\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{\beta}-\boldsymbol{\beta}^T(\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}})-(\boldsymbol{\xi}_0^T\boldsymbol{D}_{\mathbf{0}}^{\mathbf{-1}}+\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})\boldsymbol{\beta}+\boldsymbol{R}_1
\end{aligned}\tag{5.73}$$

where rest term $\boldsymbol{R}_1$is given by:

$$\begin{aligned}
\boldsymbol{R}_1&=\boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})+\hat{\boldsymbol{\beta}}^T(\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})\hat{\boldsymbol{\beta}}\\
&=\boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+(\boldsymbol{y}-\hat{\boldsymbol{y}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\hat{\boldsymbol{y}})+\hat{\boldsymbol{y}}^T\boldsymbol{C}^{-1}\hat{\boldsymbol{y}}
\end{aligned}\tag{5.74}$$

with the regression estimation for the signal $\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$.

We next write Eq. (5.73) in the following quadratic matrix form, while borrowing $\boldsymbol{D}_c^{-1}=(\boldsymbol{D}_0^{-1}+\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})$ from Eq. (5.63) and introducing $\hat{\boldsymbol{\xi}}$. The term $-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_p^{-1}\hat{\boldsymbol{\xi}}$ is added to be able to complete the square:

$$\begin{aligned}
T_1&=\left[(\boldsymbol{\beta}-\hat{\boldsymbol{\xi}})^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\xi}})\right]-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_p^{-1}\hat{\boldsymbol{\xi}}+R_1\\
&=\left[\boldsymbol{\beta}^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta}-\boldsymbol{\beta}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\boldsymbol{\beta}+\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}\right]-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_p^{-1}\hat{\boldsymbol{\xi}}+R_1
\end{aligned}\tag{5.75}$$

from which it follows that

$$\begin{aligned} \boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}} &= (\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}), \text{ and thus} \\ \hat{\boldsymbol{\xi}} &= (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{5.76}$$

Note that $\boldsymbol{\xi}_c$ (as defined in Eq. (5.63)) equals $\hat{\boldsymbol{\xi}}$ if we replace $\boldsymbol{y}$ with $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Note also that $\boldsymbol{R}_1$, $\hat{\boldsymbol{\xi}}$ and $\boldsymbol{D}_c^{-1}$ are implicit functions of ϕ but not of σ^2; $\boldsymbol{R}_1$ and $\hat{\boldsymbol{\xi}}$ depend also on the signal $\boldsymbol{y}$.

Next, we take exponents containing $\boldsymbol{\beta}$ from Eq. (5.69), now including the exp function itself, and combine it with Eq. (5.75). We also keep the $-\frac{1}{2\sigma^2}$ factor. We move the rest $\boldsymbol{R}_1$ away from the quadratic form, and to complete the square, we subtract $\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}$, giving the following equality:

$$\begin{aligned} & exp\left(\underline{-\frac{1}{2}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y})}\right) exp\left(\underline{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\xi}_0)^T(\sigma^2\boldsymbol{D}_0)^{-1}(\boldsymbol{\beta} - \boldsymbol{\xi}_0)}\right) \\ & = exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{R}_1)\right)\ exp\left(-\frac{1}{2\sigma^2}(-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}})\right) \\ & \cdot\ exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\xi}})^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\xi}})\right). \end{aligned} \tag{5.77}$$

Using Eqs. (5.77), and adding the multiplication term $|2\pi\sigma^2\boldsymbol{D}_c|^{-\frac{1}{2}}$ and its reciprocal, we rewrite (5.68) as:

$$\begin{aligned} f(\phi|\boldsymbol{y}) \propto & \int_{\sigma^2} |2\pi\sigma^2\boldsymbol{C}|^{-\frac{1}{2}}|2\pi\sigma^2\boldsymbol{D}_0|^{-\frac{1}{2}}(\sigma^2)^{-(\frac{\nu_0}{2}+1)}exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}\right) \\ & \cdot\ exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{R}_1)\right)exp\left(-\frac{1}{2\sigma^2}(-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}})\right) \\ & \cdot\ |2\pi\sigma^2\boldsymbol{D}_c|^{\frac{1}{2}}\int_{\boldsymbol{\beta}}|2\pi\sigma^2\boldsymbol{D}_c|^{-\frac{1}{2}}exp\left(-\frac{1}{2\pi\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\xi}})^T\boldsymbol{D}_c^{-1}(\boldsymbol{\beta} - \hat{\boldsymbol{\xi}})\right)\mathrm{d}\boldsymbol{\beta} \\ & \cdot\ f(\phi)\,\mathrm{d}\sigma^2 \\ = & \int_{\sigma^2} |2\pi\sigma^2\boldsymbol{C}|^{-\frac{1}{2}}|2\pi\sigma^2\boldsymbol{D}_0|^{-\frac{1}{2}}(\sigma^2)^{-(\frac{\nu_0}{2}+1)}exp\left(-\frac{\nu_0\varsigma_0^2}{2\sigma^2}\right) \\ & \cdot\ exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{R}_1)\right)exp\left(-\frac{1}{2\sigma^2}(-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}})\right)\ |2\pi\sigma^2\boldsymbol{D}_c|^{\frac{1}{2}}f(\phi)\,\mathrm{d}\sigma^2. \end{aligned} \tag{5.78}$$

As the integral over $\boldsymbol{\beta}$ encloses a multivariate normal distribution of $\boldsymbol{\beta}$, which integrates to one, we have integrated out $\boldsymbol{\beta}$, as is shown in the final step in above equation.

5.A.7 Integrating out σ^2

Comparable with the approach of bringing all terms containing $\boldsymbol{\beta}$ into a MVN in order to integrate it out, in this section we will put all terms containing σ^2 into a χ^2_{ScI} distribution, with as goal of this section to write Eq. (5.78) in the following form:

$$f(\phi|\boldsymbol{y}) \propto \boldsymbol{R}_2(\phi,\boldsymbol{y})\left(\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}\right)^{-1}\int_{\sigma^2}\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}(\sigma^2)^{-(\frac{\nu}{2}+1)}exp\left(-\frac{\nu\varsigma^2}{2\sigma^2}\right)\mathrm{d}\sigma^2, \tag{5.79}$$

with R_2 a rest term, to be determined. We begin by using Eq. (5.78), while rearranging some terms

$$\begin{aligned}
f(\phi|\boldsymbol{y}) \propto f(\phi)\int_{\sigma^2}&(2\pi\sigma^2)^{-n/2}|\boldsymbol{C}|^{-\frac{1}{2}}\\
&\cdot\ (2\pi\sigma^2)^{-k/2}|\boldsymbol{D}_0|^{-\frac{1}{2}}\\
&\cdot\ (\sigma^2)^{-(\frac{\nu_0}{2}+1)}\\
&\cdot\ exp\left(-\frac{\nu_0\varsigma_0^2+\boldsymbol{R}_1-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}}{2\sigma^2}\right)\\
&\cdot\ (2\pi\sigma^2)^{k/2}|\boldsymbol{D}_c|^{\frac{1}{2}}\,\mathrm{d}\sigma^2\\
= f(\phi)\int_{\sigma^2}&(2\pi)^{-n/2}(\sigma^2)^{-n/2}|\boldsymbol{C}|^{-\frac{1}{2}}\\
&\cdot\ |\boldsymbol{D}_0|^{-\frac{1}{2}}\\
&\cdot\ (\sigma^2)^{-(\frac{\nu_0}{2}+1)}\\
&\cdot\ exp\left(-\frac{\nu_0\varsigma_0^2+\boldsymbol{R}_1-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}}{2\sigma^2}\right)\\
&\cdot\ |\boldsymbol{D}_c|^{\frac{1}{2}}\,\mathrm{d}\sigma^2\\
= f(\phi)&|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}(2\pi)^{-n/2}\\
&\cdot\ \int_{\sigma^2}(\sigma^2)^{-(\frac{\nu_0+n}{2}+1)}\ \cdot\ exp\left(-\frac{\nu_0\varsigma_0^2+\boldsymbol{R}_1-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}}{2\sigma^2}\right)\ \cdot\ \mathrm{d}\sigma^2
\end{aligned} \tag{5.80}$$

Thus, if we set (referring to Eq. (5.63)):

$$\nu = \nu_0 + n = \nu_c \tag{5.81}$$

and

$$\begin{aligned}
\varsigma^2 &= \frac{\nu_0\varsigma_0^2+\boldsymbol{R}_1-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}}{\nu_c}\\
&= \frac{\nu_0\varsigma_0^2+\boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0+(\boldsymbol{y}-\hat{\boldsymbol{y}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\hat{\boldsymbol{y}})+\hat{\boldsymbol{y}}^T\boldsymbol{C}^{-1}\hat{\boldsymbol{y}}-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}}}{\nu_c}
\end{aligned} \tag{5.82}$$

we get a valid χ^2_{ScI} distribution.

Note the similarities, but also the differences, between ς^2 above and the earlier derived conjugate conditional distribution Eq. (5.63):

$$\varsigma_c^2 = \frac{\nu_0\varsigma_0^2 + \boldsymbol{\xi}_0^T \boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + \boldsymbol{y}^T \boldsymbol{C}^{-1}\boldsymbol{y} - \boldsymbol{\xi}_c^T \boldsymbol{D}_c^{-1}\boldsymbol{\xi}_c}{\nu_c}. \tag{5.83}$$

Substituting Eqs. (5.81) and (5.82) in Eq. (5.79) gives:

$$f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}(2\pi)^{-n/2}\left(\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}\right)^{-1}\int_{\sigma^2}\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}(\sigma^2)^{-(\frac{\nu}{2}+1)}exp\left(-\frac{\nu\varsigma^2}{2\pi\sigma^2}\right)\mathrm{d}\sigma^2. \tag{5.84}$$

Because the integral over σ^2 represents a χ^2_{ScI} distribution which integrates to one, we end up with the following marginal conditional distribution for ϕ:

$$\begin{aligned} f(\phi|\boldsymbol{y}) &\propto f(\phi)|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}(2\pi)^{-n/2}\left(\frac{(\frac{\nu}{2})^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}(\varsigma^2)^{\nu/2}\right)^{-1} \\ &= f(\phi)|\boldsymbol{D}_0|^{-\frac{1}{2}}|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}(2\pi)^{-n/2}(\frac{\nu}{2})^{\frac{-\nu}{2}}\ \Gamma(\nu/2)\ (\varsigma^2)^{-\nu/2} \\ &\propto f(\phi)|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}\ (\varsigma^2)^{-\nu/2}, \end{aligned} \tag{5.85}$$

where the last proportionality follows from removing all elements not depending on $\boldsymbol{y}$ and ϕ.

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}$$

$$\mathcal{L}\left(\boldsymbol{\beta}\right)$$

$$f(\boldsymbol{\theta}|\boldsymbol{z}) = \int f(\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{y}$$

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})}$$

$$f(\boldsymbol{y}^*|\boldsymbol{z}) = \iint f(\boldsymbol{y}^*,\boldsymbol{\theta},\boldsymbol{y}|\boldsymbol{z})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\boldsymbol{y}$$

$\pi_i \quad i = 1 \ldots n \qquad \mathcal{L}\left(\boldsymbol{\beta}\right) = \ p(\boldsymbol{y}|\boldsymbol{\beta}) =$

$$var(\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}^*) = \sigma^2\left(\boldsymbol{1} - \boldsymbol{C}^*(\phi)^T(\boldsymbol{C}(\phi))^{-1}\boldsymbol{C}^*(\phi)\right)$$

$$p(y^*|y,\phi) = MVt_{\nu_0+n}(\mu^*, \varsigma_c^2\boldsymbol{E})$$

$$\text{logit}\left(\pi_i\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{d}_i^T\boldsymbol{\beta} \qquad \widehat{\pi}_i$$

$$f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_c|^{\frac{1}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}\ (\varsigma^2)^{-\nu/2}$$

$$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)\,f(\sigma^2)$$

$$\pi_i = \text{logit}^{-1}\left(\boldsymbol{d}_i^T\boldsymbol{\beta}\right) = \frac{\exp(\boldsymbol{d}_i^T\boldsymbol{\beta})}{1+\exp(\boldsymbol{d}_i^T\boldsymbol{\beta})} \qquad \boldsymbol{p(y|\beta)}$$

$$(2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}exp\left(-\frac{1}{2}(\boldsymbol{X\beta}-\boldsymbol{y})^T(\sigma^2\boldsymbol{C})^{-1}(\boldsymbol{X\beta}-\boldsymbol{y})\right)\ \eta^2 = \frac{\tau^2}{\sigma^2}$$

$\boldsymbol{d}_i$

$$t_\nu(\beta_q;\hat\beta_q,\Sigma_q) = \frac{\Gamma[(\nu+1)/2]}{\Gamma(\nu/2)(\nu\pi)^{1/2}|\Sigma|^{1/2}}\left[1+\frac{1}{\nu}(\beta_q-\hat\beta_q)^T\Sigma_q^{-1}(\beta_q-\hat\beta_q)\right]^{-(\nu+1)/2}$$

$$f(\sigma^2) = \chi^2_{ScI}(\nu_0,\varsigma_0^2) \qquad f(\phi|\boldsymbol{y}) \propto f(\phi)|\boldsymbol{D}_c|$$

$\alpha_0 > 0, \beta_0$

π_i

$\sim MVN(\boldsymbol{X\beta}, \sigma^2\boldsymbol{C}(\phi))$

$\boldsymbol{V} = diag(\sigma_1^2,\sigma_2^2,\ldots,\sigma_n^2)$

$f(\boldsymbol{\beta},\sigma^2|\boldsymbol{y},\phi)$

[illegible]

$\propto \int_\phi$

$f_{ILc}(\bar z|\beta,\phi) = \int_{\sigma^2}$

$$|\boldsymbol{C}|^{\frac{1}{2}}\left[(\bar z - \overline{\boldsymbol{X}}\hat\beta)^T\boldsymbol{C}^{-1}(\bar z - \overline{\boldsymbol{X}}\hat\beta) + 2\beta_0\right]$$

$$\boldsymbol{D}_c = (\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$$

$MVt_\nu(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\Gamma[(\nu+n)/2]}{\Gamma(\nu/2)\nu^{n/2}\pi^{n/2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\left(1+\frac{1}{\nu}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}\right.$

$i = 1,\ldots,n$

$(\boldsymbol{P}\cdot\boldsymbol{Q}\{\boldsymbol{R}\})^T_1 = [\boldsymbol{R}^T$

[illegible]

$$T_1 = \left[\boldsymbol{R}_1 = \boldsymbol{\xi}_0^T\boldsymbol{D}_0^{-1}\boldsymbol{\xi}_0 + (\boldsymbol{y}\right.$$

$$\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{y}$$

$$= -\frac{1}{2\sigma^2}\left\{(\boldsymbol{\beta}-\boldsymbol{\xi}_0)^T\boldsymbol{D}_0^{-1}(\boldsymbol{\beta}-\boldsymbol{\xi}_0) + (\boldsymbol{y}-\boldsymbol{X\beta})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right.$$

$\varepsilon(s)$

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

$$\Sigma_q = \boldsymbol{v}[\hat\beta]_{q,q}$$

$\text{var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}\right.$

$$\alpha = \frac{1}{2}(m-k) + \alpha_0$$

$p(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} =$

$$\boldsymbol{D}_0^{-1} + \boldsymbol{X}^T\boldsymbol{C}^{-1}\boldsymbol{X})^{-1}$$

σ^2

$f_{ILa}(\bar z|\phi,\sigma^2) =$

$$f_p(\beta_q|\bar z) \propto \int f_p(\phi|\bar z)t_\nu(\beta_q;\hat\beta_q,\Sigma_q)\,\mathrm{d}\phi$$

$$\boldsymbol{\lambda}\left(s_0\right) \quad f_p(\beta|\bar z) \propto \int_{\sigma^2,\phi} f_l(\bar z|\beta,\sigma^2,\phi)f_0(\beta,\sigma^2,\phi)\,\mathrm{d}\sigma^2$$

$\boldsymbol{p(y|\beta)}$

$$\gamma(h) = \frac{1}{2}E[(\boldsymbol{Z}(s) - \boldsymbol{Z}(s+h))^2]$$

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) \propto \ p(\boldsymbol{\beta})\ p(\boldsymbol{y}|\boldsymbol{\beta})$$

$MVN(\boldsymbol{y};\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad \hat{\boldsymbol{V}} = diag(\hat\sigma$

$$f_p(\boldsymbol{z}^*|\bar{\boldsymbol{z}}) \quad \hat z(\boldsymbol{s}_0) = \sum_{i=1}^n \lambda_i z(s_i)$$

$\gamma(h)$

$$f(\boldsymbol{\beta},\sigma^2) = f(\boldsymbol{\beta}|\sigma^2)f(\sigma^2)$$

$n+p+1 \qquad n\times n$

$RMSE = \sqrt{}$

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

$s_i \int f(\boldsymbol{z}^*|\boldsymbol{z},\boldsymbol{\beta},\sigma^2,\phi)\times f(\boldsymbol{\beta},\sigma^2$

$\hat z(s_i)$

$(2\pi\sigma^2)^{-\frac{n}{2}}|\boldsymbol{C}|^{-\frac{1}{2}}e$

$$p\left(\boldsymbol{\beta}|\boldsymbol{y}\right) = \frac{p(\boldsymbol{\beta},\boldsymbol{y})}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta})}{p(\boldsymbol{y})}$$

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

$$f(\boldsymbol{z}^*|\bar{\boldsymbol{z}},\phi)$$

$$\nu = m - k$$

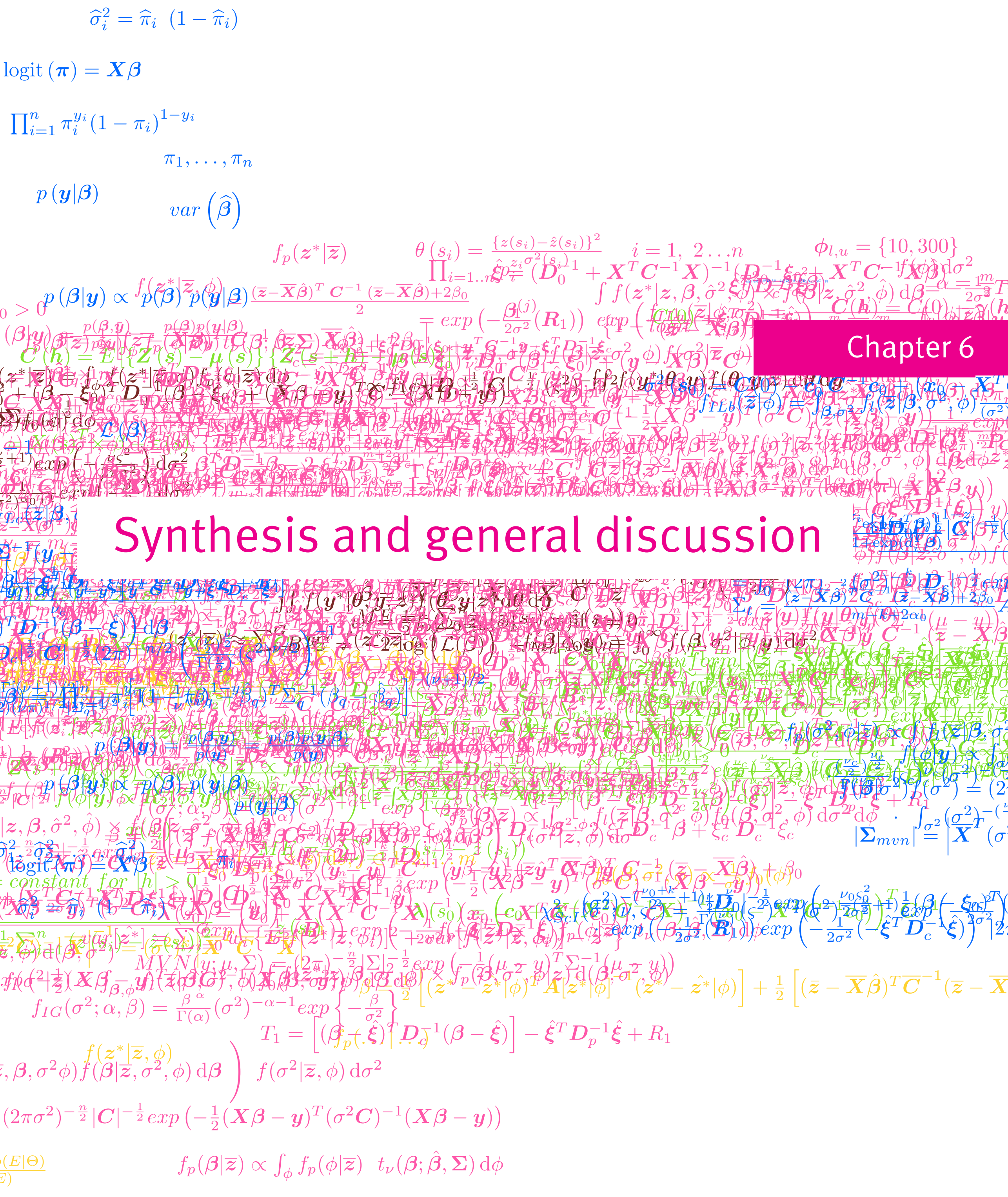
λ_i

$= \int_{\sigma^2}\left(\int_\beta f(\boldsymbol{z}^*|\boldsymbol{z}\right.$

$$\nu_t = m - k + 2\alpha_0$$

$$= exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{R}_1)\right)\ exp\left(-\frac{1}{2\sigma^2}(-\hat{\boldsymbol{\xi}}^T\boldsymbol{D}_c^{-1}\hat{\boldsymbol{\xi}})\right)$$

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



Chapter 6

Synthesis and general discussion

To meet the needs of the future, we need to think about statistically explicit and consistent ways to combine and process spatial data. In this thesis, I applied model-based geostatistics rather than a quite pragmatic approach for mapping crop yield gaps (Chapter 2); re-used soil legacy data the Bayesian way in a clay ripening study(Chapter 3); investigated the added value of taking spatial parameter uncertainty into account for mapping sorghum yields (Chapter 4) and applied hierarchical model-based geostatistics for modelling and mapping the probability of exceeding a threshold of a soil property (Chapter 5). In this final chapter I will review the results of this research and link them to the research questions and knowledge gaps formulated in Chapter 1; the overview is provided in Table 6.1. For the readers' convenience the knowledge gaps (KG) and research questions (RQ) are shortly repeated in Table 6.2. In the final section of this chapter – and also of this thesis – I will round up.

6.1 Change of support

In Chapter 2, I showed the potential of applying geostatistics for mapping the yield gap. Using 37 or 38 observation locations I calculated the potential total crop yields, and associated uncertainties in administrative areas (in this case: countries). Applying model-based geostatistics has clearly added value compared to the prevailing agro-ecological Climate Zones approach, such as a finer scale of modelling, and uncertainty quantification of predictions – both grid-wise and area-wise. Therefore, I concluded that the model-based approach offers more possibilities for both scientists and policy-makers, and is statistically much better founded (which answers *RQ 1*). However, we have no means to validate if our estimation, or our estimated uncertainty, has any connection to the real world because 1) we cannot measure a whole area, only at points and 2) because the variable of interest, modelled potential crop yields, is a quite mathematical construct based on biology, so there is nothing to physically measure at all.

I reversed the change-of-support problem in Chapter 4: I mapped grid-wise the production of a crop based on averages per province. Also here, it is important to acknowledge uncertainty: shown values and patterns are unlikely to be the perfect truth, but are only estimations. As methodology, Bayesian statistics makes it easier to consistently calculate uncertainties. Regarding prediction uncertainty, the method of the maximum

Table 6.1: Section overview.

Section	Addressed knowledge gap	Addressed research question
6.1 Change of support	iii	1, 3
6.2 Incorporating legacy information	ii	2, 3
6.3 Small data issues and algorithm behaviour		3, 4
6.4 Acceptance of model-based and/or Bayesian geostatistics in crop- and soil sciences	i, iv	1
6.5 Data, information, knowledge and wisdom		
6.6 Rounding up and looking ahead		

Table 6.2: Reminder knowledge gaps (KG) and research questions (RQ).

KG i	Lagging experiences and awareness among agronomists regarding spatial aggregation in a model-based way
KG ii	Lacking knowledge and experience with how to combine legacy data and new data
KG iii	Quantifying uncertainty in case of area-to-point prediction
KG iv	Understanding and application of Bayesian methods in pedometrics
RQ 1	What is the added value of model-based geostatistics (including point-to-area) in crop sciences?
RQ 2	Can legacy data be re-used?
RQ 3	More accurate prediction and prediction uncertainty with Bayesian statistics?
RQ 4	More accurate prediction with hierarchical model-based geostatistics?

marginal likelihood[1] was actually in most cases more accurate than the full Bayesian method; both maximum marginal likelihood and full Bayesian mostly outperformed REML, especially for smaller datasets. Regarding the accurate prediction, there was not so much difference between the compared approaches. This answers *RQ 3* and contributes to filling *KG iii*.

The algorithm used in Chapter 4 can with some slight adjustments be used for point-to-area kriging and point-to-point kriging as applied in the case study of Chapter 2, as well as for area-to-area kriging. This allows to compare different approaches – REML, maximal marginal likelihood, and a full Bayesian approach in different change of support settings, and explore the added value. Not treated in this thesis, but a logical extension, is combining data with different support into the spatial model inference (and successive prediction), which is an application of the general concept of spatial data fusion (e.g. (Bogaert and Fasbender, 2007; Castrignanò et al., 2019)). Another research path would be to incorporate non-Gaussian data types (such as the binary soil properties as discussed in Chapters 3 and 5) into change-of-support kriging in a model-based geostatistical way. This might include extra assumptions in case area count data have to be given a meaning on point level (Banerjee et al., 2015, Section 7); an interesting starting point could be the area-to-point Poisson kriging of wildlife counting data as done in Kerry et al. (2013).

6.2 Incorporating legacy information

The challenge of creating a prior from information and knowledge – the distinction between those two will be further discussed in Section 6.5 – comes back in quantifying the 'strength' of the regression parameter prior as done in Chapter 3: in this research, I 'knew' the inferred probability distribution resulting from earlier observations. However, I had to make a choice how to weigh this legacy information in relation to the new observations. In Chapter 3 the weight variable was called the uncertainty multiplication

[1] See Table 4.1, page 75: with maximum marginal likelihood I mean maximum likelihood with trend and variance integrated out, the range parameter estimated as plug-in deterministic value and – as with almost all models in this thesis – the nugget parameter and the Matérn smoothness parameter fixed to arbitrary values.

factor (defined in Section 3.3.3.1, page 48). It will be difficult to find indisputable criteria to set this variable, for two main reasons. First, this weight depends on the perceived change in soil properties over several decades, which is the topic of research itself. Second, the prior has to be weighed against the likelihood; in how far do we have to take the design of data collection (while of course being ignorant about the collected data themselves) into account? According to Gelman et al. (2017), a prior is mostly constructed in relation to the modelled likelihood, although in theory it should be totally independent. My solution was to judge the uncertainty multiplication factor using part of the new observations, de facto applying validation data for model choice, which prohibits estimating the correct prediction error based on statistical validation (Varma and Simon, 2006). For future research, we could consider nested validation approaches (see for example Krstajic et al., 2014; Pejović et al., 2018).

In this thesis, I only experimented with informative priors for the regression coefficients. For geostatistical applications, we should also be able to formulate informative priors for the spatial parameters, following the example in Truong et al. (2014). Applying informative priors can be necessary when using small sample sizes in applied research in the behavioural and social sciences (van de Schoot and Miočević, 2020); perhaps the same holds for spatial statistics. Using informative priors might increase the accuracy of the uncertainty in case of area-to-point prediction with few data (referring to Chapter 4) and perhaps also produce useful results in case of few data, to overcome the initial problems I had with hierarchical modelling in Chapter 5.

In Chapter 5, I reused a legacy map not via a prior but as a covariate, which made it possible to explore its added value to the regression model, and also to validate the legacy map directy against the observations (Table 5.1, page 136). In general however, using covariates might have a disadvantage regarding reproducibility: in the spatial domain, there are many maps with potential useful covariates around. The source, underlying assumptions and generating models of these maps are often difficult to trace. For a spatial statistician, not only smart, conscious and consistent ways to deal with data are important (as argued in the Introduction, Section 1.1), but also the data themselves should as much as possible be from a known source and have a known quality. Just using more data sources is not necessarily always better (Simmonds et al., 2020).

To summarise my answer to *RQ 2*: legacy data can indeed conveniently be used, both as prior and as covariate – but together with some assumptions. Application of legacy data did in the case studies improve the results (expressed in statistical validation metrics). More effort is needed for a wider use of, more experience with, and consciousness about legacy data (further contributing to filling *KG ii*), reaching beyond information described in probability distributions which will be discussed in Section 6.4.

The simulation research in Chapter 4 (where I investigated if model-based approaches could reconstruct the spatial simulation model settings) showed the importance of (often a-priori) decisions, such as covariate model choice, and stressed the importance of making the best possible assumptions, or integrating all possible knowledge, in case of few observations – as might be often the case with area support data (partly answering *RQ 3*). This touches the research of Truong et al. (2014). Note that with real data, if for example the stationarity assumption is violated, some models might be even more sensitive to misspecification and probably underestimate the prediction uncertainty.

On a more detailed level I found that little experience and consciousness is being developed concerning a low-informative prior for the range parameter of the spatial covariance model, although some theory exists (Berger et al., 2001) which I applied in Chapters 4 and 5. For the development of Bayesian model-based geostatistics, more attention for this difficult-to-grasp parameter and its prior is needed.

6.3 Small data issues and algorithm behaviour

In contrast with the big-data methods that nowadays get a lot of attention, one of my focus-points was how to deal with small data. To my knowledge little is known about how geostatistical algorithms actually behave with really small datasets, with the exception of research in Kerry and Oliver (2007), which advises at least 50 observation locations with Gaussian observations and REML analysis confirming earlier research in Lark (2000). Somarathna et al. (2017) concluded that with datasets approaching 100 observations the influence of model choice on prediction results increases compared to bigger datasets.

For example, it would be interesting to add a Bayesian perspective to the research of Chapter 2, possibly supported by simulation trials, because probably some uncertainty was missed with only 38 or 39 observation locations. Although, answering *RQ 3*, the simulations in Chapter 4 suggest that the increased uncertainty correctly inferred by going Bayesian (in comparison with REML parameter estimation combined with universal kriging) is on average only an issue with fewer than 20 observations in case of a spatial structure. I also found that the approach of integrating out and using a gridded calculation in parameter space as applied in Chapter 4 still functions with small datasets, in contrast with the MCMC driven algorithm as applied in Chapter 5, which did not work with small trial datasets (both real-world and simulated), as described in Section 5.5.3.2. It is unclear if this problem is entirely related to using a MCMC algorithm, or rather to the existence of a signal layer in the hierarchical model; it would be interesting to explore small data behaviour for spatial MCMC algorithms with only regression- and spatial parameters (thus: for Gaussian data), for example to investigate if in such cases the correlation between spatial parameters inhibits meaningful Bayesian inference.

Regarding the research described in Chapter 5: the low information density of binary observations appears to be a limiting factor, which might explain why textbook examples for Bayesian Generalised Linear Geostatistical Models are often Poisson or binomial with several draws per location, such as disease occurrences; see for instance the case-studies presented in Diggle and Ribeiro (2007, Section 7.6). Perhaps there are ways, for example related to information- and entropy theory, to analytically investigate the amount of information contained in spatial observations or information possibly contained in future observations according to a given sampling design. Based on this amount, it might be possible to conclude if 1) geostatistical algorithms would work; and 2) geostatistical analyses would make sense in a small data setting. In a next stage, questions 1) and 2) can be asked again combined with an informative prior in a

Bayesian setting. Such an approach would be interesting for both Gaussian and non-Gaussian settings. Not always will it be feasible to collect 1000 binary observations, as was required for the BGLGM-algorithm (in Chapter 5). As already indicated in Chapter 5: possible starting points for such research is offered in Li and Reynolds (1995); Mays et al. (2002); Nowosad and Stepinski (2019).

The issue of small data can also be understood in the context of "limited availability of relevant covariates". The case study in Chapter 5 shows that, compared to a purely covariate driven approach such as Bayesian Generalised Linear Models, the hierarchical model based geostatistics approach performs better in prediction (which answers *RQ 4* in this context), because in such a case geostatistical modelling has a clear advantage.

I want to stress here that using model-based geostatistics does certainly not exclude big data applications, although mathematical and numerical approximations, additional to those discussed in this thesis, might be needed for computational feasibility (see for example Zhang et al., 2018; Sengupta et al., 2016; Blangiardo et al., 2013).

6.4 Acceptance of model-based and/or Bayesian geostatistics in crop- and soil sciences

The rich toolbox of spatial statistics, including geostatistics is easily accessible for crop scientists, see for example the textbook by Schabenberger and Pierce (2002) including several spatial crop case studies and stressing the fact that crop sciences often cannot be separated from soil sciences and related mapping. Also, interesting connections can be made with climatological applications such as spatially modelling the temperature inside a greenhouse (Bojacá et al., 2009). Sometimes, geostatistics are considered a tool supporting Geographic information systems (GIS), which also is of interest for crop scientists (Pierce et al., 2004). However, despite the existence of textbooks and scientific papers, my personal impression (not based on any research) is that crop scientists might have less consciousness about the existence of spatial statistics, while in soil science spatial statistics is more accepted as one of the mainstream methods. To be concrete: the applied agro-ecological Climate Zone approach of spatial prediction (van Wart et al., 2013a) as mentioned in Chapter 2 is in my view a logical continuation of the mechanistic plant growth model (based on calculating plant growth with weather data, and to a lesser extent also depending on soil- and management properties) and is thus process-based, rather than model-based in the statistical sense (van Oijen et al., 2009).

Regarding *KG i*, I focused on a specific procedure and arrived at RQ 1. In Chapter 2, I showed how to calculate potential crop yields per area, including uncertainties, based on an existing dataset produced and used by plant scientists in a model-based way. As already indicated in the previous section, because of the small number of observations a Bayesian extension would perhaps have yielded more accurate uncertainties, but would have made the methodology harder approachable for the target group addressed by the specific research of Chapter 2: agronomists without geostatistical background.

In my opinion, the concept and importance of 'uncertainty' is sometimes difficult to convey to non-statisticians – agronomists and many other disciplines alike – although it is a fundamental part of environmental research (Brown, 2010). In geostatistical textbooks the amount of text about *how* to estimate/infer prediction- or parameter uncertainty exceeds by orders of magnitude the text about *what to do* with this information[2]. Sometimes examples are provided: e.g. Chilès and Delfiner (2012) mentions, in the Introduction, reasons to consider uncertainty in off-shore oil industry and shows a calculation of risk assessment in Chapter 3. For environmental scientists, Webster and Oliver (2007) mentions reasons to consider uncertainty for agriculture, but in the Introduction only. Note that in the research field of sampling design, the maximal allowed prediction uncertainty plays an important role (e.g. Vieira et al., 1983). There is also ongoing research towards visualisation of spatial uncertainty (MacEachren et al., 2005; Pebesma et al., 2007; Kinkeldey et al., 2014; Luzzi, 2016), which shows the importance and recent interest in the subject.

Concluding, and relating to *KG iv*: more effort could be given to translate the concept and importance of uncertainty towards other disciplines, but considering the added value of fine-tuning mathematics by going ‘model-based’ and ‘Bayesian’ for the sake of inferring more accurate uncertainties might be asking too much for many geostatistical users. The other way around, re-using legacy data (including ‘subjective’ informative priors) might be more appealing for users working on real-life cases while inducing resistance in more fundamental research fields. Development of Bayesian geostatistics should certainly continue (also because the philosophical concept is appealing), but in my opinion, for application-directed research the focus should for now be on re-using legacy data and the possibilities offered by hierarchical models.

On the longer term, there could also be more emphasis on the development of easily accessible applications related to model-based geostatistics (Gelfand and Banerjee, 2017, Section 7.1), rather than – for outsiders – somewhat obscure R-packages. Let me provide two examples: 1) Kerry and Oliver (2007) presume that the disappointing acceptance rate of REML algorithms in geostatistical research is due to the lack of user-friendliness; 2) for me it was quite an effort to successfully run the BGLGM-algorithm as discussed in Chapter 5, although it is implemented in a well-developed R-package. Calculation speed also adds to user-friendliness: thus replacing MCMC algorithms (which in case of BGLGM took 25-50 hours calculation time) with approaches such as INLA (Blangiardo et al., 2013) might be worth consideration for a broader acceptance. Also, more emphasis could be given to easily accessible explanations of the used principles which would help practitioners and fellow scientists, having less knowledge about spatial statistical modelling, to be aware of the background, possibilities, benefits and drawbacks of model-based tools. For this reason, I graphically showed in this thesis several of the applied principles and algorithms. With the same transparency intention towards non-statistical users, I also showed many of the equations used by the applied algorithms in this thesis.

A remark: we – statistical modellers – tend to be interested in the model parameter values, and as Bayesian modellers even in probability distributions of those model parameters. One can ask however in how far this is relevant for practitioners or for fel-

[2] Actually this PhD thesis is no exception to this.

low researchers of other disciplines. Often, their only goal in applying geostatistics is spatial prediction. According to Chilès and Delfiner (2012, Introduction), the goal of geostatistical models is descriptive rather than interpretive, although Diggle and Ribeiro (2007, Section 1.3) sets as primary scientific objectives estimation and prediction, and secondary hypothesis testing.

And as a final remark in this section, I would like to stress that there is no sharp line between Bayesian and 'non-Bayesian' (or conventional, or frequentist) statistics, neither should there be one in geostatistics. As Table 4.1 (page 75) shows, just like examples in textbooks such as Diggle and Ribeiro (2007, Chapter 7): many shades of grey exist, we can choose to estimate a point 'plug-in' value of one parameter while inferring the posterior distribution of another parameter in the same model. Also, in Chapter 5 we found that the posterior distribution for the regression parameter for the Bayesian Generalised Linear Model can be inferred using algorithms already developed for the well-known Generalised Linear Model, in case of a low-informative prior. Bayesian statistics should not be considered something alien.

6.5 Data, information, knowledge and wisdom

In Chapters 3, 4 and 5 I applied Bayesian statistics, in Chapter 3 even with informative priors. Central in the Bayesian paradigm is the notion of knowledge expressed as a probability distribution, which extends beyond only mathematics: there is also a human factor involved. Also the attitude towards uncertainty has a human component. Therefore I like to discuss a somewhat broader approach in this section.

Regarding Bayesian statistics, Truong et al. (2014) applied a formal definition of *statistical expert elicitation* to formulate informative priors in a geostatistical context. They asked three experts with a background in geostatistics and knowledge about the spatial topic of investigation to elicit geostatistical parameter distributions, resulting in very different answers; in the type of distribution itself (Gaussian, scaled-Beta, etc.) and in the distribution parameters. The relation between the real world, our perceived knowledge of it, and the expression of that knowledge in statistical probability distributions might be counter-intuitive (Spiegelhalter, 2019) and touches research fields such as psychology and science philosophy. Experiments show that children (from ca. 10 years of age on) and layman adults can follow Bayesian reasoning if supported by visualisation of frequencies (Zhu and Gigerenzer, 2006; Eichler et al., 2020); however the majority of these tests seem to be limited to binary outcomes – for example, the probability of a person being ill, given a test result and additional information. As far as I am aware, how to deal with continuous distributions is less explored.

I would like to point at the hierarchy as shown in Figure 6.1, which is used in many information theory textbooks. In my opinion, Bayesian models (and statisticians in general) tend to focus on the data and information part. On the way up in the hierarchy, in some instances it proves difficult to make concepts such as "prediction uncertainty" part of the knowledge of users and fellow non-statistical scientists, as already discussed in Section 6.4. This becomes less difficult if translated to concepts such as "ambiguity" and "equivocality" (MacEachren, 2015), or "risk assessment" and "decision taking support"

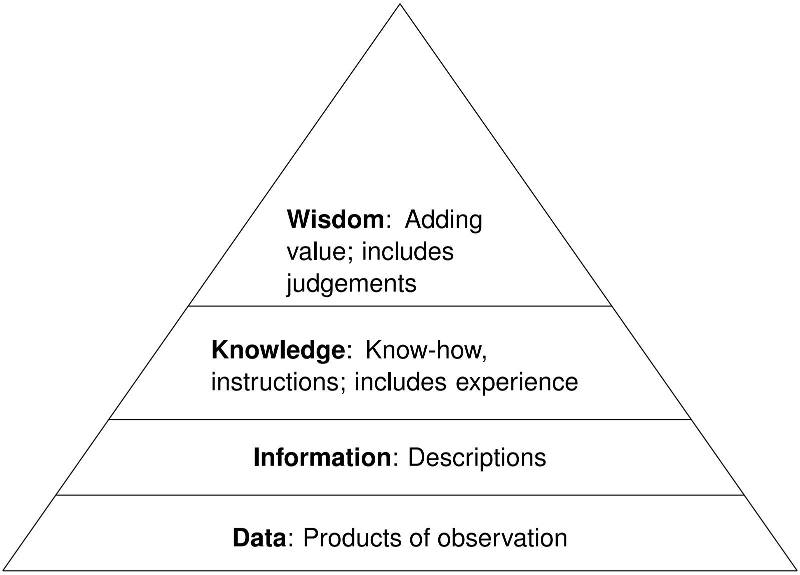

Wisdom: Adding value; includes judgements
Knowledge: Know-how, instructions; includes experience
Information: Descriptions
Data: Products of observation

Figure 6.1: The Data-Information-Knowledge-Wisdom (DIKW) hierarchy, as found in many textbooks (Rowley, 2007). An informative prior is formulated at the *information* (descriptive) layer, while we would like to incorporate *knowledge* or even *wisdom* in it.

(Kinkeldey et al., 2017; Roth, 2009). In this research, I emphasised the importance of uncertainty (most notably in Chapters 2 and 4), but I am not used to express my findings in terms of knowledge or wisdom – and perhaps this is not possible for me because each reader has a specific context. But somebody will have to make this translation.

In a Bayesian context, and depending on the research and modelling goal, we also might like to connect in the hierarchy and include less tangible, but still valuable concepts such as 'knowledge' and perhaps even 'wisdom' into 'information' expressed as an informative prior probability distribution. This happened in Chapter 3 (and was also discussed in Section 6.2): I already calculated on an information/descriptive level a prior distribution, but I needed to scale its variance, which in an ideal case should have been based on knowledge – but I actually used again data and information. Although the idea of Bayesian statistics is to express knowledge in probability distributions, in practice it seems difficult to escape the data- and information layers. Beyond the scope of this work I recommend more – multidisciplinary – research on how to cross the border from data and information to knowledge and wisdom, and vice versa. For example, to re-use data, information, knowledge and wisdom in an efficient, effective and of course also scientifically correct way, we might build on existing work such as O'Hagan and Forster (2004, Chapter 6), Kuhnert et al. (2010) and O'Hagan (2012).

6.6 Rounding up and looking ahead

Although I focused on crop- and soil applications, geostatistics has many other applications : archaeology (Lloyd and Atkinson, 2004), environmental exposure risk and disease mapping (Goovaerts, 2014), wildlife mapping (Kerry et al., 2013), mining and geology (its origin) just to name a few. Geostatisticians connect to all those disciplines using a coherent and mathematically correct framework.

In this concluding chapter, I argued that several future research paths are possible: among others research into small-data behaviour of geostatistical algorithms; research to formulate and incorporate informative priors based on knowledge and/or legacy data; research to combine informative priors with small data; the application of Bayesian model-based geostatistics for area-to-area kriging; and change-of support in non-Gaussian settings. I also argued that, to connect to practitioners and fellow scientists with little geostatistical background, research and intentions should be directed to (keep) explaining the importance and application of prediction uncertainty; to making applications user-friendly, and to making mathematical explanations transparent and approachable.

With some of the explored approaches, for example Bayesian area-to-point kriging (Chapter 3) and BGLGM (Chapter 5), the extra effort may in many cases not outweigh the benefits for most practical uses, because of little prediction accuracy gain; likewise, relating to Chapter 4, the extra effort of Bayesian modelling might not outweigh the gain in prediction uncertainty accuracy. But I also showed benefits of (Bayesian) model-based geostatistics in small data cases. In my opinion it is important to keep developing and exploring such methods and algorithms, as they are model-based and hence mathematically correct. Building on a solid mathematical foundation, mathematically correct extensions can be made, hopefully able to tackle more complicated problems; especially problems that cannot be solved with big data approaches.

Bibliography

Albert, J. (2009). *Bayesian computation with R*. New York: Springer Science & Business Media, doi:10.1007/978-0-387-92298-0.

Alexandrov, V., Eitzinger, J., Cajic, V. and Oberforster, M. (2002). Potential impact of climate change on selected agricultural crops in north-eastern Austria. *Global Change Biology* 8: 372–389, doi:10.1046/j.1354-1013.2002.00484.x.

Allard, D., Senoussi, R. and Porcu, E. (2016). Anisotropy models for spatial data. *Mathematical Geosciences* 48: 305–328, doi:10.1007/s11004-015-9594-x.

Arab, A., Hooten, M. B. and Wikle, C. K. (2017). *Hierarchical Spatial Models*. Cham: Springer International Publishing. 837–846, doi:10.1007/978-3-319-17885-1_564.

Assel, M., Sjoberg, D. D. and Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research* 1: 19, doi:10.1186/s41512-017-0020-3.

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., Boote, K. J., Thorburn, P. J., Rotter, R. P., Cammarano, D., Brisson, N., Basso, B., Martre, P., Aggarwal, P. K., Angulo, C., Bertuzzi, P., Biernath, C., Challinor, A. J., Doltra, J., Gayler, S., Goldberg, R., Grant, R., Heng, L., Hooker, J., Hunt, L. A., Ingwersen, J., Izaurralde, R. C., Kersebaum, K. C., Muller, C., Naresh Kumar, S., Nendel, C., O'Leary, G., Olesen, J. E., Osborne, T. M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M. A., Shcherbak, I., Steduto, P., Stockle, C., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., Wallach, D., White, J. W., Williams, J. R. and Wolf, J. (2013). Uncertainty in simulating wheat yields under climate change. *Nature Climate Change* 3: 827–832, doi:10.1038/nclimate1916.

Aström, K. J. and Murray, R. M. (2008). *Feedback systems: an introduction for scientists and engineers*. Princeton university press.

Bae, S. E., Shinn, T.-W. and Takaoka, T. (2014). A faster parallel algorithm for matrix multiplication on a mesh array. *Procedia Computer Science* 29: 2230 – 2240, doi:10.1016/j.procs.2014.05.208, 2014 International Conference on Computational Science.

Bakker, H. d., Schelling, J., Brus, D. and Wallenburg, C. v. (1989). *Systeem van de bodemclassificatie voor Nederland: De hogere niveaus [Soil classification for the Netherlands: higher levels]*. Wageningen: Pudoc.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Crc Press, doi:10.1201/9780203487808.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2015). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Beaudette, D. (2020). Accuracy and uncertainty for categorical predictions.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* 37: 905–938, doi:10.1214/07-AOS587.

Berger, J. O., Oliveira, V. de and Sans’o, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* 96: 1361–1374.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)* 41: 113–128, doi:10.1111/j.2517-6161.1979.tb01066.x.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and Spatio-temporal Epidemiology* 4: 33–49, doi:10.1016/j.sste.2012.12.001.

Bogaert, P. and Fasbender, D. (2007). Bayesian data fusion in a spatial prediction context: a general formulation. *Stochastic Environmental Research and Risk Assessment* 21: 695–709, doi:10.1007/s00477-006-0080-3.

Bojacá, C. R., Gil, R. and Cooman, A. (2009). Use of geostatistical and crop growth modelling to assess the variability of greenhouse tomato yield caused by spatial temperature variations. *Computers and Electronics in Agriculture* 65: 219 – 227, doi:10.1016/j.compag.2008.10.001.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (2011). *Handbook of Markov chain Monte Carlo*. Handbooks of modern statistical methods. Boca Raton, London, New York: CRC press.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 7: 434–455, doi:10.1080/10618600.1998.10474787.

Brouwer, F., Vries, F. de and Walvoort, D. (2018). Basisregistratie Ondergrond (BRO) actualisatie bodemkaart: Herkartering van de bodem in Flevoland [Key Register of the Subsurface (BRO); Update of soil map: Soil remapping in Flevoland. Tech. Rep. WOt-technical report 143, WUR, Wageningen, doi:10.18174/468672.

Brown, J. D. (2010). Prospects for the open treatment of uncertainty in environmental research. *Progress in Physical Geography: Earth and Environment* 34: 75–100, doi:10.1177/0309133309357000.

Brown, P. E. et al. (2015). Model-based geostatistics the easy way. *Journal of Statistical Software* 63: 1–24, doi:10.18637/jss.v063.i12.

Brus, D. J., Boogaard, H., Ceccarelli, T., Orton, T. G., Traore, S. and Zhang, M. (2018). Geostatistical disaggregation of polygon maps of average crop yields by area-to-point kriging. *European Journal of Agronomy* 97: 48–59, doi:10.1016/j.eja.2018.05.003.

Brus, D. J., Gruijter, J. J. de and Groenigen, J. W. van (2006). *Designing Spatial Coverage Samples Using the k-means Clustering Algorithm*. Elsevier, *Volume 31*, book section Chapter 14. 183–192, doi:10.1016/S0166-2481(06)31014-8.

Brus, D. J. and Heuvelink, G. B. M. (2007). Optimization of sample patterns for universal kriging of environmental variables. *Geoderma* 138: 86–95, doi:10.1016/j.geoderma.2006.10.016.

Brus, D. J., Kempen, B. and Heuvelink, G. B. M. (2011). Sampling for validation of digital soil maps. *European Journal of Soil Science* 62: 394–407, doi:10.1111/j.1365-2389.2011.01364.x.

Brus, D. J., Orton, T. G., Walvoort, D. J. J., Reijneveld, J. A. and 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, doi:10.1016/j.geoderma.2014.02.011.

Bussel, L. van, Müller, C., Van Keulen, H., Ewert, F. and Leffelaar, P. (2011). The effect of temporal aggregation of weather input data on crop growth models' results. *Agricultural and forest meteorology* 151: 607–619, doi:10.1016/j.agrformet.2011.01.007.

Bussel, L. G. J. van, Ewert, F., Zhao, G., Hoffmann, H., Enders, A., Wallach, D., Asseng, S., Baigorria, G. A., Basso, B., Biernath, C., Cammarano, D., Chryssanthacopoulos, J., Constantin, J., Elliott, J., Glotter, M., Heinlein, F., Kersebaum, K.-C., Klein, C., Nendel, C., Priesack, E., Raynal, H., Romero, C. C., Rötter, R. P., Specka, X. and Tao, F. (2016). Spatial sampling of weather data for regional crop yield simulations. *Agricultural and Forest Meteorology* 220: 101–115, doi:10.1016/j.agrformet.2016.01.014.

Bussel, L. G. J. van, Grassini, P., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., Groot, H. de, Saito, K., Cassman, K. G. and Ittersum, M. K. van (2015). From field to atlas: Upscaling of location-specific yield gap estimates. *Field Crops Research* 177: 98–108, doi:10.1016/j.fcr.2015.03.005.

Cassman, K. G. (1999). Ecological intensification of cereal production systems: Yield potential, soil quality, and precision agriculture. *Proceedings of the National Academy of Sciences* 96: 5952–5959, doi:10.1073/pnas.96.11.5952.

Castrignanò, A., Quarto, R., Venezia, A. and Buttafuoco, G. (2019). A comparison between mixed support kriging and block cokriging for modelling and combining spatial data with different support. *Precision Agriculture* 20: 193–213, doi:10.1007/s11119-018-09630-w.

Chen, M.-H. and Ibrahim, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica* 13: 461–476.

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician* 49: 327–335, doi:10.2307/2684568.

Chilès, J. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., second edition ed., doi:10.1002/9781118136188.

Chivers, I. and Sleightholme, J. (2015). *An Introduction to Algorithms and the Big O Notation*. Springer. doi:10.1007/978-3-319-17701-4_23.

Christensen, O. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics* 13: 702–718, doi:10.1198/106186004x2525.

Christensen, O. and Ribeiro Jr, P. (2015). Package ”geoRglm”: A package for generalised linear spatial models.

Christensen, O. F. (2002). Methodology and applications in non-linear model-based geostatistics. Ph.d. thesis.

Christensen, O. F. and Ribeiro Jr, P. (2002). geoRglm - a package for generalised linear spatial models. *R News* 2: 26–28, iSSN 1609-3631.

Christensen, O. F., Roberts, G. O. and Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics* 15: 1–17, doi:10.1198/106186006x100470.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2010). *Bayesian ideas and data analysis : an introduction for scientists and statisticians*. Chapman & Hall/CRC texts in statistical science. Boca Raton, Fla [etc.]: CRC [etc.].

Collard, F., Kempen, B., Heuvelink, G. B. M., Saby, N. P. A., Forges, A. C. Richer de, Lehmann, S., Nehlig, P. and Arrouays, D. (2014). Refining a reconnaissance soil map by calibrating regression models with data from the same map (normandy, france). *Geoderma Regional* 1: 21–30, doi:10.1016/j.geodrs.2014.07.001.

Collet, D. (1991). *Modelling Binary Data*. London: Chapman & Hall.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association* 91: 883–904, doi:10.2307/2291683.

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer, doi:10.1007/978-0-387-48536-2.

Diggle, P. J. and Ribeiro, P. J. (2002). Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling* 6: 129–146, doi:10.1080/1361593022000029467.

Diggle, P. J., Ribeiro, P. J. and Christensen, O. F. (2003). *An Introduction to Model-Based Geostatistics*. New York, NY: Springer New York. 43–86, doi:10.1007/978-0-387-21811-3_2.

Dinoloket (2020). Digitaal geologisch model [digital geological model].

Dixit, A. and Roy, V. (2017). MCMC diagnostics for higher dimensions using Kullback Leibler divergence. *Journal of Statistical Computation and Simulation* 87: 2622–2638, doi:10.1080/00949655.2017.1335313.

Eichler, A., Böcherer-Linder, K. and Vogel, M. (2020). Different visualizations cause different strategies when dealing with bayesian situations. *Frontiers in Psychology* 11: 1897, doi:10.3389/fpsyg.2020.01897.

Emery, X. and Ortiz, J. M. (2004). Shortcomings of multiple indicator kriging for assessing local distributions. *Applied Earth Science* 113: 249–259, doi:10.1179/174327504X27242.

Evangelou, E. and Roy, V. (2019). geoBayes: Analysis of geostatistical data using Bayes and empirical Bayes methods. R package version 0.6.3.

Evans, L. T. (1996). *Crop evolution, adaptation and yield*. Cambridge University Press.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters* 27: 861–874, doi:10.1016/j.patrec.2005.10.010.

Finley, A. O., Banerjee, S. and Gelfand, A. E. (2015). spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software* 63.

Fouedjio, F. and Klump, J. (2019). Exploring prediction uncertainty of spatial data in geostatistical and machine learning approaches. *Environmental earth sciences* 78: 38.

Frigessi, A. and Stander, J. (1994). Informative priors for the Bayesian classification of satellite images. *Journal of the American Statistical Association* 89: 703–709, doi:10.1080/01621459.1994.10476797.

Garthwaite, P. H., Fan, Y. and Sisson, S. A. (2016). Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics - Theory and Methods* 45: 5098–5111, doi:10.1080/03610926.2014.936562.

Gelfand, A. E. and Banerjee, S. (2017). Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and Its Application* 4: 245–266, doi:10.1146/annurev-statistics-060116-054155.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2013). *Bayesian data analysis*. Texts in statistical science. Boca Raton, FL [etc.]: Chapman and Hall/CRC, 3rd ed.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.* : 1360–1383doi:10.1214/08-aoas191.

Gelman, A., Shirley, K. et al. (2011). *Inference from simulations and monitoring convergence*. CRC Press Boca Raton, FL, *6*. 163–174.

Gelman, A., Simpson, D. and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* 19, doi:10.3390/e19100555.

Giorgi, E., Diggle, P. J. et al. (2017). PrevMap: an R package for prevalence mapping. *J Stat Softw* 78: 1–29, doi:10.18637/jss.v078.i08.

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. *International Journal of Health Geographics* 5: 52, doi:10.1186/1476-072x-5-52.

Goovaerts, P. (2008). Kriging and semivariogram deconvolution in the presence of irregular geographical units. *Mathematical Geosciences* 40: 101–128, doi:10.1007/s11004-007-9129-1.

Goovaerts, P. (2014). Geostatistics: a common link between medical geography, mathematical geology, and medical geology. *Journal of the Southern African Institute of Mining and Metallurgy* 114: 605 – 613.

Grassini, P., Van Bussel, L., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., De Groot, H., Van Ittersum, M. and Cassman, K. (2015). How good is good enough? data requirements for reliable yield-gap analysis. *Field Crops Research* 177: 49–63, doi:10.1016/j.fcr.2015.03.004.

Griffin, J. E. and Walker, S. G. (2013). On adaptive Metropolis-Hastings methods. *Statistics and Computing* 23: 123–134, doi:10.1007/s11222-011-9296-2.

Gruijter, J. de, Brus, D. J., Bierkens, M. F. and Knotters, M. (2006). *Sampling for natural resource monitoring*. Springer Science & Business Media, doi:10.1007/3-540-33161-1.

Grunwald, S., Thompson, J. A. and Boettinger, J. L. (2011). Digital soil mapping and modeling at continental scales: Finding solutions for global issues. *Soil Science Society of America Journal* 75: 1201–1213, doi:10.2136/sssaj2011.0025.

Gunnink, J., Maljers, D., Gessel, S. van, Menkovic, A. and Hummelman, H. (2013). Digital geological model (DGM): a 3D raster model of the subsurface of the Netherlands. *Netherlands Journal of Geosciences - Geologie en Mijnbouw* 92: 33–46, doi:10.1017/S0016774600000263.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. and Dougherty, E. R. (2010). Small-sample precision of roc-related estimates. *Bioinformatics* 26: 822–830, doi:10.1093/bioinformatics/btq037.

Hartigan, J. A. and Wong, M. A. C. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society* 28: 100–108, doi:10.2307/2346830.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* 61: 383–385, doi:10.2307/2334370.

Hendriks, C. M. J., Stoorvogel, J. J. and Claessens, L. (2016). Exploring the challenges with soil data in regional land use analysis. *Agricultural Systems* 144: 9–21, doi:10.1016/j.agsy.2016.01.007.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Jesus, J. Mendes de, Tamene, L. and Tondoh, J. E. (2015). Mapping soil properties of africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* 10: e0125814, doi:10.1371/journal.pone.0125814.

Heung, B., Ho, H. C., Zhang, J., Knudby, A., Bulmer, C. E. and Schmidt, M. G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* 265: 62–77, doi:10.1016/j.geoderma.2015.11.014.

Heuvelink, G. B. M. and Pebesma, E. J. (1999). Spatial aggregation and soil process modelling. *Geoderma* 89: 47–65, doi:10.1016/S0016-7061(98)00077-9.

Hodge, S. and Vieland, V. (2017). Information loss in binomial data due to data compression. *Entropy* 19: 75, doi:10.3390/e19020075.

Hoogland, T., Knotters, M., Pleijter, M. and Walvoort, D. (2014). Actualisatie van de grondwatertrappenkaart van holoceen Nederland: resultaten van het veldonderzoek [Updating ground water table map of the holocene part of the Netherlands: results of field research]. Report, Alterra Wageningen UR.

Horta, A., Pereira, M. J. a., Gonçalves, M., Ramos, T. and Soares, A. (2014). Spatial modelling of soil hydraulic properties integrating different supports. *Journal of Hydrology* 511: 1–9, doi:doi.org/10.1016/j.jhydrol.2014.01.027.

Ittersum, M. van and Rabbinge, R. (1997). Concepts in production ecology for analysis and quantification of agricultural input-output combinations. *Field Crops Research* 52: 197–208, doi:10.1016/S0378-4290(97)00037-3.

Ittersum, M. K. van, Cassman, K. G., Grassini, P., Wolf, J., Tittonell, P. and Hochman, Z. (2013). Yield gap analysis with local to global relevance—a review. *Field Crops Research* 143: 4–17, doi:10.1016/j.fcr.2012.09.009.

Jansen, M. J. (1998). Prediction error through modelling concepts and uncertainty from basic data. *Nutrient Cycling in Agroecosystems* 50: 247–253, doi:10.1023/a:1009748529970.

Jewson, S. (2004). The problem with the Brier score.

Jing, L. and De Oliveira, V. (2015). geoCount: An R package for the analysis of geostatistical count data. *2015* 63: 33, doi:10.18637/jss.v063.i11.

Jongmans, A. G., Berg, M. W. v. d., Sonneveld, M. P. W., Peek, G. J. W. C. and Saparoea, R. M. v. d. Berg van (2013). *Landschappen van Nederland: geologie, bodem en landgebruik [Landscapes of The Netherlands: geology, soil and land use]*. Wageningen: Wageningen Academic Publishers, doi:10.3920/978-90-8686-213-9.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology* 15: 445 – 468, doi:10.1007/BF01031292.

Journel, A. G. and Huijbregts, C. J. (1978). *Mining geostatistics*, *600*. Academic press London.

Kaufmann, R. von, Okigbo, B. and Oppong, E. (1983). *The environmental setting*, ANIMAL PRODUCTION AND HEALTH PAPER. Rome, Italy: FAO, book section 1.

Keesstra, S. D., Bouma, J., Wallinga, J., Tittonell, P., Smith, P., Cerdà, A., Montanarella, L., Quinton, J. N., Pachepsky, Y., Putten, W. H. van der, Bardgett, R. D., Moolenaar, S., Mol, G., Jansen, B. and Fresco, L. O. (2016). The significance of soils and soil science towards realization of the united nations sustainable development goals. *SOIL* 2: 111–128, doi:10.5194/soil-2-111-2016.

Kempen, B., Brus, D. J. and Heuvelink, G. B. M. (2012). Soil type mapping using the generalised linear geostatistical model: A case study in a Dutch cultivated peatland. *Geoderma* 189–190: 540–553, doi:10.1016/j.geoderma.2012.05.028.

Kempen, B., Brus, D. J., Heuvelink, G. B. M. and Stoorvogel, J. J. (2009). Updating the 1:50,000 Dutch soil map using legacy soil data: A multinomial logistic regression approach. *Geoderma* 151: 311–326, doi:10.1016/j.geoderma.2009.04.023.

Kerry, R., Goovaerts, P., Rawlins, B. G. and Marchant, B. P. (2012). Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. *Geoderma* 170: 347–358, doi:10.1016/j.geoderma.2011.10.007.

Kerry, R., Goovaerts, P., Smit, I. and Ingram, B. (2013). A comparison of multiple indicator kriging and area-to-point poisson kriging for mapping patterns of herbivore species abundance in kruger national park, south africa. *International Journal of Geographical Information Science* 27: 47–67, doi:10.1080/13658816.2012.663917.

Kerry, R. and Oliver, M. A. (2007). Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma* 140: 383–396, doi:10.1016/j.geoderma.2007.04.019.

Kinkeldey, C., MacEachren, A. M., Riveiro, M. and Schiewe, J. (2017). Evaluating the effect of visually represented geodata uncertainty on decision-making: systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science* 44: 1–21, doi:10.1080/15230406.2015.1089792.

Kinkeldey, C., MacEachren, A. M. and Schiewe, J. (2014). How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal* 51: 372–386, doi:10.1179/1743277414Y.0000000099.

Kitanidis, P. K. (1983). Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Resources Research* 19: 909–921, doi:10.1029/WR019i004p00909.

Kitanidis, P. K. (1986). Parameter uncertainty in estimation of spatial functions: Bayesian analysis. *Water Resources Research* 22: 499–507, doi:10.1029/WR022i004p00499.

Koster, E. A. (2009). The "european aeolian sand belt": Geoconservation of drift sand landscapes. *Geoheritage* 1: 93–110, doi:10.1007/s12371-009-0007-8.

Krstajic, D., Buturovic, L. J., Leahy, D. E. and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* 6: 10, doi:10.1186/1758-2946-6-10.

Kuhnert, P. M., Martin, T. G. and Griffiths, S. P. (2010). A guide to eliciting and using expert knowledge in bayesian ecological models. *Ecology Letters* 13: 900–914, doi:10.1111/j.1461-0248.2010.01477.x.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models*. Irwin Chicago, 5th ed.

Kyriakidis, P. C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis* 36: 259–289, doi:10.1111/j.1538-4632.2004.tb01135.x.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38: 963–74.

Lark, R. and Ferguson, R. (2004). Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging. *Geoderma* 118: 39 – 53, doi:10.1016/S0016-7061(03)00168-X.

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science* 51: 717–728, doi:10.1046/j.1365-2389.2000.00345.x.

Lark, R. M. and Cullis, B. R. (2004). Model-based analysis using reml for inference from systematically sampled data on soil. *European Journal of Soil Science* 55: 799–813, doi:10.1111/j.1365-2389.2004.00637.x.

Lark, R. M., Cullis, B. R. and Welham, S. J. (2006). On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (e-blup) with reml. *European Journal of Soil Science* 57: 787–799, doi:10.1111/j.1365-2389.2005.00768.x.

Le, N. D. and Zidek, J. V. (1992). Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. *Journal of Multivariate Analysis* 43: 351–374, doi:10.1016/0047-259X(92)90040-M.

Li, H. and Reynolds, J. F. (1995). On definition and quantification of heterogeneity. *Oikos* 73: 280–284, doi:10.2307/3545921.

Lindley, D. (2004). That wretched prior. *Significance* 1: 85–87, doi:10.1111/j.1740-9713.2004.026.x.

Lloyd, C. and Atkinson, P. (2004). Archaeology and geostatistics. *Journal of Archaeological Science* 31: 151 – 165, doi:10.1016/j.jas.2003.07.004.

Lobo, J. M., Jiménez-Valverde, A. and Real, R. (2008). Auc: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography* 17: 145–151, doi:10.1111/j.1466-8238.2007.00358.x.

Luzzi, D. (2016). Communicating spatial uncertainty to non-experts using R. Master's thesis, Wageningen University, the Netherlands.

MacEachren, A., Robinson, A., Hopper, S., Gardner, S., Murray, R. B., Gahegan, M. and Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science* 32: 139 – 160, doi:10.1559/1523040054738936.

MacEachren, A. M. (2015). Visual Analytics and Uncertainty: Its Not About the Data. In Bertini, E. and Roberts, J. C. (eds), *EuroVis Workshop on Visual Analytics (EuroVA)*. The Eurographics Association, doi:10.2312/eurova.20151104.

Magalhães, R. J. S., Clements, A. C., Patil, A. P., Gething, P. W. and Brooker, S. (2011). Chapter 5 - the applications of model-based geostatistics in helminth epidemiology and control. Academic Press, Advances in Parasitology *74*, 267 – 296, doi:10.1016/B978-0-12-385897-9.00005-7.

Malone, B. P., McBratney, A. B., Minasny, B. and Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154: 138–152, doi:10.1016/j.geoderma.2009.10.007.

Malone, B. P., Minasny, B. and McBratney, A. B. (2017). Categorical soil attribute modeling and mapping. In *Using R for Digital Soil Mapping*. Springer, 151–167.

Marchant, B. P. and Lark, R. M. (2004). Estimating variogram uncertainty. *Mathematical Geology* 36: 867–898, doi:10.1023/B:MATG.0000048797.08986.a7.

Marchant, B. P. and Lark, R. M. (2007). Optimized sample schemes for geostatistical surveys. *Mathematical Geology* 39: 113–134, doi:10.1007/s11004-006-9069-1.

Martin, A. D., Quinn, K. M. and Park, J. H. (2011). MCMCpack: Markov chain Monte Carlo in r. *2011* 42: 21, doi:10.18637/jss.v042.i09.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting* 19: 1106–1114, doi:10.1175/825.1.

Mays, D. C., Faybishenko, B. A. and Finsterle, S. (2002). Information entropy to measure temporal and spatial complexity of unsaturated flow in heterogeneous media. *Water Resources Research* 38: 49–1–49–11, doi:10.1029/2001WR001185.

McBratney, A., Field, D. J. and Koch, A. (2014). The dimensions of soil security. *Geoderma* 213: 203 – 213, doi:10.1016/j.geoderma.2013.08.013.

McBratney, A. B., Minasny, B., Stockmann, U. et al. (2018). *Pedometrics*. Springer, doi:10.1007/978-3-319-63439-5.

McBratney, A. B., Odeh, I. O., Bishop, T. F., Dunbar, M. S. and Shatar, T. M. (2000). An overview of pedometric techniques for use in soil survey. *Geoderma* 97: 293 – 327, doi:10.1016/S0016-7061(00)00043-4.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC texts in statistical science series. Boca Raton, US: CRC Press/Taylor & Francis Group.

McGrayne, S. B. (2011). *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press.

McKinley, J. M. and Atkinson, P. M. (2020). A special issue on the importance of geostatistics in the era of data science. *Mathematical Geosciences* 52: 311–315, doi:10.1007/s11004-020-09858-1.

Miller, H. J. (2004). Tobler's first law and spatial analysis. *Annals of the Association of American Geographers* 94: 284–289.

Minasny, B., Vrugt, J. A. and McBratney, A. B. (2011). Confronting uncertainty in model-based geostatistics using Markov chain Monte Carlo simulation. *Geoderma* 163: 150–162, doi:10.1016/j.geoderma.2011.03.011.

Montgomery, D. C., Peck, E. A. and Vining, G. G. (2001). *Introduction to linear regression analysis*. John Wiley & Sons.

Moraga, P., Cramb, S. M., Mengersen, K. L. and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics* 21: 27–41, doi:10.1016/j.spasta.2017.04.006.

Mullen, K., Ardia, D., Gil, D. L., Windover, D. and Cline, J. (2011). DEoptim: An r package for global optimization by differential evolution. *Journal of Statistical Software* 40: 1–26, doi:10.18637/jss.v040.i06.

Müller, H. (2007). Bayesian transgaussian kriging. In *Proc. 15th European Young Statisticians Meeting*.

Mundry, R. and Nunn, C. L. (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. *The American Naturalist* 173: 119–123, doi:10.1086/593303.

Myers, R. H., Montgomery, D. C. and Vining, G. G. (2002). *Generalized linear models: with applications in engineering and the sciences*. John Wiley & Sons.

Nahin, P. J. (2020). *Gamma and Beta Function Integrals*. Cham: Springer International Publishing. 149–180, doi:10.1007/978-3-030-43788-6_4.

Neath, A. A. and Cavanaugh, J. E. (2012). The Bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics* 4: 199–203, doi:10.1002/wics.199.

Nowosad, J. and Stepinski, T. F. (2019). Information theory as a consistent framework for quantification and classification of landscape patterns. *Landscape Ecology* 34: 2091–2101, doi:10.1007/s10980-019-00830-x.

O'Hagan, A. (2012). Probabilistic uncertainty specification: Overview, elaboration techniques and their application to a mechanistic model of carbon flux. *Environmental Modelling & Software* 36: 35 – 48, doi:h10.1016/j.envsoft.2011.03.003, thematic issue on Expert Opinion in Environmental Modelling and Management.

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics, Vol. 2B: Bayesian Inference*. Taylor & Francis.

Oijen, M. van, Thomson, A. and Ewert, F. (2009). Spatial upscaling of process-based vegetation models: An overview of common methods and a case-study for the uk. *Methods* 1: 3.

Oliver, M. and Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* 113: 56–69, doi:10.1016/j.catena.2013.09.006.

Orton, T. G., Mallawaarachchi, T., Pringle, M. J., Menzies, N. W., Dalal, R. C., Kopittke, P. M., Searle, R., Hochman, Z. and Dang, Y. P. (2018). Quantifying the economic impact of soil constraints on Australian agriculture: A case-study of wheat. *Land Degradation & Development* 29: 3866–3875, doi:10.1002/ldr.3130.

Orton, T. G., Román Dobarco, M. and Saby, N. P. A. (2017). Kriging based on areal summary statistics data: Effects of within-unit variability on predictions and uncertainties. *Spatial Statistics* 19: 42–67, doi:10.1016/j.spasta.2016.11.003.

Orton, T. G., Saby, N. P. A., Arrouays, D., Walter, C., Lemercier, B., Schvartz, C. and Lark, R. M. (2012). Spatial prediction of soil organic carbon from data on large and variable spatial supports. I. Inventory and mapping. *Environmetrics* 23: 129–147, doi:10.1002/env.2136.

Papritz, A. (2009). Limitations of Indicator Kriging for Predicting Data with Trend. In *Proceedings StatGIS 2009*.

Pardo-Igúzquiza, E. and Dowd, P. (2001). Variance–covariance matrix of the experimental variogram: Assessing variogram uncertainty. *Mathematical Geology* 33: 397–419, doi:10.1023/a:1011097228254.

Park, N.-W. (2013). Spatial downscaling of TRMM precipitation using geostatistics and fine scale environmental variables. *Advances in Meteorology* 2013: 9, doi:10.1155/2013/237126.

PDOK (2020). Dataset: Actueel hoogtebestand Nederland (AHN2) [dataset: Current dutch elevation (AHN2)].

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences* 30: 683–691, doi:10.1016/j.cageo.2004.03.012.

Pebesma, E. J., Jong, K. de and Briggs, D. (2007). Interactive visualization of uncertain spatial and spatio-temporal data under different scenarios: an air quality example. *International Journal of Geographical Information Science* 21: 515–527, doi:10.1080/13658810601064009.

Peel, M. C., Finlayson, B. L. and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.* 11: 1633–1644, doi:10.5194/hess-11-1633-2007.

Pejović, M., Nikolić, M., Heuvelink, G. B., Hengl, T., Kilibarda, M. and Bajat, B. (2018). Sparse regression interaction models for spatial prediction of soil properties in 3d. *Computers & Geosciences* 118: 1 – 13, doi:j.cageo.2018.05.008.

Pierce, F. J., Schabenberger, O. and Crandall, M. (2004). *The Spatial Dimension: Geographic Information Systems and Geostatistics*. New York: Routledge. 1172–1174, doi:10.1201/9780203757604.

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. CRC Press.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6: 7–11.

Poggio, L., Gimona, A., Spezia, L. and Brewer, M. J. (2016). Bayesian spatial modelling of soil properties and their uncertainty: The example of soil organic matter in Scotland using R-INLA. *Geoderma* 277: 69–82, doi:10.1016/j.geoderma.2016.04.026, 69.

Pons, L. J. and Zonneveld, I. S. (1965). *Soil ripening and soil classification: initial soil formation of alluvial deposits with a classification of the resulting soils*, Publication / International Institute for Land Reclamation and Improvement *13*. Wageningen: Veenman.

Pruim, R. J. (2011). *Foundations and applications of statistics: an introduction using R*, *13*. American Mathematical Soc.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis* 24: 339–355, doi:10.1093/pan/mpw014.

Ribeiro, P. J. J. and Diggle, P. J. (2006). TECHNICAL REPORT ST-99-08: Bayesian inference in Gaussian model-based geostatistics. Report, Lancaster University.

Ribeiro Jr, P., Christensen, O. and Diggle, P. (2003). geoR and geoRglm: Software for Model-Based Geostatistics.

Roberts, G. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18: 349–367, doi:10.1198/jcgs.2009.06134.

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 60: 255–268.

Rosenthal, J. S. (2011). *Optimal proposal distributions and adaptive MCMC*. Chapman and Hall/CRC. 93–112.

Rosenzweig, C. and Parry, M. L. (1994). Potential impact of climate change on world food supply. *Nature* 367: 133–138, doi:10.1038/367133a0.

Rossiter, D. G., Zeng, R. and Zhang, G.-L. (2017). Accounting for taxonomic distance in accuracy assessment of soil class predictions. *Geoderma* 292: 118 – 127, doi:10.1016/j.geoderma.2017.01.012.

Roth, M. (2013). On the Multivariate t Distribution. Report, Linköpings universitet.

Roth, R. E. (2009). A qualitative approach to understanding the role of geographic information uncertainty during decision making. *Cartography and Geographic Information Science* 36: 315–330, doi:10.1559/152304009789786326.

Roudier, P., Sultan, B., Quirion, P. and Berg, A. (2011). The impact of future climate change on West African crop yields: What does the recent literature say? *Global Environmental Change* 21: 1073–1083, doi:10.1016/j.gloenvcha.2011.04.007.

Rowley, J. (2007). The wisdom hierarchy: representations of the dikw hierarchy. *Journal of Information Science* 33: 163–180, doi:10.1177/0165551506070706.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)* 71: 319–392, doi:10.1111/j.1467-9868.2008.00700.x.

Samaniego, F. J. (2010). *Conjugacy, Self-Consistency and Bayesian Consensus*. New York, NY: Springer New York. 99–113, doi:10.1007/978-1-4419-5941-6_6.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical methods for spatial data analysis*. Chapman and Hall/CRC.

Schabenberger, O. and Pierce, F. J. (2002). *Contemporary statistical models for the plant and soil sciences*. Boca Raton: CRC Press.

Schoot, R. van de and Miocević, M. (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Taylor & Francis.

Seaman, J. W., Seaman, J. W. and Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician* 66: 77–84, doi:10.1080/00031305.2012.695938.

Searle, S. R. (1971). *Linear models*. Wiley series in probability and mathematical statistics. New York: Wiley.

Searle, S. R. (1997). *Linear models*. Wiley series in probability and mathematical statistics. New York: Wiley, doi:10.1002/9781118491782.

Sengupta, A., Cressie, N., Kahn, B. H. and Frey, R. (2016). Predictive inference for big, spatial, non-gaussian data: Modis cloud data and its change-of-support. *Australian & New Zealand Journal of Statistics* 58: 15–45, doi:10.1111/anzs.12148.

Simmonds, E. G., Jarvis, S. G., Henrys, P. A., Isaac, N. J. B. and O'Hara, R. B. (2020). Is more data always better? a simulation study of benefits and limitations of integrated distribution models. *Ecography* 43: 1413–1422, doi:10.1111/ecog.05146.

Somarathna, P., Minasny, B. and Malone, B. P. (2017). More data or a better model? figuring out what matters most for the spatial prediction of soil carbon. *Soil Science Society of America Journal* 81: 1413–1426, doi:10.2136/sssaj2016.11.0376.

Spiegelhalter, D. J. (2019). *The art of statistics : learning from data*. Pelican, an imprint of Penguin Books.

Stanaway, M. A., Mengersen, K. L. and Reeves, R. (2011). Hierarchical Bayesian modelling of early detection surveillance for plant pest invasions. *Environmental and Ecological Statistics* 18: 569–591, doi:10.1007/s10651-010-0152-x.

Steinbuch, L., Brus, D. J. and Heuvelink, G. B. M. (2018). Mapping the probability of ripened subsoils using Bayesian logistic regression with informative priors. *Geoderma* 316: 56–69, doi:10.1016/j.geoderma.2017.12.010.

Steinbuch, L., Orton, T. G. and Brus, D. J. (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. doi:10.4121/UUID:1FE0C01E-7F67-435B-A240-800579ADC6E6.

Steinbuch, L., Orton, T. G. and Brus, D. J. (2020). Model-based geostatistics from a Bayesian perspective: Investigating area-to-point kriging with small data sets. *Mathematical Geosciences* 52: 397–423, doi:10.1007/s11004-019-09840-6.

Stichting voor Bodemkartering Wageningen (1969). Bodemkaart van Nederland schaal 1 : 50.000 : Toelichting bij kaartblad 31 west Utrecht (soil map of the Netherlands 1:50,000: explanation to map 31, west Utrecht).

Stigler, S. M. (2007). The epic story of maximum likelihood. *Statist. Sci.* 22: 598–620, doi:10.1214/07-STS249.

Storn, R. and Price, K. (1997). Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization* 11: 341–359, doi:10.1023/a:1008202821328.

Supit, I., Diepen, C. A. van, Wit, A. J. W. de, Wolf, J., Kabat, P., Baruth, B. and Ludwig, F. (2012). Assessing climate change effects on European crop yields using the crop growth monitoring system and a weather generator. *Agricultural and Forest Meteorology* 164: 96–111, doi:10.1016/j.agrformet.2012.05.005.

Taffoni, G., Tornatore, L., Goz, D., Ragagnin, A., Bertocco, S., Coretti, I., Marazakis, M., Chaix, F., Plumidis, M., Katevenis, M., Panchieri, R. and Perna, G. (2019). Towards Exascale: Measuring the Energy Footprint of Astrophysics HPC Simulations. In *2019 15th International Conference on eScience (eScience)*, 403–412.

Traore, S. B., Ali, A., Tinni, S. H., Samake, M., Garba, I., Maigari, I., Alhassane, A., Samba, A., Diao, M. B., Atta, S., Dieye, P. O., Nacro, H. B. and Bouafou, K. G. M. (2014). AGRHYMET: A drought monitoring and capacity building center in the West Africa region. *Weather and Climate Extremes* 3: 22–30, doi:doi.org/10.1016/j.wace.2014.03.008.

Truong, N. (2014). Expert knowledge in geostatistical inference and prediction. Ph.D. thesis.

Truong, P. N., Heuvelink, G. B. M. and Pebesma, E. (2014). Bayesian area-to-point kriging using expert knowledge as informative priors. *International Journal of Applied Earth Observation and Geoinformation* 30: 128–138, doi:10.1016/j.jag.2014.01.019.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7: 91, doi:10.1186/1471-2105-7-91.

Vasques, G. M., Demattê, J. A. M., Viscarra Rossel, R. A., Ramírez-López, L. and Terra, F. S. (2014). Soil classification using visible/near-infrared diffuse reflectance spectra from multiple depths. *Geoderma* 223–225: 73–78, doi:10.1016/j.geoderma.2014.01.019.

Vieira, S., Hatfield, J., Nielsen, D. and Biggar, J. (1983). Geostatistical theory and application to variability of some agronomical properties. *Hilgardia* 51: 1–75, doi:10.3733/hilg.v51n03p075.

Vries, F. de, Walvoort, D. and Brouwer, F. (2017). Basisregistratie Ondergrond (BRO) : Actualisatie bodemkaart van eenheden met slappe kleilagen [Key Register of the Subsurface : Updating soil map unripened clay units]. Report, Wageningen Environmental Research.

Wackernagel, H. (2014). Geostatistics. *Wiley StatsRef: Statistics Reference Online* doi:10.1007/978-3-662-05294-5.

Wadoux, A. M. J.-C., Samuel-Rosa, A., Poggio, L. and Mulder, V. L. (2020). A note on knowledge discovery and machine learning in digital soil mapping. *European Journal of Soil Science* 71: 133–136, doi:10.1111/ejss.12909.

Walvoort, D. J. J., Brus, D. J. and Gruijter, J. J. de (2010). An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences* 36: 1261–1267, doi:10.1016/j.cageo.2010.04.005.

Wart, J. van, Bussel, L. G. J. van, Wolf, J., Licker, R., Grassini, P., Nelson, A., Boogaard, H., Gerber, J., Mueller, N. D., Claessens, L., Ittersum, M. K. van and Cassman, K. G. (2013a). Use of agro-climatic zones to upscale simulated crop yield potential. *Field Crops Research* 143: 44–55, doi:10.1016/j.fcr.2012.11.023.

Wart, J. van, Grassini, P. and Cassman, K. G. (2013b). Impact of derived global weather data on simulated crop yields. *Global Change Biology* 19: 3822–3834, doi:10.1111/gcb.12302.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley and Sons.

White, P., Gelfand, A. and Utlaut, T. (2017). Prediction and model comparison for areal unit data. *Spatial Statistics* 22: 89–106, doi:10.1016/j.spasta.2017.09.002.

Wolf, J. and Diepen, C. A. (1995). Effects of climate change on grain maize yield potential in the european community. *Climatic Change* 29: 299–331, doi:10.1007/bf01091866.

Wolf, J., Hessel, R., Boogaard, H., Wit, A. de, Akkermans, W. and Diepen, K. van (2011). *Modeling Winter Wheat Production across Europe with WOFOST—The Effect of Two New Zonations and Two Newly Calibrated Model Parameter Sets*, Advances in Agricultural Systems Modeling. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America. 297–326, doi:10.2134/advagricsystmodel2.c11.

You, L., Wood, S. and Wood-Sichra, U. (2006). Generating global crop distribution maps: from census to grid. In *Selected paper at IAEA 2006 Conference at Brisbane, Australia*, *202*, 1–16, doi:10.1016/j.agsy.2014.01.002.

You, L., Wood, S. and Wood-Sichra, U. (2009). Generating plausible crop distribution maps for Sub-Saharan Africa using a spatially disaggregated data fusion and optimization approach. *Agricultural Systems* 99: 126–140, doi:10.1016/j.agsy.2008.11.003.

Zhang, G., Liu, F. and Song, X. (2017). Recent progress and future prospect of digital soil mapping: A review. *Journal of Integrative Agriculture* 16: 2871 – 2885, doi:10.1016/S2095-3119(17)61762-3.

Zhang, L., Datta, A. and Banerjee, S. (2018). Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments. *arXiv e-prints* : arXiv:1802.00495.

Zhu, L. and Gigerenzer, G. (2006). Children can solve bayesian problems: the role of representation in mental computation. *Cognition* 98: 287 – 308, doi:10.1016/j.cognition.2004.12.003.

Zoest, V. van, Osei, F. B., Hoek, G. and Stein, A. (2019). Spatio-temporal regression kriging for modelling urban NO2 concentrations. *International Journal of Geographical Information Science* 0: 1–15, doi:10.1080/13658816.2019.1667501.

Summary

Soils and other agronomic variables can be mapped using various methods. The general focus is nowadays on machine learning (also called ‘data mining’) methods, relying on numerical approaches and abundant data – among others from remote sensing. However, while machine learning methods undoubtedly have their benefits, they also have shortcomings, such as finding irrelevant relationships, obscuring causalities and natural laws, dependence on abundant data and limited options for change of support situations. Therefore, to meet the needs of the future, we also need to think about smart, conscious and statistically consistent ways to combine and process data: data from different sources, with different uncertainty, with different spatial support, in a setting where data is not abundant. We also need ways to combine new and legacy data, including existing beliefs in a qualitative formulation. Model-based geostatistics provides tools to deal smartly, consciously and statistically sound with spatial data, and enables – unlike most machine learning methods – the integration of varying spatial support data into one stochastic model.

Geostatistics is part of the broader field of spatial statistics, and was originally designed for geological and other environmental data, having its own terminology and workflow. Geostatistics describes spatial phenomena with the help of spatial correlation, which is modelled by considering observed reality as one realisation of a spatial random process; spatial random processes are characterised by Tobler’s first law of Geography: “Everything is related to everything else. But near things are more related than distant things”.

Model-based geostatistics aims at explicitly applying formal statistical methods into the field of geostatistics. Using model-based geostatistics one can estimate or infer model parameters as well as predictions using sound methods developed in general statistics, where – apart from model choice and related assumptions – no subjective choices have to be made. Bayesian geostatistical models are part of the repertoire of model-based geostatistics.

Bayesian statisticians use probability to express their knowledge – or inversely formulated: their ignorance – about the world. By gathering data (observing events, doing experiments, etc.) their knowledge and thus the associated probability changes – hence the former name ‘inverse probability’. More mathematically, the combination of pre-observation knowledge (expressed in a ‘prior’ probability distribution) with gathered data will produce a ‘posterior’ probability distribution of the variables of interest. Such models can be extended with prediction, also as probability distributions (the ‘posterior predictive’). The Bayesian extension of model-based geostatistics facilitates an accurate estimation of parameter and prediction uncertainty because of the consistent underlying model, while allowing incorporation of pre-observation knowledge into the final results. Even more, Bayesian statistics provides a coherent framework to build a hierarchical statistical model, consisting of several layers. It does so by modelling variables which are related in one layer, while these relationships are expressed in other variables in another layer, but conveniently being part of the same model; a third layer contains prior probability distributions. This approach facilitates hierarchical model-based geostatistics, where the spatial random field forms one layer in the model, the

geostatistical model parameters form another layer, and the priors of those model parameters are the 3rd layer.

This thesis aims to feature and advance current developments in model-based geostatistics as well as Bayesian statistics in a spatial context to be used for the needs of today and the future, and also bring these developments in line with contemporary computational possibilities. Therefore, in this thesis I explore the methods and limitations of model-based geostatistics in the context of mapping soil properties and crop yields.

Following the general introduction (**Chapter 1**, in which I present the main objectives of this PhD-research), **Chapter 2** shows an example of the added value of using model-based geostatistics, and the prediction (including uncertainty) of yield gaps based on point data, for a case study with sorghum and millet in West-Africa. I used crop growth model outputs to calibrate a linear regression model using environmental covariates as predictors. The spatial regression residuals were investigated for spatial correlation. The linear regression model and the spatial correlation of residuals together were used to predict theoretical crop yield at all locations using kriging with external drift. A spatial standard deviation comes along with this prediction, indicating the uncertainty of the prediction. In combination with land use data and country borders, I summed the potential crop yield predictions to determine an area total. With spatial stochastic simulation, I estimated the uncertainty of that total production potential as well as the spatial cumulative distribution function, and I compared my results with the prevailing agro-ecological Climate Zones approach used for spatial aggregation. I concluded that the geostatistical approach can estimate a country's crop production, including a quantification of uncertainty. Using model-based geostatics offers important benefits for crop modelling scientists to explore relationships between yields and spatial environmental variables, and also assist policy makers with tangible results on yield gaps at multiple levels of spatial aggregation.

Chapter 3 explores the use of legacy information to improve the accuracy of a prediction map. I used Bayesian binomial logistic regression (BBLR) to update the map showing unripened subsoils for a reclamation area in the west of The Netherlands. Similar to conventional binomial logistic regression (BLR), in BBLR the binary target variable (i.e. the subsoil is either ripened or unripened) is modelled by a Bernoulli distribution. The logit transform of the 'probability of success' parameter of the Bernoulli distribution was modelled as a linear combination of the covariates *soil type*, *freeboard* (the desired water level in the ditches, compared to surface level) and *mean lowest groundwater table*. My research focused on quantifying the influence of informative prior distributions (inferred from legacy data) with different information levels, in combination with different sample sizes, on the resulting parameters and maps. I combined subsamples of different size (ranging from 5% to 50% of the original dataset of 676 observations) with priors representing different levels of trust in legacy data and investigated the effect of sample size and prior distribution on map accuracy. The resulting posterior parameter distributions, calculated by Markov chain Monte Carlo simulation, vary in centrality as well as in dispersion, especially for the smaller datasets. More informative priors decreased dispersion and pushed posterior central values towards prior central values. The resulting probability maps were almost similar. However, the associated uncertainty maps (showing the uncertainty of the probability parameter) were different: a more informative prior decreased prediction uncertainty. When using the ‘overall accuracy’ statistical validation metric, I found – for this case study – an optimal value for the prior information level, expressed in a variance multiplication factor. The effect of incorporating informative priors however is only detectable for smaller datasets. Bayesian binomial logistic regression proved to be a flexible mapping tool but the accuracy gain compared to conventional logistic regression was marginal and may not outweigh the extra modelling and computing effort.

I investigated the accuracy of prediction uncertainties in case of sparse data when model-based geostatistics is applied on an area-to-point kriging (ATPK) situation, illustrated with disaggregating millet crop yields in Burkina-Faso in **Chapter 4**. ATPK is a geostatistical method for creating high resolution raster maps using data of the variable of interest with a much lower resolution. However, the dataset of areal means is often considerably smaller (< 50 observations) than datasets 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, I developed a partly analytical solution, thoroughly explained and implemented in this chapter, in order to (i) explore the impact of the prior distribution on predictions and prediction error variances, (ii) investigate whether certain aspects of uncertainty can be disregarded, simplifying the necessary computations, and (iii) test the impact of various model misspecifications. I compared several approaches using simulated data, aggregated real-world point data, and a case study on aggregated crop yields in Burkina Faso. I found that the prior distribution has 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, I found some severe effects of model misspecification in terms of overly conservative or optimistic prediction uncertainties, highlighting the importance of covariance model choice.

I explored and explained an existing implementation of a Bayesian generalised linear geostatistical model (BGLGM) including possible issues and their solutions in **Chapter 5**. Using the depth of the Pleistocene sand layer in the Dutch province of Flevoland, with the depth reduced to a binary variable, I compared the BGLGM approach with the far less complicated Bayesian generalised linear model (BGLM, almost equal to the BBLR model used in Chapter 3). In general, for mapping binary spatial variables using a BGLM might be a solution if relevant environmental covariates are available. The geostatistical extension BGLGM adds spatial dependence and is thus potentially better equipped. I found that BGLGM yields considerably better statistical validation metrics compared to BGLM, especially with – as in this case – a large ($n = 1000$) observation sample but few relevant covariates. Also, the inferred posterior BGLGM spatial parameters enable the quantification of spatial relationships. However, calibrating and applying a BGLGM (as implemented in the R-package `geoRglm`) is quite demanding with respect to the minimal required sample size, tuning the algorithm, and computational costs. I replaced manual tuning by an automated tuning algorithm (which eases implementing applications) and found a sample composition that delivers meaningful results within 50 hrs calculation time. With the gained insights spatial soil practitioners and researchers can – for their specific cases – evaluate if using BGLGM is feasible and if the gain is worth the extra effort.

The final **Chapter 6** discusses and brings together the achievements as presented in this thesis, along with my recommendations for future research. Mentioned are future extensions of the applied change of support, such as Bayesian area-to-area kriging and application of data-fusion methods for combing data with different support. I also discuss advantages and challenges regarding using informative priors in spatial mapping situations. In my opinion, informative priors must be seen in relationship to the related but distinctive perceptions on data, information, knowledge and wisdom; developing and using informative priors thus connects to research fields such as information theory, communication and psychology. Informative priors can be useful in combination with small data situations. Small data situations can cause problems already on the level of algorithms, and thus inhibit any analysis because there are no calculation results at all to be judged. Future research might focus on minimal data requirement to get algorithms running combined with minimal data requirement to arrive at meaningful results – and the influence of an informative prior on both those properties. More general is the discussion if the soil- and crop science community are actually interested in the tools and possibilities delivered by model-based geostatistics and its Bayesian extensions. In my opinion, the development of easily accessible applications (rather then – for outsiders – somewhat obscure R-packages) as well as easily accessible explanations of the used principles would help practitioners and fellow scientists (having perhaps less focus on spatial statistical modelling) to be aware of the possibilities, benefits and drawbacks of those tools.

Acknowledgements

Among PhD candidates it is a well-known hypothesis that the acknowledgements are – perhaps together with the propositions and some other personal stuff – the only part of the thesis that is actually read by the majority of the receivers, of which some might be looking for their own name. I wondered if I should facilitate this by putting all the names I want to mention in alphabetical order. Other options are the more usual approach of going from formal to informal connections, or mention everybody to be acknowledged during a more descriptive part of a PhD journey. I chose to group people; as I dealt with (perceived) randomness as study subject while creating this thesis, I used the pseudo-randomness implemented in the programming language R to give some groups – and most names inside groups – a pseudo-random order.

During my PhD time, I had two experiences that made me feel like almost living in a parallel world. The first one was in autumn[3] 2018, when I was on the other side of the Earth for three months. To be exact: I was in Brisbane (Queensland, Australia) with its sunny weather, well-organised public life, and extremely friendly people everywhere. Thanks to Tom Orton, Diane and all others involved to make this such a wonderful experience! And also thanks to Ecosciences Precinct, part of Queensland Government and the University of Queensland, department of Plant Sciences, for formally making this possible.

The second parallel world experience is at the moment of writing, spring and winter 2020, at home in Wageningen. I don't know how history will name this event; at this moment we call it the "Corona crisis": almost the whole world, for the first time in over a century, is in some kind of lock-down situation because of a virus. Instead of physically being somewhere in the Gaia bulding on Wageningen Campus, I am working from home, not allowed to meet anybody face-to-face unless there is a very good reason to do so – and even then, on a physical distance. I thank my housemates (Loena, Daan, Thijs, Angelica, Cecilia, Luka, Katharina and Maartje) for this both intense and interesting life experience: being almost locked up in the same accommodation for several months. We got to know each other well, and mostly I enjoyed it.

I would like to mention my fellow PhD candidates and postdocs of the SGL chairgroup and/or closely related: Aukjen, Jalal, Femke, Abbey, Yingxia, Chantal, Anatol, Kasia, Anne-Maartje, Marcos, Cynthia, Rafael, Tijn, Alessandro, Rowin, Rocky, Selçuk, Maricke, Luciana, Marijn, Nienke, Noortje, Alexandre, Cindy, Jasper, Simona, Stephan, Bertin and Neymar. Some of you I met only once; one of you I met only once while I was a witnesses on your wedding; with others I shared an office for several years – and thus important parts of our life. Most of you are somewhere in between. Thank you all for being such nice close colleagues, and for the possibility to share experiences, emotions, opinions, courses and nice moments! It is a real pity that in those last half year we didn't have more coffeebreak & roomies' talk. I also want to mention my non-SGL Wageningen PhD colleagues for sharing experiences, for example Roos (Farming Systems Ecology) and Heleen (Animal Production Systems) – the latter I would also like to thank for being one of my paranymphs and proofreading this thesis. Still within the context of Wageningen PhD's, I would like to mention Benjamin and Dainius (both

[3] That is, autumn according to the Northern hemisphere

Laboratory of Geo-information Science and Remote Sensing) for smoothly running together the R-user group during several years.

Furthermore, I am also grateful for the insight into being a PhD candidate in a slightly different system, delivered by UK-based PhD candidates Maud and Doris. Also many thanks to you both for the cosiness during my almost yearly visits and the beautiful walks – for example along the many canals in the English Midlands.

I seem to be a person with many extensive but long-term friendships. All of you were interested in my well-being as PhD candidate, and thus I would like to mention – without the intention of being complete: Lianne, Hanna, Mathijs + Marlies, Audrey, Indira, Max, Flora, Eliane, Louise, Marije, Mali, Mark, Hedwig, Solange, Sanne, John and Jasmijn, Marleen, Irene, Kari-Anne, Joanne, Ireen, Annemarie, Jenneke, Hanneke, Karin, Ellen, Thomas, Mirre, Elanka, Caroline, Arta, Maaike, Yvonne, Keri + Johannes, Mieke, Anna, Silvia, Claudia, Boudien, Anne, Lilianne, Chantal, Femke, Anita, Cor + Justine, Anneloes, Carolina and many others. My gratitude goes also to Didi, for being the second proofreader and also for keeping me updated about science in general. Another special thanks goes to Zoë for your continuous support during tough shared life experiences, and for your willingness to be the other paranymph.

I also want to thank my colleagues in the three groups I was part of: the Soil, Geography and Landscape group; ISRIC – World Soil Information; and the Soil, Water and Landuse team of WEnR, for the nice coffee talks and the other social and/or content-wise interactions. On a more formal level, I am also grateful to the mentioned organisations for offering me this PhD in the first place. Thanks to all the others I encountered within Wageningen UR, among others Cor Langeveld, Karel Keesman and Lenny van Bussel.

An important source of support is my family: my parents Rob and Kitty, my brothers Reinier and Maarten, my sister Heleen and my siblings' numerous offspring. I thank you all for your interest and ideas!

My mental health during my PhD has certainly improved by the organisations I am part of or dealt with, such as the SFO, Idealis and Droevendaal student housing. I am happy with the many nice people I encountered.

Thanks to Martin Knotters, Titia Mulder, Sytze de Bruin and the others involved for making it possible to practise my defence. I also want to express my gratitude to the real opponents, for their time and effort.

Setting deadlines and sticking to those certainly is not my quality within a PhD context. Therefore I admire the patience of my supervisors cq. promotor and co-promotor Gerard Heuvelink and Dick Brus, and in the background Jakob Wallinga in his role as SGL group leader. You kept trust in me and in the final result. This thesis proves your trust to be justified.

Zum Abschluss: Liebe Almuth, vielen Dank für deine Hilfe beim Layout dieser Arbeit! Es ist etwas Besonderes, dass wir uns seit fast 30 Jahren kennen. Wollen wir dem noch mindestens 30 Jahre hinzufügen?

About the author

I was born in 1969 in Driebergen, The Netherlands. After finishing a practical education for biodynamic farming at the Warmonderhof (currently located in Dronten, The Netherlands), I came to Wageningen in 1993 to study Agricultural Engineering with emphasis on both ‘system control’ and ‘soil tillage in organic agriculture’. Later my academic education continued with a second master Organic Agriculture (with emphasis on technology-society interaction and education & communication). In-between my master studies, I was employed at the Louis Bolk Instituut (currently located in Bunnik, The Netherlands), both as soil & compost researcher and as computer system administrator. A large part of my work-life I was, and still are, building tailor-made complex databases and related web-based applications as part-time entrepreneur, with focus on small organisations.

This thesis gives a nice impression of my scientific contribution to the field of geostatistics. For me an important part of science is to share findings and concepts, and if possible, make them visually attractive.

I was involved as a volunteer for many years at the Wageningen organic student association (StEL), and at this moment I am involved in representing Wageningen student tenants in the student flat organisation (SFO). This includes being active in the mini-village (some call it a community) where I happen to live: Droevendaal.

My interests include sustainability and sustainable travel. I managed to use only train, ferry, bus and bicycle for all seven conferences I went to in the context of this PhD. The only exception – when I had no reasonable alternative to flying – was my research visit in Australia. Perhaps, in the future, each Wageningen PhD thesis comes with a small obligatory section ‘*Environmental impact of this PhD project*’?

One of my more playful intentions related to my PhD was to visit all university libraries in The Netherlands. Because of COVID-19, I did not succeed, but hopefully I can fulfil this quest in the future.

PE&RC Training and Education Statement

With the training and education activities listed below the PhD candidate has complied with the requirements set by the C.T. de Wit Graduate School for Production Ecology and Resource Conservation (PE&RC) which comprises of a minimum total of 32 ECTS (= 22 weeks of activities).

Review of literature (4.5 ECTS)

- Geostatistics, and Bayesian algorithms for spatial inference

Postgraduate courses (9.3 ECTS)

- Bayesian statistics; PE&RC, SENSE (2014)
- Basic statistics; PE&RC (2014)
- An introduction to Bayesian computing with INLA; geoENV, Paris (2014)
- Advanced stochastic simulations; geoENV, Paris (2014)
- Spatial sampling for mapping; PE&RC, SENSE (2015)
- Linear models; PE&RC, SENSE (2016)
- Generalized linear models; PE&RC, SENSE (2016)
- Mixed linear models; PE&RC, SENSE (2016)
- Statistical uncertainty analysis of dynamic models; PE&RC, SENSE (2017)
- R and big data; PE&RC, SENSE (2017)
- Machine learning for spatial data; PE&RC, SENSE (2018)

Laboratory training and working visits (15 ECTS)

- Bayesian area-to-point kriging; Queensland Government, Department of Environment and Science + University of Queensland; Brisbane, Australia (2018)

Invited review of (unpublished) journal manuscript (2 ECTS)

- International Journal of Geographical Information Science: spatio-temporal regression kriging (2019)
- SOIL (Copernicus): spatial variability and sampling density (2019)

Competence strengthening / skills courses (9.15 ECTS)

- Effective behaviour in your professional surroundings; WGS (2014)
- Competence assessment; WGS (2015)
- Mental coaching; WGS (2015)
- Scientific writing; Wageningen in'to Languages (2015)
- Efficient writing strategies; Wageningen in'to Languages (2016)
- Techniques for writing and presenting a scientific paper; WGS (2016)
- The essentials of scientific writing and presenting; Wageningen in'to Languages (2016)
- Reviewing a scientific paper; Wageningen in'to Languages (2016)
- Scientific artwork with Photoshop and Illustrator; Wageningen UR Library (2017)
- Adobe Indesign essential training; Wageningen UR Library (2017)
- Infographics and iconography; Wageningen UR Library (2019)
- Introduction git version control; Biometris (2017)
- Linux basic course; WUR - Facilities & Services and Shared Research Facilities (2019)
- HPC basic course; WUR - Facilities & Services and Shared Research Facilities (2019)

PE&RC Annual meetings, seminars and the PE&RC weekend (2.1 ECTS)

- PE&RC First years weekend (2014)
- WGS PhD Workshop carousel (2014)
- Wageningen PhD symposium (2015)
- PE&RC Last years weekend (2019)

Discussion groups / local seminars / other scientific meetings (12 ECTS)

- Modelling and statistics network (2014-2017)
- R Users meeting (2014-2020)
- Modelling and simulation discussion group (2019-2020)

International symposia, workshops and conferences (16.8 ECTS)

- 14th European Society of Agronomy Congress; Edinburgh (2016)
- Pedometrics; Wageningen (2017)
- EGU General Assembly; Vienna (2018)
- ISBA World Meeting; Edinburgh (2018)
- GeoENV; Belfast (2018)
- Wageningen Soil Conference; Wageningen (2019)

Lecturing / supervision of practicals / tutorials (11.1 ECTS)

- Envrionmental data collection and analysis (2014-2019)
- Spatial and temporal analysis for earth and environment (2019)

The C.T. De Wit
Graduate School
PE&RC

PRODUCTION
ECOLOGY
& RESOURCE
CONSERVATION

The research described in this thesis was co-funded by:
1) the IPOP project *Sustainable and Smart Food Supply*, supported by Wageningen University & Research, via ISRIC - World Soil Information;
2) the *Key Registry of the Dutch Subsurface*[1] project, supported by the Dutch Ministry of the Interior and Kingdom Relations and the Dutch Ministery of Agriculture, Nature and Food Quality, via Wageningen Environmental Research (former Alterra);
3) 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) in combination with the Grains Research and Development Corporation (GRDC) of Australia under project PROC-9175385, via the University of Queensland, Australia; and
4) the LEB Foundation, Wageningen.

[1] in Dutch: Basisregistratie Ondergrond

Cover design in cooperation with Almuth Jung

k
f(y|...)
p(Θ|E)
Θ
ξ_0
p(β,y)
Λ
μ = β
s_i2
h(s_i1, s_i2)
X
f_p(...|...)
p(y|β)
N
ln(σ^2)