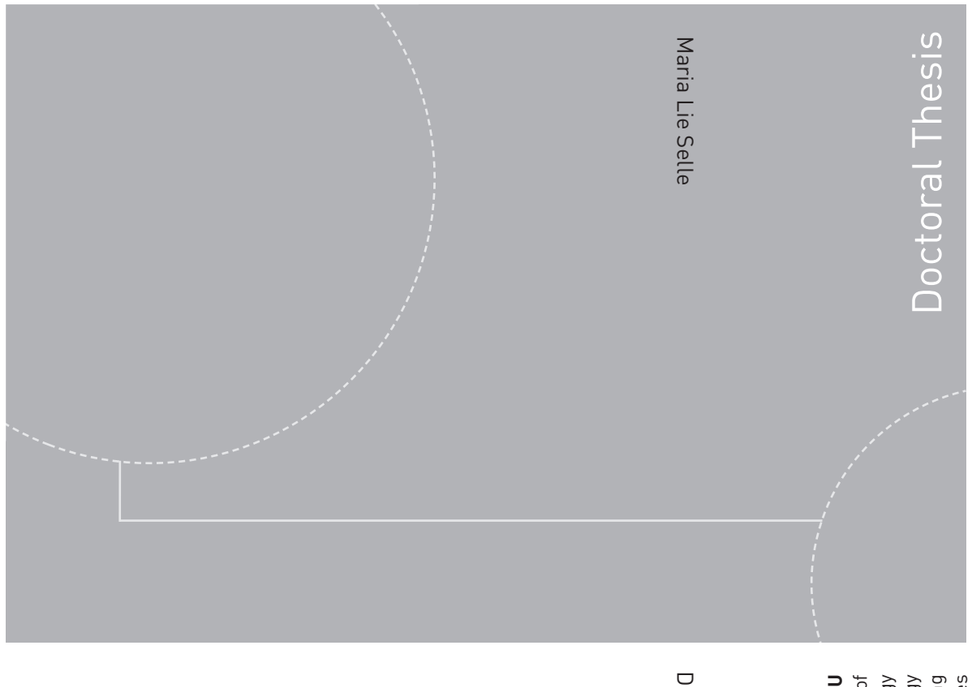
ISBN 978-82-326-4782-8 (printed version)
ISBN 978-82-326-4783-5 (electronic version)
ISSN 1503-8181



Doctoral theses at NTNU, 2020:217

Maria Lie Selle

Novel statistical variance and dependency models in quantitative genetics

Enabled by recent inference methods

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Doctoral theses at NTNU, 2020:217

NTNU

NTNU
Norwegian University of
Science and Technology

Maria Lie Selle

Novel statistical variance and dependency models in quantitative genetics

Enabled by recent inference methods

Thesis for the degree of Philosophiae Doctor

Trondheim, June 2020

Norwegian University of Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the degree of Philosophiae Doctor

Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

ISBN 978-82-326-4782-8 (printed version)
ISBN 978-82-326-4783-5 (electronic version)
ISSN 1503-8181

Doctoral theses at NTNU, 2020:217

Printed by Skipnes Kommunikasjon as

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of philosophiae doctor (PhD) at the Norwegian University of Science and Technology (NTNU), Trondheim, Norway. The research was funded by the The Research Council of Norway, and the work was carried out at the Department of Mathematical Sciences at NTNU during the years 2016-2020, with the exception of a research stay at the Roslin Institute in Edinburgh during 2019.

I am grateful to many people for their help and support during the four years I have spent working with this PhD thesis. I am particularly grateful for my main supervisor Ingelin Steinsland. Your experience and advice for the work and completion of this thesis has been invaluable. I would also like to thank my co-supervisors. Thank you Lars Rönnegård, for including me in an exciting project and for inviting me to Uppsala. I am also grateful to all co-authors, for what I think have been good collaborations leading to the papers in this thesis.

I would like to offer my special thanks to Gregor Gorjanc. Although you have not formally been my co-supervisor, you have been among my most important collaborators. Thank you for proposing interesting research projects and patiently teaching me about concepts within quantitative genetics. I am very grateful for my visit to the Roslin Institute and the AlphaGenes group, where I had some inspiring and educational months.

Assistance provided by the technical and administrative groups at the department has been greatly appreciated. I would not have been able to perform my numerical experiments without the help of the technical staff.

I would like to thank past and current PhD candidates at the department, many of whom I consider dear friends, for making the past years enjoyable and memorable. Finally, I would like to thank my friends, family, and Kristoffer for their support and encouragement.

Maria Lie Selle
Trondheim, March 2020

Contents

Chapter 1

Introduction

Statistics is "the science of learning from data, and of measuring, controlling and communicating uncertainty" (Davidian and Louis, 2012). In statistics, the goal is often to utilize observations of some sort to gain insight about a process of interest, or to do prediction related to this process, by building statistical models that represent the data-generating process. To be able to learn about a process from data we often want a model that is realistic, interpretable, and possible to draw inference from with the available data. For most realistic processes, compromises between these three are necessary. We can simplify the model, include more knowledge, add restrictions, or get more data. There can be several sources of information about a process, and to make good predictions one should use all the sources, as well as prior knowledge about (parts) of the process.

In this thesis we aim to use knowledge-based statistical methods. The word knowledge-based refers to something founded on an accumulation of facts or information. By knowledge-based statistical methods, we mean methods that aim to include the understanding of a phenomenon, such as facts, information, or descriptions, acquired through experience or by perceiving. For the use of prior knowledge, the Bayesian framework has appealing properties, and we also see that by the choice of model, a frequentist approach can be suitable for including knowledge.

Advances in technology have radically changed how much and in which ways data are collected. There has been a rapid increase in digital data collection for example through the use of mobile phones and other smart devices, geospatial technology using global positioning systems (GPS) and

geographical information systems, and whole-genome DNA sequencing. In this thesis we show how knowledge about farm locations available from GPS can enhance separation of different effects, how knowledge about genomic sequences from DNA sequencing can be used to construct genomic relations, and how knowledge about regions and processes in the genome from previous studies can be used identify important effects in different genomic regions. We show that by including this knowledge, we can make more accurate and consistent predictions.

In plant and animal breeding, artificial selection is used to improve the traits of plants and animals (Bourdon and Bourbon, 2000; Acquaah, 2009). To improve populations through selection, breeding designs and statistical methods are required to identify and utilize genetic differences between individuals for the traits of interest (Bourdon and Bourbon, 2000; Acquaah, 2009; Isik et al., 2017). Breeding programs are based on the principle that an individual's trait can provide insight to its underlying genetic value (Lynch and Walsh, 1998), and genetic and genomic evaluations involve statistical models for estimation of this genetic value, by using relevant data such as trait measurements, covariates and relationship with other individuals. The amounts of data available can be vast, so computationally fast and user-friendly methods are necessary.

This thesis develops new and combine existing statistical models and model components for variance and dependency within quantitative genetics, that allow inclusion of knowledge about the processes of interest. The models are motivated by and applied to challenges in plant and breeding, with the aim to contribute to genetic and genomic evaluation. With the enclosed papers we try to make contributions towards improving predictions of genetic effects, by proposing models that are closer to the underlying data-generating processes. Our working hypothesis has been that we can improve these predictions by using models that include knowledge about genetic and spatial processes, and that are aligned with current scientific understanding, to the extent of what is possible with available data and existing inference methods.

However, this does not mean that we aim to make "true" mathematical models to describe the processes of interest, as there is no such thing, wisely stated by Box et al. (1987) ("All models are wrong, but some are useful"), and Steyerberg et al. (2019) ("We recognize that true models do not exist. ... A model will only reflect underlying patterns, and hence

should not be confused with reality"). Further, our goal is to make predictions, not explain, which means that a simple but less true model can be a better model than a truer but less simple model (Hagerty and Srinivasan, 1991; Shmueli et al., 2010).

There are three recent statistical contributions in particular that have made inference with the proposed models possible. These are (i) the integrated nested Laplace approximations, an approximate Bayesian inference scheme introduced in Rue et al. (2009), (ii) the ability to represent Gaussian random fields on manifolds as Gaussian Markov random fields using stochastic partial differential equations (Lindgren et al., 2011), and (iii) the hierarchical likelihood for fitting hierarchical generalized linear models (Lee and Nelder, 1996). These have enabled us to do inference with the proposed models. Further, novel software for breeding program simulation (Gaynor et al., 2019), have enabled us to validate the predictions from the proposed models in settings close to reality, rather than validating the inference methods on data generated from the proposed models.

This introduction defines relevant concepts, reviews background material and provides more details than the paper format allows. The rest of this chapter is organized as follows. Section 1.1 presents the class of linear mixed models, and how prior knowledge can be included about the model parameters in both a Bayesian setting and a frequentist setting. Section 1.2 covers the two methods for performing statistical inference that are used in the thesis. Section 1.3 presents model selection, and the concept of simulation as experimentation. Section 1.4 introduces Gaussian random fields, Gaussian Markov random fields, and the explicit link between them. Section 1.5 gives a brief overview over models used in spatial statistics. Section 1.6 gives an introduction to quantitative genetics and breeding, relevant for the applications in the papers.

The main contributions of this doctoral thesis are contained in the four enclosed papers. In Chapter 2, summaries of the papers are given, identifying the scientific contributions and how they are related.

1.1 Linear mixed models

The class of models known as generalized linear mixed models (GLMM) (Fahrmeir et al., 2013) provides a range of models for the analysis of grouped data, where the differences between groups can be modeled as random effects. A subclass of GLMM are the linear mixed models (LMM), the class of models used in this thesis where the response variable is assumed to come from a Gaussian distribution. Following Fahrmeir et al. (2013), we give an introduction to LMM.

A LMM can be defined through stages in a hierarchical manner. In the first stage, the response variables are assumed to be linearly dependent on fixed effects, that are constant across subjects, and random effects, that vary across subjects. Let $(y_i, \boldsymbol{w}_i^T)$ denote the values of the response variable y and a vector of covariates $\boldsymbol{w}$ for subject i, where $i = 1, \ldots, n$. The first stage is then defined through the measurement model

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + \boldsymbol{u}_i^T \boldsymbol{\gamma} + \varepsilon_i,$$

where $\boldsymbol{\beta}$ is a vector of $(p+1)$ fixed covariate effects with vector of covariates $\boldsymbol{w}_i^T = (1, w_{i1}, \ldots, w_{ip})$, $\boldsymbol{\gamma}$ is a vector of q random effects with design vector $\boldsymbol{u}_i^T = (u_{i1}, \ldots, u_{iq})$, and ε_i is the residual term assumed to be independent and identically distributed as $\mathcal{N}(0, \sigma^2)$. The general form of the LMM in matrix notation is

$$\boldsymbol{y} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y}$ is an $n \times 1$ vector of the observations, $\boldsymbol{W}$ is an $n \times (p+1)$ matrix of covariates, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of fixed-effects coefficients including a common intercept, $\boldsymbol{U}$ is an $n \times q$ design matrix, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of random effects, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of the residuals.

In the second stage, it is assumed that the random effects are realized values of a random variable distributed according to some probability distribution, for example according to a Gaussian distribution, $\boldsymbol{\gamma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\sigma_\gamma^2)$.

In this thesis we have proposed models in both the Bayesian and frequentist framework that fit in the class of LMM. The next two sections present the LMM, first from a Bayesian perspective, and then an extension of the LMM in a frequentist framework that allows inclusion of prior knowledge on random effects.

1.1.1 Bayesian setting

Here, the LMM with two variance components is discussed from a Bayesian perspective following Sorensen and Gianola (2002) and Gelman et al. (2013). Again, the first stage of the LMM consists of the measurement model, here specified as a likelihood

$$\pi(\boldsymbol{y}|\boldsymbol{\eta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{I}\sigma^2),$$

where the linear predictor $\boldsymbol{\eta} = \boldsymbol{W}\boldsymbol{\beta} + \boldsymbol{U}\boldsymbol{\gamma}$ is often called the latent field, and σ^2 is a hyperparameter. In a Bayesian setting all parameters are assumed to come from a probability distribution, so in the second stage, both fixed and and random effects are assigned prior distributions

$$\pi(\boldsymbol{\beta}|\boldsymbol{B}\sigma_\beta^2) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}\sigma_\beta^2),$$
$$\pi(\boldsymbol{\gamma}|\boldsymbol{R}\sigma_\gamma^2) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}\sigma_\gamma^2),$$

where $\boldsymbol{B}$ and $\boldsymbol{R}$ are known, non-singular matrices, and the parameters σ_β^2 and σ_γ^2 are hyperparameters. If all distributions in the latent stage are assumed to be Gaussian, the model is known as a latent Gaussian model, a class of models assumed in several of the papers in this thesis. We return to latent Gaussian models in Section 1.2.

The third stage consists of assigning prior distributions to the hyperparameters, the parameters that control the distributions of the latent field and the likelihood. The choice of prior distributions for these parameters is important, since the priors allow the user to include knowledge about the parameters. Blangiardo and Cameletti (2015) highlight in particular two aspects which need to be taken into account when assigning prior distributions to the hyperparameters. The first is the type of distribution, which should be representative of the nature of the hyperparameters. The second is the choice of parameters of the prior distribution, which make the distribution more or less informative, and provide the level of knowledge about the hyperparameters.

The prior distributions can range from informative, expert priors to objective priors. When a statistician has access to detailed prior knowledge on the parameters of interest, this information can be used to construct informative priors (Ayyub, 2001; O'Hagan et al., 2006; Albert et al., 2012). The detailed prior knowledge can for example be distributions from

previous observations of similar phenomenons, results from previous experiments, or from consulting with experts in the field. An example of informative priors used in this thesis are the penalized complexity priors introduced by Simpson et al. (2017). The penalized complexity priors have only a single parameter that the user must choose, and this parameter controls the flexibility allowed in the model, causing the prior to be vague, weakly informative, or strongly informative.

Furthest from the informative priors are non-informative priors, also known as objective priors (Bernardo, 1979; Berger et al., 2009). Statisticians assigning objective priors to the hyperparameters aim to add as little subjective information as possible into the model. An example of an objective prior is the Jeffreys prior (Jeffreys, 1946).

Instead of assuming a prior that is non-informative along the whole support of the parameter, it can be enough to assure ignorance only on a subset of the parameters where the likelihood is far from zero. This strategy leads to a vague prior (Blangiardo and Cameletti, 2015). For instance, a vague prior distribution for a regression parameter is the $\mathcal{N}(0, 10^6)$ distribution. The prior is vague because it is nearly flat, but would nevertheless favor values closer to zero than further away from zero. In the Bayesian models in this thesis, the parameters of fixed effects, corresponding to the elements of $\boldsymbol{\beta}$ in this section, are assigned the $\mathcal{N}(0, 10^3)$ distribution as prior distribution. Although a large part of this thesis focuses on including prior knowledge into models, these parameters are not prioritized for the use of more informative priors, as there is usually sufficient information in the data to estimate the fixed effects.

A class of popular prior distributions are the conjugate priors. These prior distributions are limited in their flexibility, but yield models that are easy to treat analytically. The models are easy to treat because the conjugate prior distributions ensure that the posterior distribution belongs to the same family as the prior distribution. Since the functional form of the posterior distribution is known when applying conjugate priors, it is easy to extract summary statistics or analytically derive any other quantities of interest (Blangiardo and Cameletti, 2015).

Although the possibility of including knowledge beyond the observed data is appealing, specifying prior distributions can be challenging. One parameter that is particularly challenging is the random effects variance parameter σ_γ^2 (Gelman et al., 2006), as this parameter does not have any

simple family of conjugate prior distributions. The inverse-gamma(a, b) family is conditionally conjugate, meaning that if σ_γ^2 has a inverse-gamma prior distribution, then the posterior conditional distribution $\pi(\sigma_\gamma^2|\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{y})$ is inverse-gamma (Gelman et al., 2006). This prior is an attempt for a non-informative prior within the conditionally conjugate family with the a, b set to low values. However, in the limit where a, b become close to zero, the prior distribution yields an improper posterior density (Gelman et al., 2006). Therefore a, b must be set to reasonable values. For data sets in which low values of variance parameters are possible, inferences become sensitive to the inverse-gamma parameter choice a, b, and the prior distribution is no longer non-informative.

1.1.2 Frequentist setting

The focus now shifts to a frequentist setting, where only the data and the model are used in estimation and to make predictions. However, the choice of model can allow inclusion of knowledge about the process of interest, which we use in one of the papers of this thesis. We first repeat the LMM specified as a likelihood

$$\pi(\boldsymbol{y}|\boldsymbol{\eta}, \sigma^2) \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{I}\sigma^2). \tag{1.1}$$

If heterogeneity is expected to be present in the random effects parts of the LMM, this can be modeled with a linear predictor in the random effect variance. The second stage of the LMM then becomes

$$\pi(\boldsymbol{\gamma}|\boldsymbol{\beta}_d) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \tag{1.2}$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix with variances estimated using a linear predictor with a log link function $\log(\Lambda_{jj}) = \boldsymbol{w}_{d,j}^T \boldsymbol{\beta}_d$. Here, $\boldsymbol{\beta}_d$ is a fixed effect and the design vector $\boldsymbol{w}_{d,j}$ could contain prior information about the size of the different random effects. This prior information could be included to control the size of the random effects variance, and by this the relative importance of the random effects.

In this thesis we used model (1.2) to include information about the importance of different genomic markers (DNA sequences) in the genome. We explain the approach with an example. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)$ contain the random effect of genomic markers labeled $(1, \ldots, q)$. Assume that markers $(1, \ldots, r)$ with $r < q$ are located in a region of the genome previously shown to explain more variation than other regions, and the markers

$(r+1,\ldots,q)$ are not located in such regions. This information is in our case the prior knowledge, and is used to specify the design vector $\boldsymbol{w}_d$. We specify weight 1 for positions $(1,\ldots,r)$, i.e. $\boldsymbol{w}^T_{d,(1,\ldots,r)} = (1,\ldots,1)$, and weight 0 for positions $(r+1,\ldots,q)$, i.e. $\boldsymbol{w}^T_{d,(r+1,\ldots,q)} = (0,\ldots,0)$.

1.2 Estimation

There are two main approaches to statistical inference; Bayesian and frequentist. Bayesian inference is based on estimating the posterior distribution of the parameters in the latent stage and the hyperparameters. It is most commonly performed using Markov chain Monte Carlo (MCMC) methods (Givens and Hoeting, 2005; Gamerman and Lopes, 2006), in which samples are generated from posterior distributions by constructing a Markov chain with the target posterior as the stationary distribution. For models belonging to the class of latent Gaussian models, Bayesian inference can be performed using the integrated nested Laplace approximations method (Rue et al., 2009). This method approximates the posterior distributions without using sampling-based methods, and is much faster than MCMC methods (Rue and Martino, 2007).

Statistical inference for LMMs in the frequentist setting is usually performed using likelihood-based methods. The most common method is restricted maximum likelihood estimation (REML) (Fahrmeir et al., 2013). Other methods for inference are penalized likelihood or empirical Bayes (Fahrmeir et al., 2013), template model builder (Kristensen et al., 2015), the expectation maximization algorithm (Givens and Hoeting, 2005), and hierarchical likelihood (Lee and Nelder, 1996).

In this thesis, a Bayesian approach is taken in three of the papers, and a frequentist approach is taken in one of the papers. For Bayesian inference we use the integrated nested Laplace approximations, and for inference in the frequentist framework we use hierarchical likelihood. Introductions to both methods are given below.

1.2.1 Bayesian inference of latent Gaussian models using integrated nested Laplace approximations

This section gives a brief presentation of how the integrated nested Laplace approximations (INLA; Rue et al., 2009) are used to approximate the posterior marginal distributions for a model in the latent Gaussian model (LGM) class. In-depth descriptions of the INLA methodology can be found in Rue et al. (2009), Martins et al. (2013), Blangiardo and Cameletti (2015) and Rue et al. (2017). We first present the class of LGMs, the class of Bayesian models susceptible to INLA-based inference.

The class of LGMs includes many models, for example generalized linear (mixed) models, generalized additive (mixed) models, and spline smoothing methods. LGMs are hierarchical models, where observations $\boldsymbol{y}$ are assumed to be conditionally independent given a latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta}_1$, i.e. $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) \sim \Pi_{i\in\mathcal{I}}\pi(y_i|x_i, \boldsymbol{\theta}_1)$. The latent field $\boldsymbol{x}$ includes both fixed and random effects and is assumed to be Gaussian distributed given parameters $\boldsymbol{\theta}_2$, i.e. $\pi(\boldsymbol{x}|\boldsymbol{\theta}_2) \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{\Sigma}(\boldsymbol{\theta}_2))$. Finally, the hyperparameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, are assigned prior distributions $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.

For the computations in INLA to be both fast and accurate, the LGM has to satisfy some assumptions. Since INLA integrate over the hyperparameter space, the number of non-Gaussian hyperparameters should be low, typically less than 10, and not exceeding 20. Further, the latent field should not only be Gaussian, it should be a Gaussian Markov random field. The conditional independence property of a Gaussian Markov random field yields sparse precision matrices which makes computations in INLA fast due to efficient algorithms for sparse matrices. Lastly, each observation y_i should depend on the latent field through only one component x_i. Gaussian random fields and Gaussian Markov random fields will be covered in Section 1.4.

The main aim of Bayesian inference is to estimate the marginal posterior distribution of the variables in the model. The marginal posteriors are given as

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \tag{1.3}$$

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})\pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{1.4}$$

INLA do this by breaking down the problem into three sub-problems

1. Approximate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$
2. Approximate $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ for i of interest
3. Compute $\pi(\theta_j|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{y})$ using the results from sub-problem 1 and 2, and numerical integration

A brief summary of the steps is given below. For details we refer to Rue et al. (2009) and Martins et al. (2013).

The marginal posterior distribution for $\boldsymbol{\theta}$ in sub-problem 1 is approximated starting from the identity

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \tag{1.5}$$

The numerator in (1.5) is easy to compute, but the denominator is in general not available in closed form and must be approximated. However, when the model has a Gaussian likelihood, the full conditional $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ and its marginals $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ are also Gaussian. This implies that for each value of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ can be computed exactly, without the need of approximations. In this thesis we assume a Gaussian likelihood in all our models, meaning that we have no approximations in sub-problem 1.

For the approximation in sub-problem 2, there are different options for non-Gaussian likelihoods, but with a Gaussian likelihood, this step simplifies as sub-problem 1. Rather than approximating $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, INLA compute these Gaussian marginals exactly. Since a Gaussian likelihood ensures exact distributions of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$, the only source of error from INLA is then from the numerical integration.

Finally, in sub-problem 3, INLA use numerical integration to solve the integrals in (1.3) and (1.4), with respect to $\boldsymbol{\theta}$. INLA have three different options for exploring the $\boldsymbol{\theta}$ space. The first is using a grid search around the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This option is the default option when the dimension of $\boldsymbol{\theta}$ is 1 or 2, and is also the most accurate. An illustration of the grid search approach for a $\boldsymbol{\theta}$ of dimension two is shown in Figure 1.1. The mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ is located and the principal component directions are explored by grid search to locate the majority of the probability mass. The second option is to use the central composite design (Box and Wilson, 1951), which cleverly locates fewer points around the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. This option

is the default strategy for dimensions of $\boldsymbol{\theta}$ larger than two. The last option is to ignore the variability around the hyperparameters and to use only the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

Putting it all together, the INLA computation scheme for a model with a Gaussian likelihood is (Martino and Riebler, 2019)

1. Explore the $\boldsymbol{\theta}$ space through $\pi(\boldsymbol{\theta}|\boldsymbol{y})$. Find the mode of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and a series of points $\{\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^K\}$ in the area with high density of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.
2. Compute $\pi(\boldsymbol{\theta}^1|\boldsymbol{y}), \ldots, \pi(\boldsymbol{\theta}^K|\boldsymbol{y})$ for the K chosen support points using (1.5).
3. Compute $\pi(x_i|\boldsymbol{\theta}^1, \boldsymbol{y}), \ldots, \pi(x_i|\boldsymbol{\theta}^K, \boldsymbol{y})$ for the K chosen support points.
4. Solve (1.3) and (1.4) via numerical integration. The integral in (1.4) is solved as

$$\pi(x_i|\boldsymbol{y}) = \sum_{k=1}^{K} \pi(x_i|\boldsymbol{\theta}^k, \boldsymbol{y})\pi(\boldsymbol{\theta}^k|\boldsymbol{y})\Delta_k, \tag{1.6}$$

   where Δ_k are appropriate weights depending on the chosen support points. The integral in (1.3) is solved similarly.

1.2.2 Estimation using hierarchical likelihood

Inference is also performed using hierarchical likelihood (h-likelihood; Lee and Nelder, 1996). The method is suited for hierarchical generalized linear models, and allows fixed and random effects in a linear predictor for the variance parameters. The fitting algorithm is implemented in Rönnegård et al. (2010b), and an overview of estimation with GLMMs via h-likelihood is described in Lee et al. (2018). The rest of this section gives a brief outline of h-likelihood theory following Rönnegård et al. (2010b) with more detail given in Lee and Nelder (1996), Lee et al. (2018), and Alam et al. (2015).

The conditional log-likelihood for $\boldsymbol{y}$ given $\boldsymbol{\gamma}$ for the model in (1.1) has the so-called canonical form

$$l(\boldsymbol{y}|\boldsymbol{\gamma}; \boldsymbol{\theta}^*, \phi) = \frac{\boldsymbol{y}\boldsymbol{\theta}^* - b(\boldsymbol{\theta}^*)}{\phi} + c(\boldsymbol{y}, \phi),$$

where $\boldsymbol{\theta}^*$ is the parameter of interest and $\phi = \sigma^2$. In the case of a Gaussian model $\boldsymbol{\theta}^* = \boldsymbol{\mu}^*$, where $\boldsymbol{\mu}^*$ is the conditional mean of $\boldsymbol{y}$ given $\boldsymbol{\gamma}$, and $\boldsymbol{\mu}^* =$

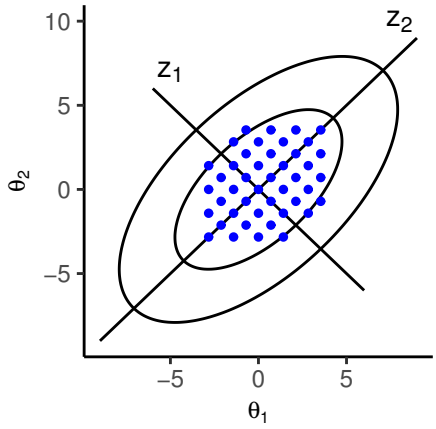


Figure 1.1: Illustration of the exploration of the posterior marginal for $\boldsymbol{\theta}$. The mode is located and the principal component directions z_1 and z_2 are explored by a grid search (blue points) to locate the majority of the probability mass

$\boldsymbol{W\beta} + \boldsymbol{U\gamma}$. In the canonical form, the function $b(\boldsymbol{\theta}^*)$ satisfies $\mathrm{E}(\boldsymbol{y}|\boldsymbol{\gamma}) = b'(\boldsymbol{\theta}^*)$ and $\mathrm{Var}(\boldsymbol{y}|\boldsymbol{\gamma}) = b''(\boldsymbol{\theta}^*)\phi$, and $c(\boldsymbol{y}, \boldsymbol{\phi})$ is a known function. The h-likelihood is then defined as

$$h = l(\boldsymbol{y}|\boldsymbol{\gamma}; \boldsymbol{\theta}^*, \phi) + l(\boldsymbol{v}|\alpha),$$

where $\boldsymbol{v} = v(\boldsymbol{\gamma})$ is a strict monotonic link function specified such that the random effects occur linearly in the linear predictor $\boldsymbol{\eta}^* = \boldsymbol{\mu}^*$, and $l(\boldsymbol{v}|\alpha)$ is the log density for $\boldsymbol{v}$ with parameter α. The adjusted profile h-likelihood is

$$h_p = \left(h + \frac{1}{2} \log |2\pi H^{-1}| \right) |_{\beta=\hat{\beta}, v=\hat{v}},$$

where H is the Hessian matrix of the h-likelihood h.

In order to estimate the model parameters, Lee and Nelder (1996) suggested a two-step procedure. First, the h-likelihood h is maximized with respect to the $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for given variance parameters. Next the adjusted profile h-likelihood h_p is maximized to estimate the dispersion parameters ϕ and α for given $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$. Iteration between the two steps continue until convergence.

The fitting algorithm implemented in Rönnegård et al. (2010b) for estimating parameters in the model from (1.1) and (1.2) using the h-likelihood (Lee and Nelder, 1996) is described below.

---

1. Start by setting $\boldsymbol{\Lambda} = \boldsymbol{I}\sigma^2_\gamma$, and set starting values for σ^2_γ and σ^2.

   Iterate from 2. to 4. until convergence:

2. Estimate $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ from the following augmented linear model

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{W} & \boldsymbol{U} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon} \\ -\boldsymbol{\gamma} \end{pmatrix} \tag{1.7}$$

   using weighted least squares with weight matrix

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{I}\frac{1}{\sigma^2_\varepsilon} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Lambda}^{-1} \end{pmatrix}$$

3. To estimate the elements in $\boldsymbol{\Lambda}$ and σ^2, compute the hat matrix $\boldsymbol{H}_{(n+q)\times(n+q)}$ for the model in (1.7). Then

   a. Fit a gamma distributed generalized linear model with response $\hat{\gamma}_i/(1-H_{jj})$, log link, linear predictor $\boldsymbol{w}_d^T\boldsymbol{\beta}_d$, and weights $(1-H_{jj})/2$. Here H_{jj} is jth the diagonal element in the hat matrix $\boldsymbol{H}$ with $j = n+i$.

   b. Fit a gamma distributed generalized linear model with response $\hat{\varepsilon}_i/(1-H_{ii})$, identity link, linear predictor λ, and weights $(1-H_{ii})/2$.

4. From 3. we have $\hat{\Lambda}_{ii} = \exp(\boldsymbol{w}^T_{d,i}\hat{\boldsymbol{\beta}}_d)$, and $\hat{\sigma}^2 = \hat{\lambda}$.

Weighted least squares is an extension of ordinary least squares regression, where each observation is weighted according to some criterion (Fahrmeir et al., 2013). The model in (1.7) can be rewritten as

$$\boldsymbol{y}_a = \boldsymbol{T}_a\boldsymbol{\delta} + \boldsymbol{e} \tag{1.8}$$

The weighted least squares solutions for the model (1.8) is then

$$\hat{\boldsymbol{\delta}} = (\boldsymbol{T}_a^T \boldsymbol{V}^{-1} \boldsymbol{T}_a)^{-1} \boldsymbol{T}_a^T \boldsymbol{V}^{-1} \boldsymbol{y}_a$$

The use of h-likelihood is generally not accepted by all statisticians, with the main criticism for the h-likelihood being non-invariance of inference with respect to transformation (Rönnegård et al., 2010a). Non-invariance here means that the h-likelihoods of two equivalent models are not equivalent. However, a restriction that random effects occur linearly in the linear predictor is implied in the h-likelihood, and assures invariance (Lee et al., 2007, 2018).

1.3 Model selection

In this section we discuss different approaches for performing model selection, the task of selecting a statistical model from a set of candidate models, given data (Claeskens et al., 2008; Leeb and Pötscher, 2009; Claeskens, 2016). We distinguish between two different motives for statistical modeling and selecting models; to predict or increase understanding, which determine the approach for evaluating the models. The two approaches are often connected, but the predictive versus explanatory distinction has an impact on the statistical modeling (Shmueli et al., 2010).

When the goal of statistical modeling is to make predictions, we must determine whether a model, built on a training set, can be used to make predictions to support decision making. When the goal is to increase the understanding of a process of interest, we must determine whether a statistical model is a good representation of the truth. We discuss both approaches in this section, with most focus on making predictions, since this is the focus of the enclosed papers in this thesis.

As mentioned in the beginning of this introduction, the papers in this thesis aim to make contributions towards improving predictions of certain genetic effects. These effects are not observable from data, but are components of the response variables in our models. Because the effects we aim to predict are not observable, simulation studies have been important to evaluate predictions from our computer models. Simulation studies are computer experiments often used by statisticians to understand the behavior of the statistical models and methods, where the parameters of interest are known from the process generating the data (Ripley, 1987;

Gentle, 2006). What we can call the computer models, are the proposed statistical models for making predictions, and simulation models are the models used to generate data where the parameters of interest are known. A part of this section is dedicated to simulation of data from simulation models.

1.3.1 Making predictions

The assessment of goodness-of-fit of statistical models through their predictive performance is often done by evaluating the accuracy of point predictions. This can be done using the root mean-square error (RMSE) (Claeskens, 2016), which is simply the square root of the mean of the squared difference between the mean (posterior) estimate and the true/observed value. The predictive ability of a model for a set of observations is then typically calculated using an average over each point prediction, and we choose the model with the lowest average RMSE.

The RMSE does not assess the prediction uncertainty. To do this, the whole predictive distribution of the model must be evaluated, which is known as probabilistic forecasting (Gneiting and Raftery, 2007). The continuous ranked probability score (CRPS; Gneiting and Raftery, 2007) measures a combination of bias and sharpness of the posterior distribution, by taking into account the whole predictive distribution, and is a popular score function for this kind of forecasting

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(u) - 1\{y \leq u\})^2 \, du,$$

where F is the predictive cumulative distribution and y is the true/observed value (Gneiting and Raftery, 2007). In this thesis we use the CRPS for model evaluation in most papers, where the predictive cumulative distribution F is approximated with a Gaussian distribution with mean and variance from the posterior distribution of the estimates. The two scores mentioned so far (RMSE, CRPS) are negatively oriented, meaning that lower values of the scores, indicate a better predictive ability.

Correlation is an important measure of prediction accuracy in plant and animal breeding (Bourdon and Bourbon, 2000; Lado et al., 2013; Ferrão et al., 2017; Elias et al., 2018). In this thesis the sample correlation coefficient, also known as the Pearson product-moment correlation coefficient (Walpole et al., 2012), is used to estimate the linear association

between two variables. The correlation works with point estimates and ignores uncertainty of the estimates, and a high value of this metric is desired. The reason why correlation is popular to assess model performance in plant and animal breeding is explained in Section 1.6.

In this thesis we make predictions that are both in-sample and out-of-sample. The in-sample predictions are based on data used in the model construction, whereas the out-of-sample predictions are based on data not used in the model construction. For out-of-sample predictions, the data are separated in two data sets; the training data and the testing data (Friedman et al., 2001). Usually, a large part of the data is used in the parameter estimation, known as training, and a smaller subset of the available data is reserved for evaluation of the predictions, known as testing.

A popular approach for performing training and testing systematically is cross-validation (Friedman et al., 2001; Konishi and Kitagawa, 2008; James et al., 2013). In cross validation the data is first partitioned into r subsets of similar size. Next, training and testing is performed in several rounds starting with the first subset as testing set and the combined remaining $r - 1$ subsets as training set. In the second round, the second subset is used as testing set, and the combined remaining $r - 1$ subsets as training set. This continues until all subsets have been used as testing sets. Based on the estimates from the training set, we obtain predictions for the testing set, and can combine (e.g. average) the predictive performance over all rounds to get a total estimate of the model's predictive performance.

1.3.2 Increase understanding of a phenomenon

In many scientific fields, for example the social sciences, statistical methods are used to increase the understanding of a process of interest, to discover causal effects of variables, and models are built to be good representations of the truth (Shmueli et al., 2010). When this is the case, model selection is focused towards variable selection, the process of selecting a subset of relevant variables for use in model construction (Dunson, 2008; Heinze et al., 2018). This is done by testing whether the model assumptions fit with the data, by using various model choice criteria to select the most promising model among candidate models (Claeskens et al., 2008), and not by evaluating predictions.

There are several popular diagnostics and approaches, and we mention some of these here. For example are analysis of residuals and the coefficient of determination R^2 play important roles in evaluation of regression models (Walpole et al., 2012), and the Akaike information criterion is one of the most widely used criteria for model choice within likelihood-based inference (Fahrmeir et al., 2013). The deviance information criterion (DIC; Spiegelhalter et al., 2002) is widely used to compare model fit between different hierarchical Bayesian models while also assessing the model complexity. It is defined as

$$\text{DIC} = \overline{D(\theta)} + p_D,$$

where $\overline{D(\theta)}$ is the posterior mean deviance and p_D is the effective number of parameters in the model. The "spike-and-slab" regression (Mitchell and Beauchamp, 1988) is a Bayesian variable selection technique that is useful when the number of possible variables is larger than the number of observations, and the reversible jump MCMC (Green, 1995) can be used to move between models with different numbers of variables.

1.3.3 Simulation as experimentation

For the applications presented in this thesis, the goal is to predict certain random effects in the models. When modeling real data, the true random effects are not known, and it is therefore not possible to evaluate the predictions from this real data. Simulation studies are therefore performed in all papers of the thesis. In these computer experiments, data are created by pseudorandom sampling (Ripley, 1987; Gentle, 2006). This way it is possible to understand the behavior of the statistical models and methods because the parameters of interest are known from the process generating the data. With a large number of realizations (sets of artificial data) for each set of parameters, it is possible to experiment with different models, and make model comparisons.

The statistical model used for simulation of data can be either *convenient* or *mechanistic* (Ripley, 1987). By this we mean that the simulation model is either a simplified version of the process of interest, or aimed to represent the actual mechanisms of the process, respectively. The convenient simulation model is often the same model as the computer model, whereas a mechanistic model is a more complex model than the computer

model. Although we use the term mechanistic, we do not mean that the mechanistic simulation model is not stochastic, merely that the simulation model is based on the mechanisms of the process of interest. Both convenient and mechanistic simulation models can be used to help understanding the behavior of the statistical computer models, to predict, or to aid decision making (Ripley, 1987). However, with data simulated from the convenient model the evaluation mostly informs whether the simulation program and the computer model do what they are intended to do. On the other hand, evaluation with data generated from the mechanistic model allows testing of the performance and behavior of a computer model with data more similar to observed data. Most of the data generated in this thesis is generated from mechanistic models, more realistic and complex processes than the models used in estimation.

Simulation of genomic data is commonly done using mechanistic models that rely on the underlying genomic processes discovered during the past century (Zhang et al., 2015b; Faux et al., 2016; Xu et al., 2013). It is for example known that only a few of many so-called genomic markers may have causal effect on the traits of interest, so simulation of the genetic effects is performed according to this using a mechanistic model. In the computer models however, the effects of genomic markers are usually assumed to come from distributions where all markers can be assigned an effect (Meuwissen et al., 2001; Muir, 2007), because the genetic processes are too complex to model. In this thesis we have used the AlphaSimR software (Gaynor et al., 2019) which allows stochastic simulations of breeding programs to the level of DNA sequence for every individual. We return to genomic markers and the computer models commonly used within plant and animal breeding in Section 1.6.

1.4 Gaussian random fields

Gaussian random fields (GRFs) play an important role in statistics, especially in spatial statistics, and in this thesis GRFs play an important role as model components. This section presents GRFs, a subclass of GRFs known as Gaussian Markov random fields, and the stochastic partial differential equation approach (Lindgren et al., 2011), which allows representing GRFs as GMRFs.

A GRF is a random function over an arbitrary domain involving Gaus-

sian probability density functions of the variables. Let $\{x(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a stochastic process where $D \in \mathbb{R}^d$. The process $\{x(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is then a GRF if for any $k \geq 1$ and any locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k \in D$, $(x(\boldsymbol{s}_1), \ldots, x(\boldsymbol{s}_k))$ is normally distributed (Rue and Held, 2005). The mean is $\boldsymbol{\mu}(\boldsymbol{s}) = E(x(\boldsymbol{s}))$, and the covariance function is $C(\boldsymbol{s}, \boldsymbol{t}) = Cov(x(\boldsymbol{s}), x(\boldsymbol{t}))$, and usually the dimension d of the domain D is 1,2 or 3.

GRFs can be included as random effects in LMMs to model the dependency in point referenced data, where the field is usually assumed to be stationary and isotropic. A GRF is stationary if $\boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{\mu}$ for all $\boldsymbol{s} \in D$, and $C(\boldsymbol{s}, \boldsymbol{t})$ only depends on $\boldsymbol{s} - \boldsymbol{t}$. If in addition the covariance function only depends on the Euclidean distance between $\boldsymbol{s}$ and $\boldsymbol{t}$, the field is isotropic. In this thesis, the GRFs are assumed to be stationary and isotropic.

For the GRF to be a valid probability model, the covariance function must be positive definite (Rue and Held, 2005). To ensure positive definiteness, it is common to use established positive definite covariance functions such as exponential, Gaussian, powered exponential, and Matérn covariance functions (Rue and Held, 2005). Among the most popular is the Matérn covariance function (Matérn, 1960; Guttorp and Gneiting, 2006)

$$C(\boldsymbol{s}, \boldsymbol{t}) = \frac{\sigma_s^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa \|\boldsymbol{t} - \boldsymbol{s}\|\right)^{\nu} K_{\nu} \left(\kappa \|\boldsymbol{t} - \boldsymbol{s}\|\right), \tag{1.9}$$

where K_ν is the modified Bessel function of the second kind and order $\nu > 0$, $\|\cdot\|$ denotes the Euclidean distance in $\mathbb{R}^d$, and σ_s^2 is the marginal variance. For the scaling parameter $\kappa > 0$, an empirically established relation to the range parameter is $\kappa = \sqrt{8\nu}/\rho$, where the range parameter $\rho > 0$ describes the distance where the correlation between two points is near 0.1. The parameter ν determines the mean-square differentiability and the smoothness of the field. This value is generally difficult to identify from data, so it is usually fixed (Lindgren et al., 2011).

1.4.1 Gaussian Markov random fields

Inference with GRFs is computationally expensive because it requires factorization of dense precision matrices (Rue and Held, 2005). Gaussian Markov random fields (GMRFs) do not incur this penalty because the Markov property ensures sparse precision matrices. This excellent computational property makes GMRF modeling popular in a range of statistical

areas. Rue and Held (2005) mention several applications within structural time series analysis, analysis of longitudinal and survival data, graphical models, semiparametric regression and splines, image analysis, and spatial statistics. Here, we introduce GMRFs following Rue and Held (2005), and in Section 1.4.2 we show how GMRFs can be used to represent GRFs to make computations with GRFs efficient.

Before introducing GMRFs further, two concepts need to be introduced; conditional independence and undirected graphs. For three random variables X, Y and Z, we denote the conditional independence between X and Y given Z as

$$X \perp Y \mid Z,$$

meaning that conditional on Z, Y and X are independent. The same can be expressed as $\pi(x, y|z) = \pi(x|z)\pi(y|z)$.

The conditional independence structure of GMRFs can be represented using undirected graphs. Let the undirected graph $\mathcal{G}$ be a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes in the graph and $\mathcal{E}$ contains the edges $\{i, j\}, i, j \in \mathcal{V}$ and $i \neq j$. If and only if $\{i, j\} \in \mathcal{E}$, there is an undirected edge between nodes i and j. If $\mathcal{V} = \{1, \ldots, n\}$, the graph is labeled. An example of an undirected graph is shown in Figure 1.2.

The random vector $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a GMRF with respect to the labeled graph $\mathcal{G}$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$, if and only if the density of $\boldsymbol{x}$ has the form

$$\pi(\boldsymbol{x}) = (2\pi)^{n/2}|\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{1.10}$$

and $Q_{ij} \neq 0$ if and only if $\{i, j\} \in \mathcal{E}$ for all $i \neq j$ (Rue and Held, 2005).

A GMRF is usually parameterized with the precision matrix because it is sparse. The precision matrix also has the property that it gives information about the conditional independence of the GMRF. An element in the precision matrix, Q_{ij}, is zero if and only if x_i and x_j are conditionally independent given all other nodes $\boldsymbol{x}_{-ij}$. This means that the non-zero pattern in $\boldsymbol{Q}$ determines the graph $\mathcal{G}$, and that for a given graph $\mathcal{G}$, the non-zero terms in $\boldsymbol{Q}$ can be determined.

The Markov properties of a GMRF are related to the conditional independence of the corresponding graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We now present three

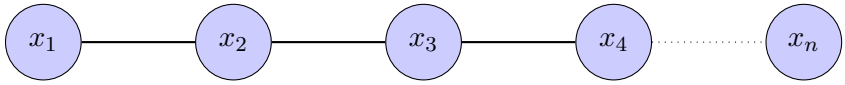


Figure 1.2: A linear undirected graph

Markov properties that are all equivalent for a GMRF. These are the pairwise Markov property

$$x_i \perp x_j \mid \boldsymbol{x}_{-ij}, \quad \text{if } \{i,j\} \notin \mathcal{E} \text{ and } i \neq j,$$

the local Markov property

$$x_i \perp \boldsymbol{x}_{-\{i,ne(i)\}} \mid \boldsymbol{x}_{ne(i)}, \quad \text{for every } i \in \mathcal{V},$$

and the global Markov property

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B \mid \boldsymbol{x}_C, \quad \text{for all disjoint sets A, B and C, where C separates A and B, and A and B are non-empty.}$$

Here $ne(i)$ refers to the neighbors of x_i, the nodes with a direct edge to x_i.

The computational advantages of GMRFs compared to GRFs comes from the sparsity of the precision matrix, and the use of sparse matrix algorithms. In general, the cost of factorizing a dense $n \times n$ (covariance) matrix is $O(n^3)$. The cost of factorizing the precision matrix of a GMRF depends on the GMRF itself, but typical costs are $O(n)$ for one-dimensional GMRFs, $O(n^{3/2})$ for two-dimensional GMRFs, and $O(n^2)$ for three-dimensional GMRFs. Details and algorithms for computations with GMRFs are given in Rue and Held (2005).

The auto-regressive process of order 1

The auto-regressive process of order 1 is a simple example of a GMRF on a linear graph (Rue and Held, 2005). The linear graph in Figure 1.2 corresponds to an auto-regressive process of order 1, and the model can be expressed as

$$\begin{aligned} x_i &= \rho x_{i-1} + \varepsilon_i \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2(1-\rho^2)), \quad |\rho| < 1 \end{aligned} \tag{1.11}$$

The model specifies that the output variable depends linearly on the previous values and a stochastic term. By representing (1.11) in the conditional form, the assumption about conditional independence can be seen directly

$$x_i \mid x_{i-1} \sim \mathcal{N}(\rho x_{i-1}, \sigma^2(1-\rho^2)),$$
$$x_1 \sim \mathcal{N}(0, \sigma^2), \quad |\rho| < 1,$$

for $i = 2, \ldots, n$. In this form of the model, the assumption about conditional independence is easier to see; x_i and x_j with $1 \leq i < j \leq n$ and $j - i > 1$, are conditionally independent given $\{x_{i+1}, \ldots, x_{j-1}\}$.

The joint density of $\boldsymbol{x}$ has the form given in (1.10) with a tridiagonal precision matrix given by

$$\boldsymbol{Q} = \frac{1}{\sigma^2(1-\rho^2)} \begin{pmatrix} 1 & -\rho & & & \\ -\rho & 1+\rho^2 & -\rho & & \\ & \ddots & \ddots & \ddots & \\ & & -\rho & 1+\rho^2 & -\rho \\ & & & -\rho & 1 \end{pmatrix}$$

The tridiagonal form of the matrix is due to the conditional indepence of x_i and x_j given the rest for $|i - j| > 1$.

1.4.2 The stochastic partial differential equation approach

In this section we present the stochastic partial differential equation (SPDE) approach introduced by Lindgren et al. (2011). The approach is a combination of results within stochastic process theory (Whittle, 1954, 1963) and numerical methods for solving partial differential equations. Lindgren et al. (2011) showed how to provide an explicit link between some GRFs in the Matérn class and GMRFs, using an approximate stochastic weak solution to (linear) stochastic partial differential equations. Their approach enables continuous modeling with GRFs by a GMRF representation, leading to fast computations.

The approach is based on the equation

$$(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{1.12}$$

where $\Delta = \sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2}$, $\mathcal{W}$ is Gaussian white noise, κ controls the range and α controls the smoothness. The stationary solutions of (1.12), are

GRFs with Matérn covariance function (Whittle, 1954, 1963). The Matérn covariance function was presented in (1.9), and its parameters are coupled with the SPDE in (1.12) with $\alpha = \nu + d/2$, and

$$\sigma_s^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}},$$

where ν is the smoothness parameter and σ_s^2 is the marginal variance of the Mátern covariance function.

The continuous solution $x(\boldsymbol{s})$ is approximated by the finite element method, an approximate numerical method to solve partial differential equations. The domain of interest is discretized into non-overlapping triangles, and an approximation to the GRF is built on a basis function representation on the discretized domain. The full description of how GRFs are represented using GMRFs via the SPDE approach can be found in Lindgren et al. (2011).

1.5 Spatial statistics

Spatial statistics is a sub-field of statistics that uses spatially referenced data (Gelfand et al., 2010). The field is a cornerstone in petroleum and hydrology, and is also used for applications within many other fields, such as agriculture, ecology and epidemiology. By learning from spatial data, it becomes possible to make predictions for locations where no data have been observed, and to understand the underlying spatial processes generating the data.

Gelfand et al. (2010) divide spatial statistics into three fields; continuous spatial variation, discrete spatial variation, and spatial point patterns. Models for continuous spatial variation assumes a continuous process in space, with observations at a discrete set of locations − known as point-referenced data. Discrete spatial variation deals with lattice data, pixel data and areal unit data, while in models for spatial point patterns the spatial locations are considered as random events. In this thesis we cover models for both continuous and discrete spatial variation.

The branch of statistics covering models for continuous spatial variation is often referred to as geostatistics. The models in geostatistics grew out from the work of Krige (1951), known for the kriging technique, and Matérn (1960), with the Matérn covariance function. An important model

for continuous spatial variation is the Gaussian process, where the random vector of the of the spatial locations is assumed to be a GRF. In this thesis we use the SPDE approach to represent GRFs as GMRFs to do efficient modeling of continuous spatial variation.

Statistical modeling of discrete spatial variation is often done using a Markov random field (Guyon, 1995), where the Gaussian case is the previously introduced GMRF (Rue and Held, 2005). Models that have been heavily used to model discrete spatial variation are the conditional and intrinsic auto-regressions (Besag, 1974). Among those, the most commonly used are Gaussian conditional auto-regressions (CAR) which are GMRFs, and intrinsic auto-regressions which are conditional auto-regressions with singular precision matrices (Gelfand et al., 2010). An example of a Gaussian conditional auto-regression on the line is the auto-regressive process or order 1, and the corresponding graph in Figure 1.2.

1.6 Breeding and quantitative genetics

Quantitative genetics is the study of quantitative traits, and is a cornerstone in both evolutionary biology and breeding (Lynch and Walsh, 1998; Sorensen and Gianola, 2002). In quantitative genetics it is of interest to study measurable traits, predict certain random effects known as breeding values, and identify regions of DNA associated with a particular trait. In plant and animal breeding, estimated breeding values are used to select individuals for future breeding, in order to increase the population mean for some traits of interest (Bourdon and Bourbon, 2000; Acquaah, 2009; Isik et al., 2017).

The applications in this thesis have all been within plant and animal breeding, with the aim to contribute to improve estimation and prediction of different genetic effects. This section starts with a summary of some basic concepts in quantitative genetics, and continues with more introduction to plant and animal breeding, where breeding trials and the most common statistical models within breeding are presented.

1.6.1 Some important concepts in genetics

The DNA (deoxyribonucleid acid) contains the genetic information of a living organism. It is composed of two chains of nucleotides, which again

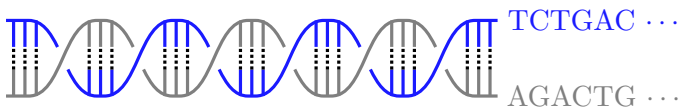


Figure 1.3: A simple illustration of the DNA with the nucleobases

are composed of one of four nitrogen-containing nucleobases (cytosine, guanine, adenine or thymine) (Ziegler et al., 2010). A specific segment of the DNA is known as a locus, and the base variant in a specific locus is known as an allele. A group of alleles in an organism that were inherited together from a single parent is known as a haplotype, and the combination or pair of alleles in an organism at a particular locus is known as a genotype. Figure 1.3 shows a simple illustration of the DNA double helix and some nucleobases. In this figure, the sequence TCTGAC is one haplotype, and AGACTG is another haplotype, both consisting of six alleles. The combination T/A is an example of a genotype.

A quantitative trait locus (QTL; Lynch and Walsh, 1998) is a locus or a region of several loci in the genome that contains genes affecting quantitative traits. By locating QTL in the genome, work can be done to determine what the genes in the QTL code for. The concept of QTL is particularly relevant for one of the papers in this thesis, where we include information in the models about potential QTL in different regions of the DNA.

A genetic marker (or genomic marker) is a gene or DNA sequence with known location in a chromosome that can be used to identify individuals. An example of a genomic marker is a single-nucleotide polymorphism (SNP). A SNP is a type of genetic variation, defined as a locus where the type of nucleotide present can differ between individuals (Ziegler et al., 2010). For example, in a specific locus, the cytosine (C) nucleotide may appear in most individuals, but in a minority of individuals, the position is occupied by the adenine (A) nucleotide. This means that there is a SNP at this specific locus, with the two possible nucleotide variations, C or A.

Linkage disequilibrium is a non-random relationship between alleles at two or more loci (Lynch and Walsh, 1998). Under linkage disequilibrium, haplotypes do not occur at the frequencies expected when the alleles are independent. Positive linkage disequilibrium exists when two alleles occur together on the same haplotype more often than expected, and negative

Table 1.1: Example of 5 haplotypes spanning 7 mutations from Kelleher et al. (2019). The original alleles are coded as 0 and mutated alleles are coded as 1

		Locus						
		1	2	3	4	5	6	7
Haplotype	a	1	0	0	1	1	0	0
	b	1	0	0	0	1	1	0
	c	1	0	0	0	1	1	0
	d	0	1	0	0	0	0	1
	e	0	1	1	0	0	0	1

linkage disequilibrium exists when alleles occur together on the same haplotype less often than expected. Because of this non-random relationship between alleles, non-coding markers in the genome can be correlated with QTL, and in this way "capture" effects affecting traits.

A phylogeny, or a phylogenetic tree, is a directed diagram showing the evolutionary relationships between different biological species or other entities (Morrison, 2016). The phylogeny is based on similarities and differences in physical or genetic properties. Haplotypes can for example be connected in a phylogeny, where the edge between two haplotypes indicate a mutation in an allele. In one of the papers in this thesis, we construct models for phylogeny. An example of 5 haplotypes spanning 7 loci from this paper is given in Table 1.1. The alleles are coded using 0 or 1, rather than the letters indicating the nucleobases. An example of a plausible phylogeny for the haplotypes is shown in Figure 1.4, where haplotypes are denoted as nodes with allele sequences. Relationships between haplotypes are denoted as edges, and mutated sites are denoted with a number on edges. For example, the haplotype i has allele sequence `0000000`, and the haplotype g with sequence `1000100` differs from the haplotype i due to mutations in loci 5 and 1.

In quantitative genetics, measurable traits of interest are known as phenotypes. These are assumed to consist of a genetic part and an environmental part, sometimes with an interaction term (Conner et al., 2004). When a phenotype is caused by several genes, the phenotype usually has continuous distribution in a population, and is often assumed to be Gaussian distributed. In this thesis, the response variable in our statistical

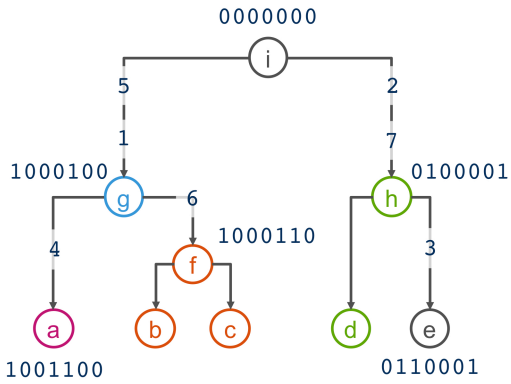


Figure 1.4: Phylogenetic tree for the haplotypes in Table 1.1

models are phenotypic observations, which are assumed to be Gaussian distributed. Further, we assume that there are no interaction effects between the genetic and environmental part of the phenotype.

The total phenotypic variance in a population is divided into components representing the genetic and environmental variation. Further, the genetic variation can be divided into additive and non-additive genetic variation, $\sigma_g^2 = \sigma_a^2 + \sigma_n^2$ (Conner et al., 2004). The effects behind the additive genetic variation, the additive effects, are commonly known as breeding values. These values are sums over allele substitution effects over the unobserved genotypes of causal loci. Non-additive effects are defined as the remaining genetic effects not captured by the additive values. Statistically, the non-additive effects capture variation due to allele interactions within and between loci (Conner et al., 2004).

1.6.2 Plant and animal breeding trials

The goal of both plant and animal breeding is to identify the individuals in a population with the highest genetic value for some traits of interest (Bourdon and Bourbon, 2000; Acquaah, 2009). By selecting these individuals for future breeding, the population mean for the traits of interest can be moved in the desired direction, for example towards higher grain yield for plants, or higher milk yield for cattle. To determine which individuals to select, a breeding program is performed through several phases:

(i) defining breeding goal, (ii) collecting phenotypes, genotypes and pedigrees, (iii) genetic evaluation to estimate and predict breeding values, (iv) selection of parents for next generation based on the estimated breeding values, (v) mating in an animal breeding program or re-planting genetic lines in a plant breeding program, and (vi) evaluation of the program with respect to the genetic diversity maintained and realized response to selection. Detailed descriptions of breeding programs within plant and animal breeding can be found in Acquaah (2009) and Bourdon and Bourbon (2000), respectively. In this thesis we have focused our contributions to phase (iii) estimation and prediction of breeding values. To be able to estimate and predict breeding values, methods for separating the genetic variation from environmental variation are required. This is done both through the design of the breeding program, and using statistical models applied to the gathered data, estimating genetic effects.

Estimates of genetic effects, and the additive and non-additive components have different applications in breeding (Acquaah, 2009). The estimates of additive effects are used to identify parents of the next generation, because additive values indicate the expected change in mean genetic value in the next generation. Estimates of genetic effects are used to identify individuals for commercial production, because genetic values indicate the expected phenotypic value. Estimates of genetic values are particularly valuable in plant breeding where individual genotypes can be effectively cloned, whereas additive effects are usually more of interest in animal breeding.

The development of experimental design in plant breeding was pioneered by R.A. Fisher. Thes plant breeding designs are based on replications, randomization and blocking (Fisher, 1926, 1935). The experiments are designed to control environmental error and to give reasonable confidence that the differences between genetic lines will be detected. Time and resources limit the experimental design, so environmental effects are included in the models for estimation of breeding values (Isik et al., 2017).

The most popular spatial model for environmental effect within agriculture is the separable auto-regressive model (Cullis and Gleeson, 1991; Gilmour et al., 1997). When different varieties are planted in a lattice consisting of rows and columns, this model can be used to capture the environmental effect of the field. The precision matrix for the model is constructed as the Kronecker product (Neudecker, 1969) between the pre-

cision matrix for an auto-regressive model along the rows, and the precision matrix for an auto-regressive model along the columns. We use this models for discrete spatial variation in one of the papers in this thesis. The response to selection determines how fast a breeder can advance the population towards higher population mean for some trait of interest through selection (Acquaah, 2009). The value of the response to selection is the difference between the mean phenotypic value of the offspring of the selected parents, and the whole of the parental generation before selection. Factors that influence this value are the total phenotypic variation in the population σ_p^2, the (narrow sense) heritability of the trait of interest σ_a^2/σ_p^2, and the proportion of the population that is selected for the next generation (Acquaah, 2009). Since a breeder wants to advance the population to higher trait means, the response to selection is important to control.

In Section 1.3 correlation was mentioned as an important score of predictive performance of statistical models in plant and animal breeding. More specifically, it is the correlation between the true breeding value and the estimated breeding value breeders are interested in. This is because this value is related to the response to selection (Lynch and Walsh, 1998; Bourdon and Bourbon, 2000). By choosing models and methods that yield predictions with high correlations with the true breeding values, a breeder can increase the response to selection.

1.6.3 The animal models

The models in this thesis are based on the animal model and the extensions that are presented in this section. The animal model is a GLMM that uses pedigree information to partition the observed phenotypic variance into different genetic and environmental components. This model is widely used in both animal and plant breeding (Lynch and Walsh, 1998; Bourdon and Bourbon, 2000), and in this thesis it is used to estimate and predict breeding values, where we have assumed Gaussian distributed phenotypic observations.

Under the simplest form of the animal model, the phenotypic observation for individual i is expressed as

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + a_i + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{w}_i$ is a vector of covariate effects for individual i, $\boldsymbol{\beta}$ contains fixed effects, a_i is the breeding value, and ε_i is a residual effect. The breeding value is modeled as a random effect assuming $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$ given the pedigree, where σ_a^2 is the additive genetic variance in the base population. The elements of $\boldsymbol{A}$ are $A_{ij} = 2\Theta_{ij}$, where Θ_{ij} is the coefficient of coancestry between individuals i and j, a measure of expected relatedness (Lynch and Walsh, 1998). The residual effects are assumed to the identical and independently distributed according to $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$. Solving the animal model using Henderson's mixed model equations (Henderson, 1950), assuming known covariance matrices for $\boldsymbol{a}$ and $\boldsymbol{\varepsilon}$, provide best linear unbiased predictions (BLUP; McCulloch, 2003) for the breeding values. Because of this, the animal model is often referred to as a BLUP model.

The animal model can be extended to model the effect of genomic markers such as SNPs. With current technology, genome-wide information about an individual can readily be obtained, either through SNP-array genotyping or sequencing platform (LaFramboise, 2009). Since the genome-wide information has become abundant, modeling this data has become the standard in plant and animal breeding. The application of this modeling has been shown to improve genetic gains in breeding during the last decade (Meuwissen et al., 2001; Hickey et al., 2017; Ibanez-Escriche and Simianer, 2016). This powerful tool, using high-density SNP marker panels, can be used to predict breeding values, and is then known as genomic prediction (Meuwissen et al., 2001). The extended animal model that includes SNP marker effects, is commonly referred to as the SNP-BLUP model (Koivula et al., 2012)

$$y_i = \boldsymbol{w}_i^T\boldsymbol{\beta} + \boldsymbol{z}_i^T\boldsymbol{u} + \varepsilon_i,$$

where $\boldsymbol{z}_i$ is a vector of length q containing the genotype coding of the SNP marker values of individual i, and $\boldsymbol{u}$ are the random effects of the q SNPs, usually modeled as independent following a Gaussian distribution, $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma_u^2)$. Typically there are many more SNP-markers than trait observations, which raises identification issues (de los Campos et al., 2009; Gianola et al., 2009).

Instead of modeling the marker effects directly, SNP markers are sometimes used to model the genomic relationship between individuals. This is commonly referred to as a genomic BLUP (GBLUP; Wang et al., 2018) model. The GBLUP model substitutes the pedigree-based relationship

matrix $\boldsymbol{A}$ with the marker-based relationship matrix $\boldsymbol{G}$, and is given by

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + g_i + \varepsilon_i,$$

where the genetic effect is assumed to be $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}\sigma_g^2)$, where $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}^T/k$. $\boldsymbol{Z}$ is a column centered genotype matrix, and $k = 2\sum_l \varrho_l(1-\varrho_l)$ with ϱ_l as the allele frequency of marker l (VanRaden, 2008).

A common inference method for the animal model is REML, often performed using the implementation in the ASReml program (Butler et al., 2009). MCMC methods have been popular within breeding research, but is rarely used within the breeding industry, probably due to the computational costs. Animal models with genetic dependency explained by the pedigree have a GMRF structure (Steinsland and Jensen, 2010), which means that INLA can be used and is very efficient (Holand et al., 2013; Steinsland et al., 2014). Animal models with genetic dependency explained by SNP markers belong in the class of LGMs, but the precision matrix $\boldsymbol{G}^{-1}$ is dense. Estimation can still be done using INLA, but it is not possible to take full advantage of INLA's sparse matrix computational benefits.

Chapter 2

Scientific papers

The previous chapter introduced a number of topics relevant for the scientific papers in this thesis. This chapter identifies the scientific contributions of the papers, how they are connected, and present their importance and relevance first as a unit, then separately.

The enclosed scientific papers aim to contribute knowledge-based statistical models for learning from data in the field of quantitative genetics and breeding. We have aimed to use and propose models that are realistic and aligned with scientific understanding, that are interpretable, and possible to draw inference from with the available data and existing methods.

Thesis goal

With this thesis we have tried to fill gaps in literature for models in quantitative genetics that allow inclusion of prior knowledge. The scientific papers in this thesis have all been motivated by challenges in plant and animal breeding, where an important goal is to estimate and predict breeding values. From an applied point of view, the aim of this thesis is to contribute to improving estimates and predictions of breeding values and haplotype effects.

Working hypothesis

Our working hypothesis, the temporarily accepted basis for research, has been that we can improve predictions of random genetic effects by using

models that include knowledge about genetic and spatial processes, and that are aligned with current scientific understanding, to the extent of what is possible with available data and existing inference methods.

To be able to get good estimates of breeding values and haplotype effects, it is necessary with a well designed breeding program, so that an individual's trait can provide insight to its underlying genetic value, and to use statistical models and methods that are able to separate the genetic effects from the environmental effects. Our focus is on proposing statistical models for improving the breeding values and haplotype effects, and the design of breeding programs is therefore outside the scope of the thesis.

Enablers for scientific innovation

There are several available scientific contributions that have made inference and evaluation with the proposed models possible. The Bayesian framework has appealing properties for inclusion of prior knowledge, and the use of novel inference methods like INLA and the SPDE approach enable fast and accurate inference with models that was not feasible 10-20 years ago. The hierarchical generalized linear models framework is also suited for inclusion of knowledge through the variance components of random effects, using h-likelihood.

An important enabler for evaluation with realistic data, is the recent development of software for breeding program simulation (e.g., Faux et al., 2016; Gaynor et al., 2019). Further, recent technological advances have made genome-wide information abundant (LaFramboise, 2009), and the availability of such data has inspired the development of tools over the last decades such as genomic prediction (Meuwissen et al., 2001; Ibanez-Escriche and Simianer, 2016; Hickey et al., 2017).

Innovation

The models and model components we propose in the scientific papers make up the innovation of the scientific approach, and are the main contributions of the thesis. We propose variance and dependency models for genetic and spatial effects, that are in line with what is known about nature, to the extent of what is possible with current inference methods and data gathering methods. We propose models in both the Bayesian hierarchical framework and in the frequentist hierarchical generalized linear

model framework.

Prior information is included as prior distributions, or by choice of model for variances. For the priors of spatial effects, we use knowledge about spatial processes being continuous. Knowledge about size of parameters is included via prior distributions based on discussions with geneticists. Haplotypes are modeled using the knowledge of expected similarities between them, and knowledge about the importance of different genetic markers is included via linear predictors for the variances.

Evaluation

We evaluate the proposed models relative to commonly used models using simulated data, and use the results from real data as complement to the simulation results. The simulated data are mainly generated from mechanistic models imitating the genetic processes in nature and breeding procedures to the extent of what is feasible, and not from the proposed computer models that are simplified versions of nature. This leads to fair comparisons between the models.

To evaluate the models we have used mostly correlation and CRPS to compare predictive performance. The use of correlation is standard in the breeding community because it gives breeders insight about the response to selection. However, from a statisticians point of view, it is important to control and communicate uncertainty in estimates. Because of this, we propose using the CRPS, to evaluate probabilistic forecasts.

Since we have been able to generate data from mechanistic models, and make fair model comparisons, we have not focused on comparing inference methods, for example comparing the results using INLA and MCMC methods.

Documentation

Documentation is important to ensure reproducible results, and to make the research verifiable for other scientists. The papers in this thesis strive to give a clear description of data sets and data simulation procedures, inference methods, evaluation criteria, and chosen parameter values. The papers contain online supplementary material with coding examples that together with the papers themselves should be sufficient to ensure reproducibility and verification of the results in the papers. For the applications

with real data, some of the data sets are available online, whereas other are available on request.

The scientific papers

The following papers constitute the scientific contribution of this thesis.

Paper I: Selle, M.L., Steinsland, I., Hickey, J.M., and Gorjanc, G. (2019). "Flexible modelling of spatial variation in agricultural field trials with the R package INLA", published in *Theoretical and Applied Genetics*. Electronic supplementary material available from `link.springer.com/article/10.1007/s00122-019-03424-y#Sec31`.

Paper II: Selle, M.L., Steinsland, I., Powell, O., Hickey, J.M., and Gorjanc, G. (2020). "Modeling environmental variation in genetic evaluations for smallholder breeding programs".

Paper III: Mouresan, E.F, Selle, M., and Rönnegård, L. (2019). "Genomic Prediction Including SNP-Specific Variance Predictors", published in *G3: Genes, Genomes, Genetics*. Electronic supplementary material available from `doi.org/10.25387/g3.9247832`.

Paper IV: Selle, M.L., Steinsland, I., Lindgren, F., Brajkovic, V., Cubric-Curik,V., and Gorjanc, G. (2020). "Hierarchical modeling of haplotype effects based on a phylogeny", submitted to *Frontiers in Genetics*. Electronic supplementary material available from `doi.org/10.6084/m9.figshare.12024450`.

All papers present models or modeling approaches that aim to improve current methods by taking a knowledge-based approach, which is the statistical contribution of this thesis. The papers all strive to improve predictions of breeding values and haplotype effects in quantitative genetics with respect to correlation and CRPS. This is done through the development of new and combining existing variance and dependency models in new ways that allow inclusion of prior knowledge about different processes.

Overall, we conclude that including prior knowledge into the statistical models, in the form of knowledge about size of parameters or effects, spatial locations and distribution of the underlying processes, or expected similarities or relationships between effects, can improve predictions for

the random effects of interest. The results and conclusions from the papers allow and encourage scientists to use their expert knowledge, and to use models that are interpretable and realistic.

Paper I: "Flexible modelling of spatial variation in agricultural field trials with the R package INLA"

This paper analyzes agricultural field experiments using different established spatial dependency models in a Bayesian framework. The three models are: an independent row and column effects model, a separable first-order auto-regressive model (Cullis and Gleeson, 1991; Gilmour et al., 1997), and a GRF with Matérn covariance function represented as a GMRF via the SPDE approach.

The main contribution of this paper is to show how to include a continuous spatial model to agricultural field trials which allows for flexibility, is interpretable, and is easily available in the R (R Core Team, 2017) package implementing the INLA method. The flexibility opens opportunities for new field trial designs, and the interpretable parameters of the Matérn covariance function can allow plant breeders to get a better understanding of the underlying spatial processes affecting the observed phenotypes in the agricultural field trials. The separation of different effects in the phenotype allow better estimates of the breeding value, which we see from the results.

Knowledge about the underlying spatial process causing environmental variation in the phenotypes is included in the model via the prior distribution for the spatial effects. The processes in nature causing changes in soil fertility, watering and soil depth, are expected to be continuous. A GRF model is therefore suggested as the prior for the spatial effects, in addition to the two other more common models, and a case without a spatial model.

The main results show that the estimates of genetic effects can be improved by accounting for spatial dependency in trials irrespective of the magnitude of the spatial variation. The highest improvement is achieved when spatial variation is modeled using either the discrete first-order auto-regressive model or the continuous GRF.

The findings are based on estimation using simulated data. The simulated genetic effects are generated using a mechanistic model using the R package AlphaSimR (Faux et al., 2016; Gaynor et al., 2019). The simula-

tion of the spatial effects is performed assuming an underlying GRF, due to the assumptions about continuous spatial variation mentioned above. All models are also tested on simulated data where the spatial effects are generated from the discrete separable first-order auto-regressive model. The conclusions are the same as the ones drawn from findings with simulated data from the GRF.

Hyperparameters that are variance parameters are assigned prior distributions according to the inverse-gamma distribution. Although we experienced that changing from the inverse-gamma distribution to the penalized complexity priors changed the posterior for one of the variance parameters, we kept the results with the inverse-gamma priors to avoid overwhelming the target audience, which are geneticists, bio-technologists and breeders.

Paper II: "Modeling environmental variation in genetic evaluations for smallholder breeding programs"

This paper contributes to filling the gap for spatial modeling in animal breeding programs. There seems to be extensive literature on modeling genotype-by-environment effects (Tiezzi et al., 2017; Yao et al., 2017; Schultz and Weigel, 2019), but modeling spatial dependency in the environmental effect between herd locations on its own is not commonly done. The effects of herds are usually assumed to be fixed categorical effects or independent random effects. In this paper however, knowledge about farm location is included in the model to enhance separation of genetic and environmental effects, which is the key contribution of this paper. By this, we aim to improve genetic evaluation of smallholder animal breeding programs by including a spatially dependent environmental effect, and we consider the improvement for different strengths of genetic relatedness in the population.

The proposed models are applied to both simulated data and real cattle data. The simulated data are generated from a mechanistic model, using the R package AlphaSimR (Faux et al., 2016; Gaynor et al., 2019) to simulate the genetic effects. Three different scenarios for breeding the animals imitate different breeding strategies controlling the strength of genetic relatedness. The simulated spatial effects are linear combinations of eight sampled Matérn GRFs. This is because we want to mimic several layers of environmental effects such as different climatic effects. Further-

more, we are assuming the Mátern GRF as prior for the spatial effects, and we want the simulated spatial effects to come from a different model than the Mátern GRF.

We take a Bayesian approach to modeling, so knowledge about the relative size of the effects acting on the phenotype are included in the model by setting prior distributions for the variance parameters. Based on conversations with geneticists, who are experts in the field, we have set informative priors for the variance parameters.

The results show that including both an independent model and a spatially dependent model for herds gives the best separation of genetic effects and environmental effects. Powell et al. (2019) propose using genomic markers rather than pedigree information to model the dependency between individuals. While the pedigree is only able to capture the expected relationship between individuals, the genomic data are able to capture the realized relationship between individuals. The results from our paper support the findings of Powell et al. (2019), and show that a spatial model is less important to include when using genomic markers to capture relationships between individuals. However, our results also suggest that there is prospect for a better separation of environmental and genetic effects when including a spatially dependent herd model in addition.

Paper III: "Genomic prediction including SNP-specific variance predictors"

This paper proposes a general model for genomic prediction using a link function approach within the hierarchical generalized linear model framework, that can include external information on genomic markers. The motivation for this research was the increasing amount of available biological information on genomic markers (e.g., NCBI et al., 2020), and the expected increase of this information as the current technology for obtaining whole-genome information is becoming more available. Although a large number of methods have been developed already for genomic prediction (Meuwissen et al., 2001; de los Campos et al., 2009; Habier et al., 2011; Gianola, 2013; Zhang et al., 2014), there has been a gap in literature for a general linear mixed model to include explanatory variables for SNP-specific variances, that allow both continuous and categorical variables. This paper contributes to fill that gap.

Our aim is to assess the accuracy of the proposed variance models un-

der different genetic architecture. We conclude that the proposed models are able to improve estimation of breeding values relative to the standard SNP-BLUP model without external information. As more accurate information on genomic markers will become available, we believe that the proposed models will become more useful.

The proposed models fit in the class of LMM in the presented frequentist setting. Prior knowledge about genomic markers is included in the model by specifying covariates for the variance of the random marker effects. This gives the model a flexibility that is in line with our knowledge about the genetic process.

The conclusions in this paper are mainly based on results from simulation studies, where genomic data are simulated from a mechanistic model imitating different genetic architecture. Applying the proposed model to a real data set allows us to compare the model with an approach suggested by Zhang et al. (2015a), and we find that our results are comparable with theirs.

Paper IV: "Hierarchical modeling of haplotype effects based on a phylogeny"

This paper proposes a model component for modeling haplotype effects within quantitative genetics that is based on the phylogeny between the haplotypes. The motivation for this model came from the work of Kelleher et al. (2019), who are building phylogenies on large data sets, and the fact that most haplotypes are similar in effect, due to most mutations in the genome not having causal effects on phenotypes.

The main contribution of this paper is the development of an autoregressive model of order one that hierarchically models haplotype effects by leveraging phylogenetic relationships between the haplotypes described with a directed acyclic graph. The model, which we have called the haplotype network model, yields a sparse precision matrix for the haplotypes.

There is extensive literature on estimating haplotype effects (Templeton et al., 1987; Balding, 2006; Thompson, 2013; Morris and Cardon, 2019). One issue with estimating these effects is that there is usually an uneven distribution of haplotypes in a population (Ewens, 1972, 2004; Walsh and Lynch, 2018). This means that few haplotypes are frequently observed in individuals, but most haplotypes are observed in only a few individuals. Estimating the effects of rare haplotypes is therefore chal-

lenging. However, by using knowledge about the genetic processes that create a "network" of haplotypes, we are able to infer the effects of these rare haplotypes much better than using models that assume independent haplotypes, since effects of similar haplotypes are expected to be similar. This shows the importance and relevance of the paper.

A Bayesian framework is chosen, which makes it possible to incorporate knowledge about the genetic processes. From knowledge about mutations and haplotypes, it is expected that most haplotypes have similar effect when they are only a few mutations apart. Therefore we choose a prior for the auto-correlation parameter that has most mass close to 1.

Bibliography

Acquaah, G. (2009). *Principles of plant genetics and breeding*. John Wiley & Sons.

Alam, M., Rönnegård, L., and Shen, X. (2015). Fitting conditional and simultaneous autoregressive spatial models in hglm. *The R Journal*, 7(2):5–18.

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., Rousseau, J., et al. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis*, 7(3):503–532.

Ayyub, B. M. (2001). *Elicitation of expert opinions for uncertainty and risks*. CRC press.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature reviews genetics*, 7(10):781.

Berger, J. O., Bernardo, J. M., Sun, D., et al. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Bourdon, R. M. and Bourbon, R. M. (2000). *Understanding animal breeding*, volume 2. Prentice Hall Upper Saddle River, NJ.

Box, G. E., Draper, N. R., et al. (1987). *Empirical model-building and response surfaces*, volume 424. Wiley New York.

Box, G. E. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the royal statistical society: Series B (Methodological)*, 13(1):1–38.

Butler, D., Cullis, B. R., Gilmour, A., and Gogel, B. (2009). ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.

Claeskens, G. (2016). Statistical model choice. *Annual review of statistics and its application*, 3:233–256.

Claeskens, G., Hjort, N. L., et al. (2008). Model selection and model averaging. *Cambridge Books*.

Conner, J. K., Hartl, D. L., et al. (2004). *A Primer of Ecological Genetics*. Sinauer Associates Incorporated.

Cullis, B. and Gleeson, A. (1991). Spatial analysis of field experiments − an extension to two dimensions. *Biometrics*, pages 1449–1460.

Davidian, M. and Louis, T. A. (2012). Why statistics? *Science*, 336:12.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. *Genetics*, 182(1):375–385.

Dunson, D. B. (2008). *Random effect and latent variable model selection*. Springer.

Elias, A. A., Rabbi, I., Kulakow, P., and Jannink, J.-L. (2018). Improving genomic prediction in cassava field experiments using spatial analysis. *G3: Genes, Genomes, Genetics*, 8(1):53–62.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112.

Ewens, W. J. (2004). *Mathematical population genetics 1*. Springer-Verlag, New York, NY, 2 edition.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: models, methods and applications*. Springer Science & Business Media.

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., and Hickey, J. M. (2016). AlphaSim: Software for breeding program simulation. *The plant genome*, 9(3):1–14.

Ferrão, L. F. V., Ferrão, R. G., Ferrão, M. A. G., Francisco, A., and Garcia, A. A. F. (2017). A mixed model to multiple harvest-location trials applied to genomic prediction in coffea canephora. *Tree Genetics & Genomes*, 13(5):95.

Fisher, R. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33:503–513.

Fisher, R. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Chapman and Hall/CRC.

Gaynor, R. C. G., Gorjanc, G., Wilson, D., Money, D., and Hickey, J. M. (2019). *AlphaSimR: Breeding Program Simulations*. R package version 0.9.0.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. et al. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534.

Gentle, J. E. (2006). *Random number generation and Monte Carlo methods*. Springer Science & Business Media.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, 194(3):573–596.

Gianola, D., de los Campos, G., Hill, W. G., Manfredi, E., and Fernando, R. (2009). Additive genetic variability and the Bayesian alphabet. *Genetics*, 183(1):347–363.

Gilmour, A. R., Cullis, B. R., and Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 269–293.

Givens, G. and Hoeting, J. (2005). *Computational Statistics (Wiley Series in Computation Statistics)*. Wiley New Jersey.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Guttorp, P. and Gneiting, T. (2006). Studies in the history of probability and statistics XLIX On the Matérn correlation family. *Biometrika*, 93(4):989–995.

Guyon, X. (1995). *Random fields on a network: Modeling, statistics, and applications*. Springer Science & Business Media.

Habier, D., Fernando, R. L., Kizilkaya, K., and Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics*, 12(1):186.

Hagerty, M. R. and Srinivasan, V. (1991). Comparing the predictive powers of alternative multiple regression models. *Psychometrika*, 56(1):77–85.

Heinze, G., Wallisch, C., and Dunkler, D. (2018). Variable selection−A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449.

Henderson, C. (1950). Estimation of genetic parameters. *Annals of Mathematical Statistics*, 21:309–310.

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., Jones, C., Canales, C., Grattapaglia, D., Bassi, F., et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature genetics*, 49(9):1297.

Holand, A. M., Steinsland, I., Martino, S., and Jensen, H. (2013). Animal models and integrated nested Laplace approximations. *G3: Genes, Genomes, Genetics*, 3(8):1241–1251.

Ibanez-Escriche, N. and Simianer, H. (2016). Animal breeding in the genomics era [Special issue]. *Animal Frontiers*, 6. (Eds.).

Isik, F., Holland, J., and Maltecca, C. (2017). *Genetic data analysis for plant and animal breeding*, volume 400. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature genetics*, 51(9):1330–1338.

Koivula, M., Strandén, I., Su, G., and Mäntysaari, E. A. (2012). Different methods to calculate genomic predictions—comparisons of BLUP at the single nucleotide polymorphism level (SNP-BLUP), BLUP at the individual level (G-BLUP), and the one-step approach (H-BLUP). *Journal of dairy science*, 95(7):4065–4073.

Konishi, S. and Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. (2015). TMB: Automatic differentiation and Laplace approximation. *arXiv preprint arXiv:1509.00660*.

Lado, B., Matus, I., Rodríguez, A., Inostroza, L., Poland, J., Belzile, F., del Pozo, A., Quincke, M., Castro, M., and von Zitzewitz, J. (2013). Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. *G3: Genes, Genomes, Genetics*, 3(12):2105–2114.

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic acids research*, 37(13):4181–4193.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(4):619–656.

Lee, Y., Nelder, J. A., and Noh, M. (2007). H-likelihood: Problems and solutions. *Statistics and Computing*, 17(1):49–55.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2018). *Generalized linear models with random effects: Unified analysis via H-likelihood*. Chapman and Hall/CRC.

Leeb, H. and Pötscher, B. M. (2009). Model selection. In *Handbook of Financial Time Series*, pages 889–925. Springer.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 73(4):423–498.

Lynch, M. and Walsh, J. B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Assocs., Inc.

Martino, S. and Riebler, A. (2019). Integrated nested Laplace approximations (INLA). *arXiv preprint arXiv:1907.01248*.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83.

Matérn, B. (1960). Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsintitut*, 49.

McCulloch, C. E. (2003). Generalized linear mixed models. In *NSF-CBMS regional conference series in probability and statistics*, pages i–84. JSTOR.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157(4):1819–1829.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Morris, A. P. and Cardon, L. R. (2019). *Genome-Wide Association Studies*, chapter 21, pages 597–550. John Wiley & Sons, Ltd.

Morrison, D. A. (2016). Genealogies: Pedigrees and phylogenies are reticulating networks not just divergent trees. *Evolutionary biology*, 43(4):456–473.

Muir, W. (2007). Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, 124(6):342–355.

NCBI, National Library of Medicine (US), and National Center for Biotechnology Information (1988 (accessed February 10, 2020)). *National Center for Biotechnology Information (NCBI)[Internet]*. https://www.ncbi.nlm.nih.gov/.

Neudecker, H. (1969). A note on Kronecker matrix products and matrix equation systems. *SIAM Journal on Applied Mathematics*, 17(3):603–606.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.

Powell, O., Mrode, R., Gaynor, R. C., Johnsson, M., Gorjanc, G., and Hickey, J. M. (2019). Genomic data enables genetic evaluation using data recorded on LMIC smallholder dairy farms. *bioRxiv*, page 827956.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ripley, B. D. (1987). *Stochastic simulation*. John Wiley & Sons.

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H. A., and Strandberg, E. (2010a). Genetic heterogeneity of residual variance−Estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution*, 42(1):8.

Rönnegård, L., Shen, X., and Alam, M. (2010b). hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*, volume 104 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC.

Rue, H. and Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, 137(10):3177–3192.

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*, 71(2).

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4:395–421.

Schultz, N. and Weigel, K. (2019). Inclusion of herdmate data improves genomic prediction for milk-production and feed-efficiency traits within North American dairy herds. *Journal of dairy science*, 102(12):11081–11091.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Sorensen, D. and Gianola, D. (2002). *Likelihood, Bayesian, and MCMC methods in quantitative genetics*. Springer, New York, NY.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series B (statistical methodology)*, 64(4):583–639.

Steinsland, I. and Jensen, H. (2010). Utilizing Gaussian Markov Random Field Properties of Bayesian Animal Models. *Biometrics*, 66:763–771.

Steinsland, I., Larsen, C. T., Roulin, A., and Jensen, H. (2014). Quantitative genetic modeling and inference in the presence of nonignorable missing data. *Evolution*, 68(6):1735–1747.

Steyerberg, E. W. et al. (2019). *Clinical prediction models*. Springer.

Templeton, A. R., Boerwinkle, E., and Sing, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila. *Genetics*, 117(2):343–351.

Thompson, K. L. (2013). *Using ancestral information to search for quantitative trait loci in genome-wide association studies*. PhD thesis, The Ohio State University.

Tiezzi, F., de Los Campos, G., Gaddis, K. P., and Maltecca, C. (2017). Genotype by environment (climate) interaction improves genomic prediction for production traits in us holstein cattle. *Journal of dairy science*, 100(3):2042–2056.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability and statistics for engineers and scientists*, volume 9. Pearson.

Walsh, B. and Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Wang, J., Zhou, Z., Zhang, Z., Li, H., Liu, D., Zhang, Q., Bradbury, P. J., Buckler, E. S., and Zhang, Z. (2018). Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. *Heredity*, 121(6):648–662.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, pages 434–449.

Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2):974–994.

Xu, Y., Wu, Y., Song, C., and Zhang, H. (2013). Simulating realistic genomic data with rare variants. *Genetic epidemiology*, 37(2):163–172.

Yao, C., De Los Campos, G., VandeHaar, M., Spurlock, D., Armentano, L., Coffey, M., De Haas, Y., Veerkamp, R., Staples, C., Connor, E., et al. (2017). Use of genotype × environment interaction model to accommodate genetic heterogeneity for residual feed intake, dry matter intake, net energy in milk, and metabolic body weight in dairy cattle. *Journal of dairy science*, 100(3):2007–2016.

Zhang, Z., Erbe, M., He, J., Ober, U., Gao, N., Zhang, H., Simianer, H., and Li, J. (2015a). Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. *G3: Genes, Genomes, Genetics*, 5(4):615–627.

Zhang, Z., Li, X., Ding, X., Li, J., and Zhang, Q. (2015b). GPOPSIM: A simulation tool for whole-genome genetic data. *BMC genetics*, 16(1):10.

Zhang, Z., Ober, U., Erbe, M., Zhang, H., Gao, N., He, J., Li, J., and Simianer, H. (2014). Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PloS one*, 9(3).

Ziegler, A., König, I. R., and Pahlke, F. (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons.

Paper I

Flexible modelling of spatial variation in agricultural field trials with the R package INLA

Selle, M. L., Steinsland, I., Hickey, J. M., and Gorjanc, G. (2019) published in *Theoretical and Applied Genetics*

Theoretical and Applied Genetics
https://doi.org/10.1007/s00122-019-03424-y

ORIGINAL ARTICLE

Flexible modelling of spatial variation in agricultural field trials with the R package INLA

Maria Lie Selle[1] · Ingelin Steinsland[1] · John M. Hickey[2] · Gregor Gorjanc[2]

Received: 12 April 2019 / Accepted: 6 September 2019

Abstract

Key message **Established spatial models improve the analysis of agricultural field trials with or without genomic data and can be fitted with the open-source R package INLA.**

Abstract The objective of this paper was to fit different established spatial models for analysing agricultural field trials using the open-source R package INLA. Spatial variation is common in field trials, and accounting for it increases the accuracy of estimated genetic effects. However, this is still hindered by the lack of available software implementations. We compare some established spatial models and show possibilities for flexible modelling with respect to field trial design and joint modelling over multiple years and locations. We use a Bayesian framework and for statistical inference the integrated nested Laplace approximations (INLA) implemented in the R package INLA. The spatial models we use are the well-known independent row and column effects, separable first-order autoregressive (AR1 ⊗ AR1) models and a Gaussian random field (Matérn) model that is approximated via the stochastic partial differential equation approach. The Matérn model can accommodate flexible field trial designs and yields interpretable parameters. We test the models in a simulation study imitating a wheat breeding programme with different levels of spatial variation, with and without genome-wide markers and with combining data over two locations, modelling spatial and genetic effects jointly. The results show comparable predictive performance for both the AR1 ⊗ AR1 and the Matérn models. We also present an example of fitting the models to a real wheat breeding data and simulated tree breeding data with the Nelder wheel design to show the flexibility of the Matérn model and the R package INLA.

Introduction

In plant breeding, the main goal is to select individuals with the best performance as new market varieties or to select individuals with the best genetic potential as parents of the next generation. To this end, breeders use field trials to estimate genetic and breeding values of individuals. Spatial variation is common in such trials, and if not accounted for it can impact the estimation. There can be several sources of spatial variation in a field trial, such as changes in fertility, watering and soil depth. Other sources of spatial variation that often occur are external influences due to the way plots are treated, for example the effect of drilling, spraying and harvesting. This extraneous variation can be handled by the addition of further effects in a model, such as column or row effects.

Traditionally, spatial variation has been accounted for by using control plots, replications and blocks. These approaches do not account for fine-grained spatial variability, in particular they do not account for dependency

Communicated by Ian Mackay.

The Research Council of Norway, Grant Number: 250362; The UK Biotechnology and Biological Sciences Research Council, Grant Numbers: BB/L020467/1 and BBS/E/D/30002275.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00122-019-03424-y) contains supplementary material, which is available to authorized users.

✉ Maria Lie Selle
maria.selle@ntnu.no

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[2] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Edinburgh, UK

Published online: 18 September 2019

between neighbouring blocks and plots within blocks, which can affect the estimation of genetic values. Several models have been proposed to model spatial variation. One of the most widely used is the separable first-order autoregressive (AR1 ⊗ AR1) model introduced by Cullis and Gleeson (1991) and extended by Gilmour et al. (1997). It has been shown to fit well in many trials (e.g. Gilmour et al. 1997; Rodríguez-Álvarez et al. 2018). There are other models that can correct for spatial variation. For example, there is a whole class of Gaussian intrinsic models based on the seminal work of Besag and Higdon (1999), which have not gained much traction in plant breeding applications. Much has also been done on smoothing techniques, among which the recent SpATS approach explores two-dimensional smooth surfaces through the use of tensor product P-splines (Rodríguez-Álvarez et al. 2018). Nearest neighbour models are reviewed by Piepho et al. (2008), and the use of spatial kernels is also common (Elias et al. 2018; Mao et al. 2019).

Most of the popular spatial methods in plant breeding use lags between plot locations as a distance, while continuous spatial variation is not commonly addressed. If observations are irregularly spaced, the autoregressive and other models assuming equal spacing are not applicable. However, there are extensions to the autoregressive model, using covariance functions known as the power model and the exponential model (Schabenberger and Gotway 2017). The kernel methods presented in Elias et al. (2018) also use covariance functions based on Euclidean distance between plots.

In this paper, we limit the focus to spatial variation in agricultural field trials such as changes in fertility, watering and soil, not to the spatial variation occurring due to the way plots are treated. We model this spatial variation using different models with publicly available open-source software. We fit the common column and row effects and the separable first-order autoregressive AR1 ⊗ AR1 model (Cullis and Gleeson 1991; Gilmour et al. 1997). In addition, we fit a Gaussian random field (Matérn) model to the field trial via the stochastic partial differential equation (SPDE) approach introduced by Lindgren et al. (2011).

For inference, we use the Bayesian numerical approximation procedure known as the integrated nested Laplace approximations (INLA) introduced by Rue et al. (2009) with further developments described in Martins et al. (2013). The method is implemented in the R package INLA where models are fit with the `inla()` function with the same ease as using the base R functions `lm()` or `glm()`. INLA calculates marginal posteriors for all model parameters (fixed and random effects and hyper-parameters) and linear combinations of effects without using sampling-based methods such as Markov chain Monte Carlo (MCMC). It is based on numerical approximations and numerical methods for sparse matrices and is much faster than sampling-based methods (Rue and Martino 2007).

INLA has previously been compared with several other methods for statistical inference. One of these is Mathew et al. (2015) who compared INLA, MCMC (as implemented in the R package MCMCglmm; Hadfield et al. 2010) and restricted maximum likelihood (REML) (as implemented in the ASReml program; Butler et al. 2009), and found that INLA can be used for rapid and accurate estimation of genetic parameters. The computation time for INLA and REML was about the same and significantly shorter than with MCMC, which was also the conclusion of Holand et al. (2013). Huang et al. (2017) compared INLA and REML for spatial models and showed that the performance of INLA–SPDE was comparable to REML. We emphasize that these comparisons are not straightforward because different programs implement different computational methods as well as different models. For example, the R package INLA implements a full Bayesian analysis (using the INLA method), as does the R package MCMCglmm (using the MCMC method), while the ASReml program implements an empirical Bayes analysis (using a two-stage method where first hyper-parameters are estimated and then using these estimates the fixed and random effects are estimated). Gianola et al. (1986) and Sorensen and Gianola (2007) describe these differences in great detail.

The R package INLA is flexible with respect to the field trial design and to including several years and locations in the analysis. For example, it can fit designs beyond the standard lattice design, which we demonstrate with the Nelder wheel design used in forestry (Parrott et al. 2012). For a recent review and comprehensive treatment of the R package INLA, see Bakka et al. (2018) and Krainski et al. (2018).

The objective of this article was to test established spatial models for analysing agricultural field trials using the open-source R package INLA. This R package allows us to fit multi-trial data where designs vary between trials and do not necessarily have to be regular. With a simulation study, we show that the Matérn model performs equally well as the AR1 ⊗ AR1 model. Further, using the package enables full Bayesian analysis. We also fitted the models on wheat data from Lado et al. (2013) and on a simulated tree breeding data set with the Nelder wheel design to further demonstrate the flexibility of the Matérn model and SPDE approach implemented in the R package INLA.

Material and methods

In this section, we present the data for a simulated wheat breeding programme, a real wheat field trial and a simulated tree breeding trial with the Nelder wheel design. We also present the used statistical models, studied cases, how we

inferred model parameters and how we evaluated the different models.

Experimental design and data

Simulated wheat data

To evaluate and compare the proposed models, we have simulated a wheat breeding programme and corresponding field trials using the R package AlphaSimR (Faux et al. 2016; Gaynor et al. 2019). The simulation followed closely our previous work (Gaynor et al. 2017; Gorjanc et al. 2018), where we simulated a wheat-like genome and 30 years of a wheat breeding programme with field trials.

The ancestral wheat-like genome had 21 chromosomes, each with 1000 single nucleotide polymorphism markers and 1000 quantitative trait loci. Each year in the breeding programme was based on 100 crosses between 50 parental inbred lines with 100 doubled-haploid lines per cross, resulting in a total 10,000 lines. These were planted in headrows, and the 1000 best individuals were planted in a preliminary yield trial with 0.25 heritability. The 100 best went through a final stage of planting and selection. The 50 best individuals from the preliminary yield trial and the following stages were used as parents in the next year of the breeding programme. Selection was based on phenotype with the exception of the preliminary yield trial in years 20 through 30, where we used the estimated breeding value.

We have focused our attention to the preliminary yield trial, because this stage has low replication, which makes modelling of spatial variation important. The 1000 lines in the preliminary yield trial were planted in two locations, with plots randomly assigned, ensuring that each line was planted once in each location so that the two locations were considered as replicates. The fields in the two locations had the same design, plots arranged in a lattice with 50 rows and 20 columns. The distance between columns was twice as large as the distance between rows causing long and narrow plot shape.

We let the years 1 through 19 serve as burn-in years for the breeding programme, and for years 20 through 30 the plots in the preliminary yield trial were assigned spatially dependent effects. We sampled plot spatial effects from a Matérn model generated via the SPDE approach with a spatial range of 10 units. We varied the proportion of variation due to spatial effects to be 0%, 50%, or 75% of the residual variance, that is, with 50% a half of variation between plots was due to spatial effects and a half due to other unknown effects (plot residual). More detailed description of the Matérn model and the SPDE approach is given in the "Spatial effect" and "The SPDE approach to spatial modelling" sections. To simulate yield phenotypes, we summed the year, location, individual genetic, spatial dependent plot and independent plot residual effects. We sampled year and location effects from a Gaussian distribution with an expected value of 0 and variance equal to residual variance. Individual genetic effects were based on quantitative trait loci genotypes and corresponding allele substitution effects (Faux et al. 2016; Gaynor et al. 2019). We standardized the yield phenotype before the data analysis, by centring with the mean and scaling with the standard deviation across both locations within the same year.

The reason for simulating spatial effects from the Matérn model generated via the SPDE approach was that this generated realistic geostatistical spatial processes—the true underlying spatial variation in a field is more likely a continuous process rather than discrete process. However, we also simulated spatial effects according to the AR1 $\otimes$ AR1 model. We varied the proportion of variation due to spatial effects to be 0%, 50%, or 75% of the residual variance, and we set the autocorrelation parameter to be 0.8 in both row and column directions. This autocorrelation corresponds to a range of 10 units.

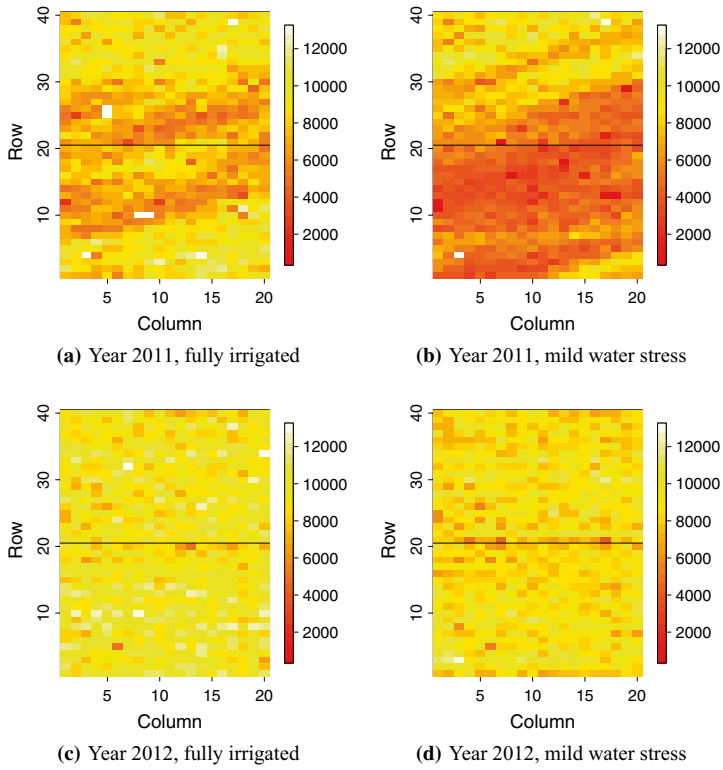
Chilean wheat data

We used parts of the wheat field trial data presented in Lado et al. (2013) and used by Rodríguez-Álvarez et al. (2018) as shown in Fig. 1. The data consisted of 384 advanced lines from wheat breeding programmes in Chile and Uruguay in years 2011 and 2012, and 16 additional lines that were not genotyped. The advanced lines were evaluated in the Santa Rosa region under two different levels of water supply: full irrigation (FI) and mild water stress (MWS). We analysed the total grain yield harvested within each plot.

The experimental design was an alpha-lattice with 20 incomplete blocks, with each block containing 20 genotypes. Two replicates were used for each year and irrigation level, so that each trial had 40 rows and 20 columns, and the lines were assigned the same plot for each year and irrigation level. According to Rodríguez-Álvarez et al. (2018), the replicates were placed such that the first/second 20 rows corresponded to the first/second replicate. This is indicated by the horizontal line in Fig. 1. Plots were twice as long as they were wide and consisted of five rows 2 m long and 0.2 m distance among the rows.

This gave four data sets each with 800 observations. The 384 genotyped lines had 102,324 genome-wide markers. We imputed missing genotypes with the average allele dosage and computed the VanRaden (2008) genomic relationship matrix among the 384 advanced lines. For the 16 lines not genotyped, but with phenotypic observations, we assumed a genomic relationship of zero between themselves and the 384 advanced lines.

One line had missing phenotypic observations for all replicates in 2011, and five other lines had missing phenotypic



(a) Year 2011, fully irrigated

(b) Year 2011, mild water stress

(c) Year 2012, fully irrigated

(d) Year 2012, mild water stress

Fig. 1 Grain yield in the Chilean wheat data (Lado et al. 2013)

observation for one replicate each. We standardized the yield phenotype before the data analysis, by centring with the mean and scaling with the standard deviation across all locations for multi-trial models and for each trial separately for the single-trial models.

Simulated tree data with the Nelder wheel design

We also simulated data with a design used by tree breeders to test the effect of multiple planting densities on tree growth, known as the Nelder wheel design (Parrott et al. 2012). We chose this particular design to show the flexibility of the R package INLA and the SPDE approach. The Nelder wheel design is circular with rings radiating outward with increasing distance. Spokes connect the centre with the furthest ring, and at the intersections of spokes and rings, a tree is planted. The variable planting densities within a single trial eliminate the need for separate trials for each planting density.

In the simulation, we tested 10 different planting densities with 30 planted trees for each density. The inner circle had a radius of 10, and the 9 subsequent circles had a radius of 1.15 times the radius of the previous circle (Fig. 2).

We simulated the phenotype for each tree as a sum of the intercept with a value of 10, the tree density covariate multiplied by a regression coefficient of 10, a spatial effect simulated from a Matérn model using the SPDE approach,

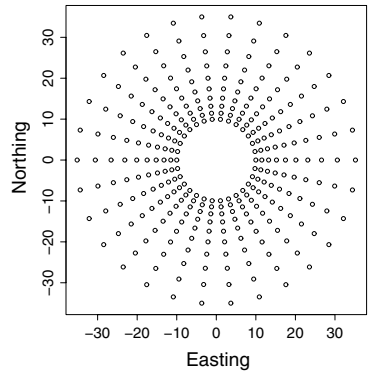


Fig. 2 Depiction of the Nelder wheel plot design

and a Gaussian residual with zero mean and variance 0.5. The simulated field (spatial effects) had a variance of 0.5 and a range of 10. There was no other effects to the design, that is, no treatment other than density, since modelling other genetic and environmental effects was illustrated with the simulated wheat data and the Chilean wheat data.

The growing area available to each tree i was calculated from:

$$\text{Growing area}\,(i) = \frac{\theta r_i^2 (k - k^{-1})}{2},$$

where θ is the angle between rays in radians and r_i is the radius of circle i. The factor k is 1.15. The planting density was then calculated as the inverse of the growing area.

Statistical models

We assumed to have n plots such that a single field trial was indexed by the rows and columns of an $r \times c$ array. There were $m \leq n$ different lines planted in these plots. The observed phenotype $y(s_i)$ was assumed to be a realization of a random variable $Y(s_i)$ in plot coordinates $s_i \in \mathbb{R}^2, i = 1, \ldots, n$. We considered the following general linear model:

$$y(s_i)|\eta(s_i), \sigma_e^2 \sim \mathcal{N}(\eta(s_i), \sigma_e^2), \tag{1}$$

with

$$\eta(s_i) = \beta_0 + \boldsymbol{w}_i\boldsymbol{\beta} + g_j + x(s_i), \tag{2}$$

where β_0 is an intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{n_f})$ is a vector of effects with a known covariate vector $\boldsymbol{w}_i$ for plot i with $\beta_i \sim \mathcal{N}(0, \sigma^2_{\beta_i})$ (for example year or location effects), g_j is the genetic effect for individual $j = 1, \ldots, m$ tested in the plot i and $x(s_i)$ is the spatial effect for the plot.

Genetic effect

We assumed that the genetic effect g_j was a sum of an additive genetic effect (breeding value) a_j and a non-additive (residual) genetic effect n_j. For non-additive genetic effects, we assumed an independent prior distribution $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m\sigma_n^2)$. For additive genetic effects, we assumed that they were fully explained by genome-wide markers such that $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\sigma_a^2)$, where $\boldsymbol{A}$ is a relationship matrix. We calculated the relationship matrix as $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^T/k$, where $\boldsymbol{Z}$ is a column-centred genotype matrix of dimension $m \times p$, p is the number of markers, and $k = 2\sum_l q_l(1 - q_l)$ with q_l being allele frequency at marker l (VanRaden 2008). An equivalent model for the additive genetic effects was to use the genotype matrix directly, letting $\boldsymbol{a} = \boldsymbol{Z}\boldsymbol{u}$, where $\boldsymbol{u}$ are marker effects $\boldsymbol{u} \sim \mathcal{N}(0, \boldsymbol{I}_p\sigma_u^2)$.

Genome-wide marker data contain substantial amount of shared information among related individuals due to shared genome segments. Therefore, we could compress it to reduce model dimension while retaining information, which saved computation time (e.g. Jolliffe 1986). With singular value decomposition, we obtained:

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T,$$

where $\boldsymbol{U}$ is a unitary matrix of dimension $(m \times m)$, $\boldsymbol{S}$ is the diagonal matrix $(m \times n)$ of singular values and $\boldsymbol{V}$ is an $(n \times n)$ matrix of eigenvectors. We used the principal components (the columns of $\boldsymbol{ZV}$) corresponding to the largest singular values of $\boldsymbol{S}$ and chose p^* components that explained approximately 95% of the variation in $\boldsymbol{Z}$. That is, we replaced the $\boldsymbol{Z}$ by $\boldsymbol{Z}^* = \boldsymbol{ZV}(:, 1 : p^*)$ of dimension $m \times p^*$. The linear predictor from (2) then became:

$$\eta_j(s_i) = \beta_0 + \boldsymbol{w}_i\boldsymbol{\beta} + \boldsymbol{z}_j^*\boldsymbol{u}^* + n_j + x(s_i), \tag{3}$$

where $\boldsymbol{z}_j^*$ is the jth row vector of $\boldsymbol{Z}^*$ for individual j and $\boldsymbol{u}^* \sim \mathcal{N}(0, \mathbf{I}_{p^*}\sigma^2_{u^*})$ are principal component effects.

Spatial effect

We tested the independent row and column effects model, the separable first-order autoregressive (AR1 $\otimes$ AR1) model and a Gaussian random field (Matérn) model via the SPDE approach. The independent row and column model and separable autoregressive model are based on a discretization of the field and model only a finite collection of spatial random variables. For these models, we omit the s_i in $x(s_i)$ and use x_i. This is to emphasize that these models use neighbouring plots as opposed to the Gaussian random field which is a continuous spatial process and for which we use the notation $x(s_i)$.

Row and column effects model

Row and column effects can model the underlying smooth spatial field as well as external variation due to field management. We assumed:

$$x_i = r_i + c_i,$$

where $r_i \sim \mathcal{N}(0, \sigma_r^2)$ is the row effect and $c_i \sim \mathcal{N}(0, \sigma_c^2)$ is the column effect of plot $i, i = 1, \ldots, n$.

Separable autoregressive model, AR1 $\otimes$ AR1

The autoregressive model of order 1 (AR1) for the Gaussian vector $\boldsymbol{x} = (x_1, \ldots, x_r)$ is defined as:

$$\begin{aligned} x_1 &\sim \mathcal{N}(0, \sigma_x^2/(1 - \rho^2)), \\ x_i|x_{i-1} &\sim \mathcal{N}(\rho x_{i-1}, \sigma_x^2), \quad i = 2, \ldots, r, \end{aligned}$$

where $|\rho| < 1$.

For modelling the influence of neighbouring plots along rows and columns, the autoregressive model in each

direction was combined into a two-dimensional first-order separable autoregressive model (Cullis and Gleeson 1991; Gilmour et al. 1997), denoted as AR1 ⊗ AR1. In this model the spatial effect vector $\boldsymbol{x}$ of length n was modelled as:

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}\sigma_x^2),$$

with $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_\mathrm{r} \otimes \boldsymbol{\Sigma}_\mathrm{c}$. The matrices $\boldsymbol{\Sigma}_\mathrm{r}$ and $\boldsymbol{\Sigma}_\mathrm{c}$ are the covariance matrices of first-order autoregressive processes in row and column direction, respectively, and ⊗ is the Kronecker product. The model had two dependency parameters, one in each direction, ρ_r and ρ_c, and a variance parameter σ_x^2.

Gaussian random fields and the Matérn model

In the model described above, the spatial variation was modelled as discrete, meaning that the model only considers data on a fixed field trial layout, possibly allowing the distance between rows to be different from the distance between columns. Assuming a continuous field for the spatial variation is, however, more realistic and allows the spatial variation to be modelled at any observed distance or field trial layout.

Continuously indexed Gaussian random fields play an important role in spatial statistical modelling and geostatistics. In the field $\mathscr{D} \in \mathbb{R}^d$ with coordinates $\boldsymbol{s} \in \mathscr{D}$, the continuously indexed Gaussian random field $x(\boldsymbol{s})$ has a joint Gaussian distribution for all finite collections $\{x(\boldsymbol{s}_i)\}$. The Gaussian random field is specified through the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma} = C(\boldsymbol{s}_i, \boldsymbol{s}_j)$.

In this study, we used $\boldsymbol{\mu} = 0$ and the Matérn covariance function, which is the most important covariance function in spatial statistics (Stein 2012). We refer to this Gaussian random field model as the Matérn model. The Matérn covariance function between locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^d$ was:

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\sigma_s^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa \|\boldsymbol{s}_j - \boldsymbol{s}_i\|\right)^\nu K_\nu \left(\kappa \|\boldsymbol{s}_j - \boldsymbol{s}_i\|\right), \tag{4}$$

where K_ν is the modified Bessel function of the second kind and order $\nu > 0$. The parameter κ can be expressed as $\kappa = \sqrt{8\nu}/\rho$, where $\rho > 0$ is the range parameter describing the distance where the correlation between two points was near 0.1, and σ_s^2 is the marginal variance. The parameter ν determined the mean-square differentiability of the field. The SPDE approach is a computationally efficient way to fit the Gaussian random field (Matérn) model (Lindgren et al. 2011), which we describe in the "The SPDE approach to spatial modelling" section.

Prior distributions

We used a full Bayesian approach to estimation which requires prior distributions for all parameters. The model consisted of two layers of parameters. The first layer consisted of fixed and random effects, for which we have specified most prior distributions above. In addition, a Gaussian prior with mean 0 and variance 1000 was assigned to the intercept and covariate effects, meaning $\sigma_{\beta_i}^2 = 1000$. The second layer consisted of the variance/dispersion parameters and other (spatial) parameters controlling the first layer and the likelihood for the data, i.e. all variance parameters, the parameters of the AR1 ⊗ AR1 and the Matérn models. For parameters in this layer, which we refer to as the hyperparameters, we used the default priors of the R package INLA. These are proper, but weak priors. For variance parameters, this was an inverse gamma prior with shape 1 and inverse scale 5×10^{-5}, which has 95% percentiles at approximately 0.009 and 0.010. In the separable autoregressive model, the same inverse gamma prior was set for the marginal variance $\sigma_x^2/(1-\rho^2)$. The transformed variable $\log((\rho+1)/(\rho-1))$ was assigned a Gaussian prior with mean 0 and standard deviation 0.15, which has 95% percentiles at approximately -0.15 and 0.15 for ρ. Priors for the Matérn model were specified for the parameters κ and τ that control spatial range and variance; see the "The SPDE approach to spatial modelling" section. We used the default joint Gaussian prior on $\log(\kappa)$ and $\log(\tau)$ with mean 0 and identity covariance matrix, so that $\log(\kappa)$ and $\log(\tau)$ were independent (Blangiardo and Cameletti 2015) and automatically scale to the size of the field.

Case studies

Simulation study

We fitted the model (1) with two versions of the linear predictor (3) to the preliminary yield trial of each simulated breeding programme—without and with genome-wide markers. The two linear predictors were:

$$\eta(s_i^k) = \beta_0 + \boldsymbol{w}_i\boldsymbol{\beta} + x(s_i^k) + n_j, \tag{5}$$

$$\eta(s_i^k) = \beta_0 + \boldsymbol{w}_i\boldsymbol{\beta} + x(s_i^k) + \boldsymbol{z}_j^*\boldsymbol{u}^*, \tag{6}$$

where β_0, $\boldsymbol{w}_i\boldsymbol{\beta}$, g_j, $\boldsymbol{z}_j^*\boldsymbol{u}^*$ were as described as in the "Statistical models" section. The linear predictors differed in that model (5) assuming that individuals were genetically independent, whereas model (6) used genome-wide marker data to model the genetic dependency. The linear predictors included both trials simultaneously. The k in s_i^k indicated that the plot coordinates s_i were in field k, where $k = 1, 2$, and a fixed effect of location was included in $\boldsymbol{w}_i\boldsymbol{\beta}$. Otherwise, the two locations were assumed to be independent realizations from the same distribution, and we used all three spatial models described in the "Spatial effect" section to fit spatial variation. We also fitted a model where the spatial effect was omitted, which we

denoted as the NoSpatial model. Since the distance between columns was twice as large as the distance between rows, we accounted for this with the Matérn model, by appropriately scaling the column coordinates. The matrix Z^* was constructed using $p^* = 500$ principal components of the singular value decomposition of the centred genotype matrix Z.

Chilean wheat data

Using the data sets from the four trials (Lado et al. 2013) presented in the "Experimental design and data" section, we fitted the model (1) with different versions of the linear predictor (3). The four linear predictors were:

$$\begin{array}{lll} \eta(s_i) = \beta_0 + x(s_i) & + n_j, & \text{W1: wheat model 1} \\ \eta(s_i) = \beta_0 + x(s_i) & + \boldsymbol{z}_j^* \boldsymbol{u}^* + n_j, & \text{W2: wheat model 2} \\ \eta(s_i^k) = \beta_k + x(s_i^k) & + n_j, & \text{W1M: use all trials} \\ \eta(s_i^k) = \beta_k + x(s_i^k) & + \boldsymbol{z}_j^* \boldsymbol{u}^* + n_j, & \text{W2M: use all trials} \end{array}$$

where β_0, $x(s_i)$, $\boldsymbol{z}_j^* \boldsymbol{u}^*$, and n_j are as described in the "Statistical models" section. As with the simulation study, we used the three spatial models described above and the NoSpatial model. The linear predictors W1M and W2M included all four trials simultaneously, and therefore the intercept β_k, $k = 1, \ldots, 4$, was trial specific to capture fixed year and irrigation effects. Further, the k in s_i^k indicated that the plot coordinates s_i were in field k. The four trials in Fig. 1 showed quite different spatial patterns with respect to dependency in distance and variance, so it was not reasonable to assume that they were realizations from the same distribution. However, assigning separate variance and parameters controlling the spatial dependency to each trial increased the number of hyper-parameters considerably. We therefore modelled the spatial effect in the trials from 2011 as independent realizations from the same underlying distribution, and the same for the 2012 trials, because these showed most similar behaviour. This gave two sets of spatial parameters in the model, one set for the 2011 trials and one set for the 2012 trials. We emphasize that this decision was driven by observation of the data. The matrix Z^* was constructed using $p^* = 280$ principal components of the singular value decomposition of the centred and scaled genotype matrix Z.

Nelder wheel plot

To analyse the simulated tree data, we fitted the model (1) with the following linear predictor:

$$\eta(s_i) = \beta_0 + w_i \beta + x(s_i),$$

where β_0 is the intercept, β is a density effect, and a Matérn model is assumed for the spatial effect $x(s_i)$. We also fitted a model where the spatial effect was omitted.

SPDE, inference and evaluation of case studies

The SPDE approach to spatial modelling

Modelling with Gaussian random fields is computationally challenging because they give rise to dense precision matrices that are numerically expensive to factorize in the estimation procedures (Rue and Held 2005). Gaussian Markov random fields do not incur this penalty because they have a sparse precision matrix due to their Markov property. Lindgren et al. (2011) showed how to construct an explicit link between (some) Gaussian random fields and Gaussian Markov random fields by showing that the approximate weak solution of the SPDE:

$$\begin{aligned} &(\kappa^2 - \Delta)^{\alpha/2} x(\boldsymbol{s}) = \mathcal{W}(\boldsymbol{s}), \\ &\boldsymbol{s} \in \mathbb{R}^d, \alpha = \nu + d/2, \quad \kappa > 0, \quad \nu > 0, \end{aligned} \tag{7}$$

is a Gaussian random field with Matérn covariance function as given in (4). Here, $\mathcal{W}(\cdot)$ is the Gaussian white noise, Δ is the Laplacian, α is a smoothness parameter, κ is the scale parameter in (4), d is the dimension of the spatial domain and τ is a parameter controlling the variance. The parameters of Matérn covariance are linked to the SPDE through:

$$\sigma_s^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}$$

where $\nu = \alpha - d/2$, and we use $\alpha = 2$ and $d = 2$.

A Gaussian Markov random field approximation described in Lindgren et al. (2011) is enabled by solving the SPDE in (7) by the finite element method. Further details on the SPDE approach to spatial modelling can be found in Lindgren et al. (2011).

Bayesian inference with INLA and the R package INLA

Statistical inference is carried out using the INLA method introduced in Rue et al. (2009), which is implemented for use in `R` (R Core Team 2018) in the R package `INLA` (available at www.r-inla.org). In this section, we give a short introduction to the class of models known as latent Gaussian models and how INLA can be used to approximate the posterior marginal distributions for such models. For an in-depth description of INLA, useful sources are Rue et al. (2009), Martins et al. (2013) and the recent review by Rue et al. (2017).

The class of latent Gaussian models includes many models, for example generalized linear (mixed) models, generalized additive (mixed) models and spline smoothing methods. Latent Gaussian models are hierarchical models in which observations $\boldsymbol{y}$ are assumed to be conditionally independent given a latent Gaussian random field $\boldsymbol{x}$ and hyper-parameters

$\boldsymbol{\theta}_1$, that is, $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) \sim \Pi_{i \in \mathscr{I}} \pi(y_i|x_i, \boldsymbol{\theta}_1)$. The latent field $\boldsymbol{x}$ includes both fixed and random effects and is assumed to be Gaussian-distributed given hyper-parameters $\boldsymbol{\theta}_2$, that is, $\pi(\boldsymbol{x}|\boldsymbol{\theta}_2) \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{\Sigma}(\boldsymbol{\theta}_2))$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are known as hyper-parameters and control the Gaussian field and the likelihood for the data. These are usually variance (dispersion) parameters for simple models, but can also include other parameters, for example autocorrelation. The hyper-parameters must also be assigned a prior density to completely specify the model.

The main aim of Bayesian inference is to estimate the marginal posterior distribution of the variables in the model, that is, $\pi(\theta_j|\boldsymbol{y})$ for hyper-parameters and $\pi(x_i|\boldsymbol{y})$ for location parameters. INLA computes approximations to these densities quickly and with high accuracy. Laplace approximations are applied to integrals that are Gaussian or close to Gaussian, and for non-Gaussian problems, conditioning is done to break down the approximations into smaller sub-problems that are almost Gaussian.

For the computations in INLA to be both quick and accurate, the latent Gaussian models have to satisfy some additional assumptions. Since INLA integrates over the hyper-parameter space, the number of non-Gaussian hyper-parameters $\boldsymbol{\theta}$ should be low, typically less than 10, and not exceeding 20. Further, the latent field should not only be Gaussian, it must be a Gaussian Markov random field. The conditional independence property of a Gaussian Markov random field yields sparse precision matrices which makes computations in INLA fast due to efficient algorithms for sparse matrices. Lastly, each observation y_i should depend on the latent Gaussian field through only one component x_i.

The R package INLA can be installed from within `R`. It is run using the `inla()` function with three mandatory arguments: a data frame containing the data, a formula much like the formula for the standard `lm()` function in `R` and a string indicating the likelihood family. The default is Gaussian with the identity link. The following call generates an object of type `inla`:

```
fit <- inla(data = Data, formula = Formula, family = "Gaussian")
```

Prior distributions are specified through additional arguments. Several tools to manipulate models and likelihoods exist as described in tutorials at the Web page www.r-inla.org and the books by Blangiardo and Cameletti (2015), Krainski et al. (2018). The `R` scripts used for the fitted models and the tree breeding simulation are available in Online Resource 1. Specifically we provide `R` code for all the fitted models to the real wheat data and the simulation and analysis of the tree breeding data with the Nelder wheel design.

Here, we show how to fit an: (1) Row + Col model, (2) AR1 row and AR1 col model, (3) AR1 ⊗ AR1 model and (4) Matérn model. The data should be stored in a data frame or list. Here, the data frame `Data` has one row for each observation with columns containing the phenotype, id for each line and row and column in the field. The id for each line is included twice because we want to model the genetic effect with and without genetic markers.

```
head(Data)
```

y	IndI	IndMI	Row	Col
0.254	275	275	1	1
0.029	325	325	1	2
-0.056	119	119	1	3
-0.128	138	138	1	4
0.531	296	296	1	5
-0.672	257	257	1	6

In the formula below, we indicate that each line should be modelled both with an independent normal distributed effect and using marker effects for the markers stored in `Gen`, the approach described in the "Genetic effect" section.

```
Formula <- y ~ f(IndI, model = "iid") + f(IndMI, model = "z", Z = Gen)
```

To include a spatial model, one of the following functions can be added to `Formula`.

```
(i): f(Row, model = "iid") + f(Col, model = "iid")

(ii): f(Row, model = "ar1") + f(Col, model = "ar1")

(iii): f(Row, model = "ar1", group = Col, control.group = list(model = "ar1"))

(iv): f(Field, model = Spde)
```

Here, `f()` indicates a random effect with a specific model. The `group` argument nests the random effect within each level of the group factor, and the `control.group` argument specifies the model between the group levels. The models with formula including either of effects (1)–(3) are fitted with the call to `inla()` as described above. The SPDE approach (4) requires a few additional stages which we show in the full code available in Online Resource 1.

Evaluation of model performance

We evaluated the models using the correlation between the true and estimated values, the continuous rank probability score (CRPS), by identifying the top individuals, and the residual variance.

We used the CRPS to take into account the whole posterior predictive distribution, that is, to compare the estimated posterior means with the true/observed values while accounting for the uncertainty of estimation. The CRPS is defined as (Gneiting and Raftery 2007):

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(u) - 1\{y \leq u\})^2 \, \mathrm{d}u,$$

where F is the cumulative distribution of the estimator of interest and y is the observed value. The CRPS is negative oriented, so the smaller the CRPS the closer the estimated value is to the observed/true value. For readers not familiar with the CRPS, three plots in Fig. 3 show the cumulative distribution functions for estimates and the observed value of 1.0. In Fig. 3a, the estimate is close to the true value and the area between the curves is small and so is the CRPS. In Fig. 3b, the estimated mean is equal to the true value, but the large uncertainty due to estimation causes a large area between the curves, and hence a larger CRPS than in Fig. 3a. In Fig. 3c, the uncertainty of the estimation is small, but the estimated mean is further from the true value, causing the area and the CRPS to be large.

For the simulated data, we computed the correlation and the CRPS between true and estimated breeding value. We also quantified how many of the ten best individuals were among the estimated top 100 individuals.

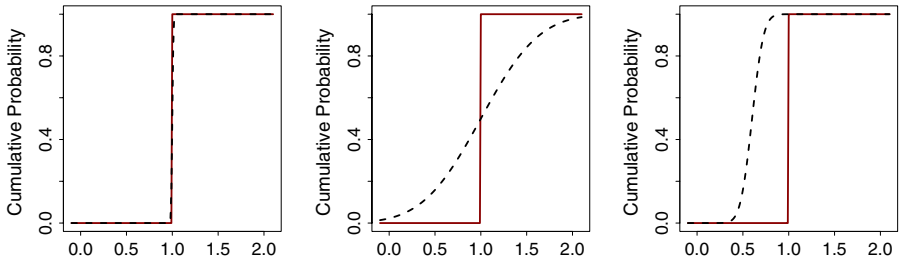
For the real data, we did not know the true breeding value, and it was therefore not possible to validate the estimated breeding values. We therefore focused on the residual variance from each model as a measure of the unexplained variance. This value can be seen as a proxy for the coefficient of determination (R^2), a measure on how much of the data variance is explained by a given model (Gelman and Hill 2006).

Results

In this section, we present the results from the three cases presented in the "Case studies" section. In the results from the simulation study, we compare correlation, CRPS and top ranking of individuals between the spatial models. In the results from the real data, we present estimated genetic variances, marker variances and residual variances and compare these between the different models. In the results from the simulated tree breeding data, we present the posterior distribution of all parameters and the estimated spatial effect.

Simulation study

This section presents the results from the simulation study. The models were evaluated using the correlation and CRPS between the true and estimated breeding value and using the number of the top ten individuals that were among the top 100 ranked individuals when considering estimated breeding value (posterior mean). In this section, all tables have three scenarios indicating the proportion of environmental



(a) Estimate $\mathcal{N}(1, 0.01^2)$; CRPS=0.002

(b) Estimate $\mathcal{N}(1, 0.5^2)$; CRPS=0.116

(c) Estimate $\mathcal{N}(0.6, 0.1^2)$; CRPS=0.344

Fig. 3 Cumulative distribution function (CDF) of the observation (true value = 1; solid line) and of estimate (dashed line)

variance due to spatially structured variation in the data, 0.00, 0.50 or 0.75, while the total variance was the same. Proportion of spatial variation therefore indicates how much of total environmental variance was due to structured spatial noise and unstructured noise (see the "Simulated wheat data" section).

In this section, we only present the results from data with spatial effects generated from the Matérn model via the SPDE approach. Tables with results for correlation, CRPS and ranking, based on data with spatial effect generated from the AR1 ⊗ AR1 model, are given in the Online Resource 2. These results show similar tendencies to the ones presented in this section. Tables showing average estimates for residual variance, genetic and marker variance and other spatial hyper-parameters based on data with spatial effect generated from the Matérn model are given in the Online Resource 2.

In Table 1, the average correlation is presented, in Table 2 the average CRPS is presented, and in Table 3 the average number of the top ten individuals that are among the top 100 ranked individuals is presented. The average was taken over 100 independent realizations of the breeding programme described in the "Simulated wheat data" section. We note that genomic data improved the correlation, CRPS and the average number of the top ten individuals for all models and proportions of spatial variance. We further note that modelling the spatial variation also improved these metrics. Below, we go through each table in detail.

Across all metrics, the Matérn and AR1 ⊗ AR1 stand out as best to model the spatial variation. These had the highest correlation when spatial variation was present as seen in Table 1. When there was no spatial variation, the two models did not perform worse than not including a spatial effect. The performance increased as the extent of spatial variation increased. The CRPS results in Table 2 show lower CRPS for the Matérn model and the AR1 ⊗ AR1 models compared to the NoSpatial and Row + Col models. These results are in line with the correlation results with one exception for the AR1 ⊗ AR1 model. We also note an improvement in CRPS with increasing extent of spatial variation.

The average number of the top ten individuals among the top 100 ranked individuals is given in Table 3. The Matérn and AR1 ⊗ AR1 models again had better results when there was a spatial variation in the data and when genome-wide markers were used—in this setting there were on average between 6 and 8 of the top ten individuals among the top 100 ranked individuals. As expected, the NoSpatial showed

Table 1 Correlation between the simulated true and estimated breeding value in the preliminary yield trial by the proportion of spatial variation, the spatial model and using genome-wide markers

Genome-wide markers	No			Yes		
Prop. of spatial var	0.00	0.50	0.75	0.00	0.50	0.75
NoSpatial	0.39	0.39	0.39	0.62	0.61	0.62
Row + Col	0.39	0.41	0.42	0.62	0.63	0.64
AR1 ⊗ AR1	0.39	0.47	0.56	0.62	0.68	0.74
Matérn	0.39	0.47	0.57	0.62	0.68	0.74

The standard error was around 0.002

Table 2 CRPS between the simulated true and estimated breeding value in the preliminary yield trial by the proportion of spatial variation, the spatial model and using genome-wide markers

Genome-wide markers	No			Yes		
Prop. of spatial var	0.00	0.50	0.75	0.00	0.50	0.75
NoSpatial	0.149	0.149	0.149	0.114	0.115	0.114
Row + Col	0.169	0.142	0.138	0.114	0.113	0.111
AR1 ⊗ AR1	0.169	0.127	0.117	0.114	0.108	0.100
Matérn	0.148	0.127	0.117	0.114	0.107	0.099

The standard error was around 0.0002

Table 3 Average number of the top ten individuals among the top 100 ranked individuals in the preliminary yield trial by the proportion of spatial variation, the spatial model and using genome-wide markers

Genome-wide markers	No			Yes		
Prop. of spatial var	0.00	0.50	0.75	0.00	0.50	0.75
NoSpatial	3.89	3.82	3.89	6.32	6.41	6.43
Row + Col	3.89	4.05	4.20	6.33	6.59	6.77
AR1 ⊗ AR1	3.89	4.81	5.81	6.32	7.36	8.07
Matérn	3.89	4.80	5.85	6.32	7.38	8.15

The standard error was around 0.05

no improvement when the degree of spatial variation was increased and the Row + Col model showed only a little improvement with respect to all evaluations.

We also evaluated predictions of breeding values for 1000 doubled-haploid individuals that were genotyped, but not phenotyped. These individuals served to test out-of-sample prediction, which we could perform using estimated genome-wide marker effects. The average correlation between the true and predicted breeding value is presented in Table 4, where AR1 ⊗ AR1 and Matérn again had the highest correlation. For the CRPS in Table 5, we see a similar trend as for the phenotyped individuals; however, the improvement with the higher degree of spatial variation is now less dominant. Finally, the average number of the top ten individuals among the 100 ranked individuals is given in Table 6. These results improved with the Matérn and AR1 ⊗ AR1 model and with the increasing spatial variation. The results for the non-phenotyped doubled-haploid lines showed lower correlation, higher CRPS and lower number of the top ten individuals captured than in the preliminary yield trial. This is expected as we had not observed any phenotype data on the doubled-haploid lines.

Table 5 CRPS between the simulated true and predicted breeding value for the non-phenotyped doubled-haploid lines by the proportion of spatial variation and the spatial model

Prop. of spatial var	0.00	0.50	0.75
NoSpatial	0.128	0.128	0.129
Row + Col	0.128	0.128	0.127
AR1 ⊗ AR1	0.128	0.126	0.122
Matérn	0.128	0.126	0.122

The standard error was around 0.00004

Table 6 Average number of the top ten individuals among the top 100 ranked individuals for the non-phenotyped doubled-haploid lines by the proportion of spatial variation and the spatial model

Prop. of spatial var	0.00	0.50	0.75
NoSpatial	3.37	3.38	3.35
Row + Col	3.37	3.51	3.60
AR1 ⊗ AR1	3.37	3.99	4.67
Matérn	3.37	3.97	4.75

The standard error was around 0.06

Chilean wheat data

In this section, we present results from fitting the models W1, W2, W1M and W2M to the Chilean wheat data. We present the estimated genetic variances, marker variances and residual variances from the different spatial models. These are shown in Fig. 4. We also present the posterior predicted phenotype from model W2 for the 2011 trial with full irrigation. Tables showing estimates for residual variance, genetic and marker variance, and other spatial hyperparameters are given in the Online Resource 2.

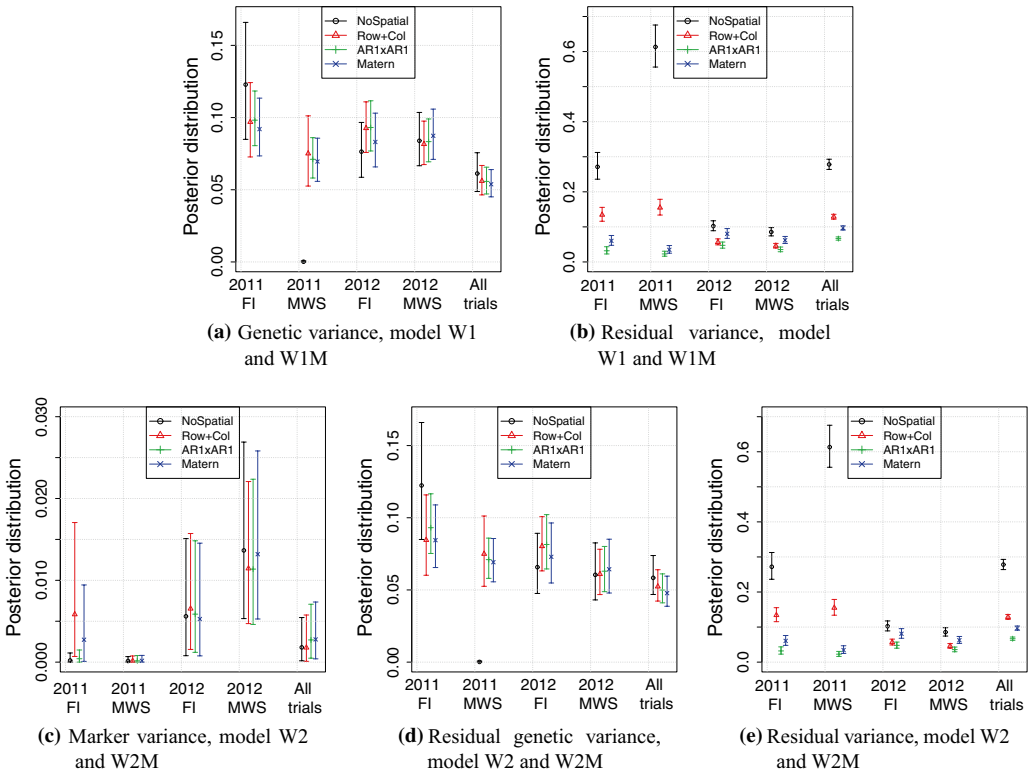
We first focus on the results from fitting the models without genome-wide markers (models W1 and W1M), which are shown in Fig. 4a, b. The estimated genetic variances were similar within each trial except for the NoSpatial case which assigned all variation to the residual variance in the trial from 2011 with mild water stress (MWS), indicating a very bad model fit. Between the trials, there was more variation between the estimates of genetic variance; however, most 95% confidence intervals overlap between the different models and trials with a few exceptions. The uncertainty in the genetic variance was reduced when all trials were analysed together (W1M), which was expected as more data were used in this model. For the residual variance, we expected that it would differ both between models and trials as they described the amount of variation not explained by the structured model terms. As expected, the residual variance from NoSpatial was the largest as this model cannot explain spatial variation. The AR1 ⊗ AR1 model had the lowest residual variance, closely followed by the Matérn model in the 2011 trials. When all trials were analysed jointly, the residual variance increased slightly for the AR1 ⊗ AR1 and the Matérn.

We now focus on the results for models including genome-wide markers (models W2 and W2M) in Fig. 4c–e. We note that marker variance estimate had large uncertainty and was lower in 2011, particularly in the medium-water stress condition. The genetic variance not captured by markers (Fig. 4d) became more similar between the different trials compared to model W1 (as summarrized in Fig. 4a). The residual variance did not change significantly, indicating that the markers captured the variation that was already captured by the genetic effect modelled in W1 and W1M. However, with genome-wide markers we captured the genetic dependency between individuals with the model, which makes it possible to predict genetic value for non-phenotyped individuals as shown in the previous subsection.

Table 4 Correlation between the simulated true and predicted breeding value for the non-phenotyped doubled-haploid lines by the proportion of spatial variation and the spatial model

Prop. of spatial var	0.00	0.50	0.75
NoSpatial	0.36	0.36	0.36
Row + Col	0.36	0.37	0.38
AR1 ⊗ AR1	0.36	0.42	0.47
Matérn	0.36	0.42	0.48

The standard error was around 0.004



(a) Genetic variance, model W1 and W1M

(b) Residual variance, model W1 and W1M

(c) Marker variance, model W2 and W2M

(d) Residual genetic variance, model W2 and W2M

(e) Residual variance, model W2 and W2M

Fig. 4 Posterior variances in the models for Chilean wheat data. The top panels are for models that do not use genome-wide marker data (W1 and W1M) and the bottom panels for models that use genome-wide marker data (W2 and W2M)

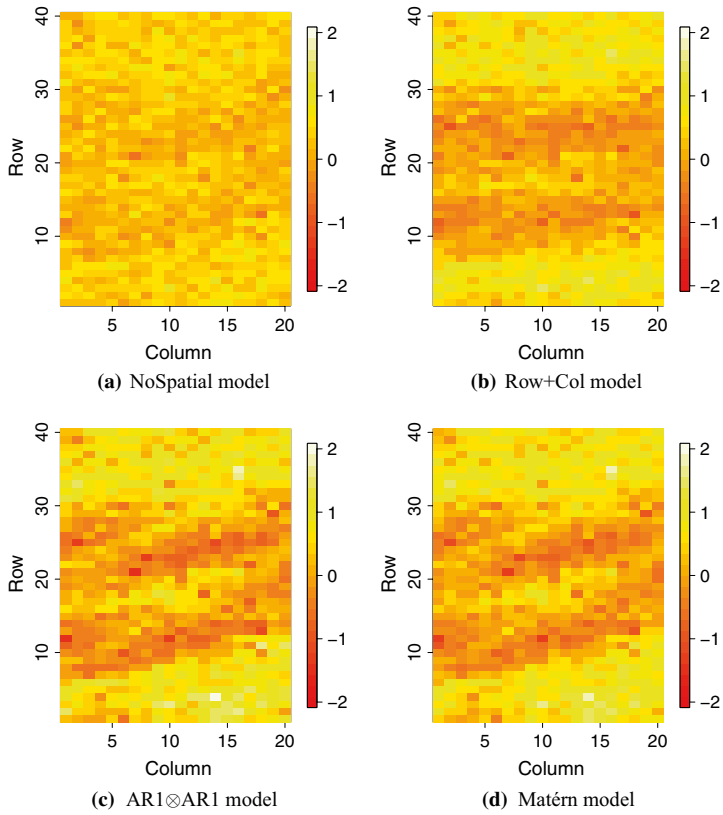
We show the fitted values from model W2 for the 2011 full irrigation trial in Fig. 5. These show how the AR1 ⊗ AR1 and Matérn models managed to capture the spatial pattern in the observations, whereas the NoSpatial model and Row + Col model could not. Since we do not know the true spatial effects for the data, we cannot know for a fact that this spatial variability is real. However, from the simulation study we showed that the models accounting for spatial variability do not perform worse than the NoSpatial model when there is no spatial variability acting on the phenotype. Note that the scale here is different from the one in Fig. 1 since models were fitted to standardized data.

Nelder wheel plot

In this section, we present the results from fitting the model presented in the "Nelder wheel plot" section to the simulated tree breeding data. In Fig. 6, the posterior distributions for the intercept, fixed density effect, spatial range, spatial variance and residual variance from the Matérn model are presented along with the true values used in simulating the data. For all parameters, the posterior distribution contained the true values and the distribution modes were close to the true values for the Matérn model.

For the NoSpatial model, the true effect of density is barely covered by the 95% confidence interval of the posterior distribution (Fig. 6b), and the true intercept is not covered (Fig. 6a). The posterior residual variance is approximately twice as large as the true residual variance in Fig. 6c. This is expected as the NoSpatial model cannot account for the spatial variation, and we therefore expect it to perform worse than the Matérn model in this comparison.

In Fig. 7, we show the simulated spatial effect, the posterior mean spatial effect and the standard deviation of the estimate. The mean estimate resembled closely the true spatial field, especially in locations where we had observations. The standard deviation was the smallest where we had



(a) NoSpatial model **(b)** Row+Col model

(c) AR1⊗AR1 model **(d)** Matérn model

Fig. 5 Posterior fitted values from the model W2 for trial 2011 FI using all three methods of spatial correction and no spatial correction

observations and where the observations were more densely observed.

Discussion

The objective of this paper was to test established spatial models for analysing agricultural field trials using the open-source R package INLA. We have fitted both spatial and genetic effects jointly in a simulated wheat trial data, a real wheat data set and a simulated tree breeding data set with the Nelder wheel design. Here, we highlight three points for discussion: (1) the importance of modelling spatial variation in agricultural field trials, (2) the flexibility of the R package INLA and the SPDE approach to model multiple trials and years as well as non-standard designs and non-standard phenotype distributions and (3) the limitations of the R package INLA to estimate large numbers of hyper-parameters and to fit genomic models.

Modelling spatial variation

With the analysis of simulated wheat data sets, we showed that the estimates of genetic effects can be improved by accounting for spatial dependency in trials irrespective of the magnitude of the spatial variation. This is in line with the other studies (Elias et al. 2018; Rodríguez-Álvarez et al. 2018; Velazco et al. 2017; Piepho et al. 2008). We observed the greatest improvements with both the AR1 ⊗ AR1 model (Cullis and Gleeson 1991; Gilmour et al. 1997) and the Matérn model using the SPDE approach (Lindgren et al. 2011). We measured this improvement with the correlation and continuous rank probability score (CRPS) between the true and estimated effects as well as the average number of the top ten individuals that were among the 100 ranked individuals based on the estimates. When we attempted to model non-existing spatial variation, the results were not significantly worse compared to not modelling it. This observation suggests that the AR1 ⊗ AR1 model and the Matérn model are good default spatial models that do not overfit

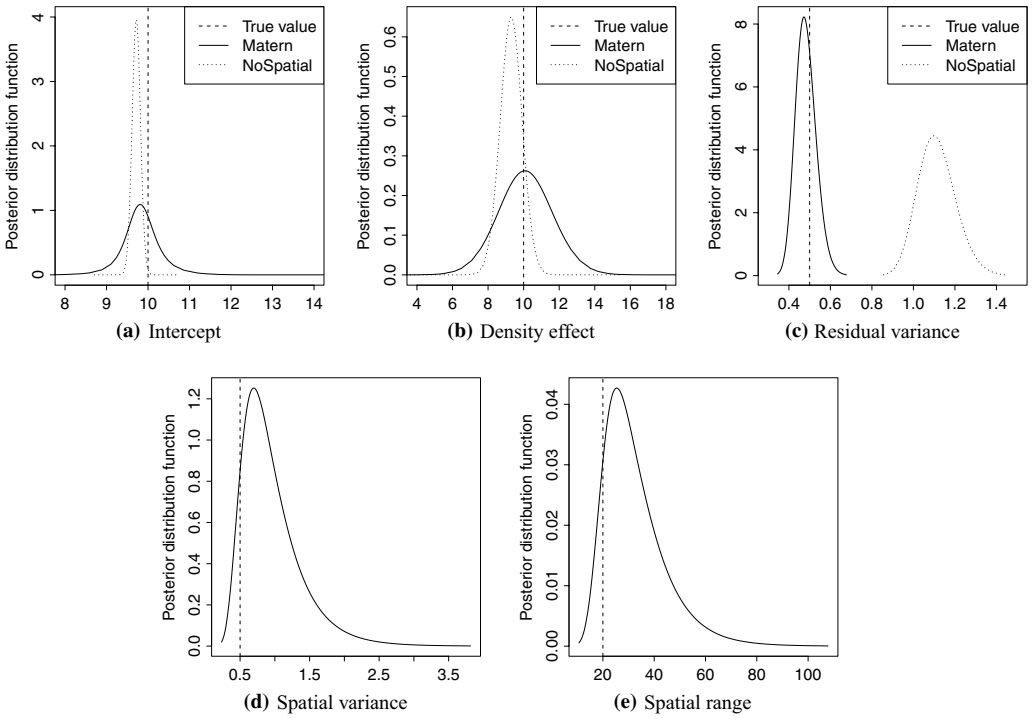


Fig. 6 Posterior distributions from model fitted to simulated tree breeding data. Full and dotted curves represent the posterior distribution and the straight dashed line the true values

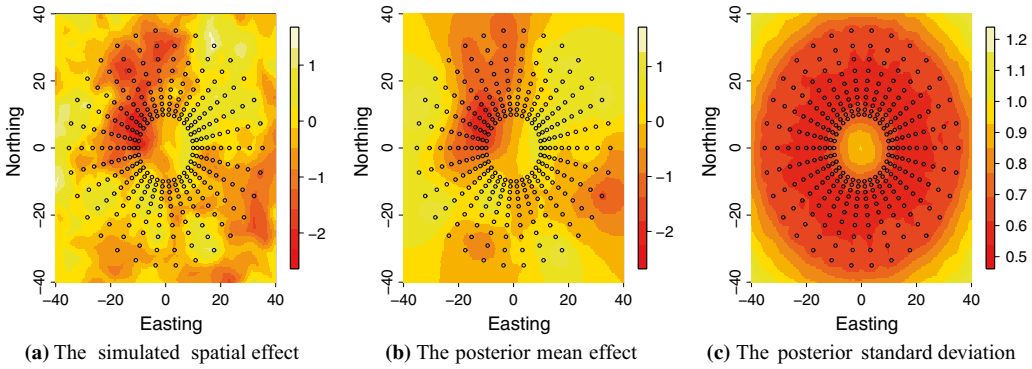


Fig. 7 Simulated spatial effect, posterior mean spatial effect and posterior standard deviation of spatial effect in the Nelder wheel example. Black circles indicate tree positions

the data. A reviewer pointed out that field trial design and management effects should be modelled in addition to spatial effects. When this is required (e.g. Borges et al. 2019; González-Barrios et al. 2019), the demonstrated R package INLA can easily accommodate this via its general model formulae functionality, that is, by adding block and sub-block effects and row and column effects. These effects can be modelled either as fixed or as random effects.

Flexibility of the R package INLA

Through modelling the real wheat data, we demonstrated the flexibility of the R package INLA to model both the genetic and spatial effects for several trials simultaneously. We treated the spatial variation in each trial as an independent realization of the chosen spatial model. By modelling spatial and genetic effects across several trials jointly with one model, we did not lose any information as we would if spatial effects were estimated first and then subtracted from the data (Schulz-Streeck et al. 2013). Furthermore, there is a large potential in modelling all trials jointly because this approach enables reduction of the required number of replicates per individual per trial and therefore test more individuals (Bernal-Vasquez et al. 2014). It also makes it possible to estimate location and year effects, which can be helpful for future management of the trial locations.

With the Nelder wheel design, we demonstrated the flexibility of the Matérn model using the SPDE approach with respect to the field trial design. This flexibility arises from the continuous modelling of spatial effects with the Matérn model as compared to the discrete approach of other standard models. The Nelder wheel example is a very special case and does not resemble standard agricultural field trials, which largely have a regular lattice layout of plots. We have nevertheless included this example to demonstrate the flexibility of the Matérn model and the R package INLA. This approach can be used for regular as well as non-regular designs, which can be useful in special settings, for example, when plot sizes differ (Archbold et al. 1987), when design is non-standard as in the Nelder wheel design (Parrott et al. 2012), when spatial correlation is not expected to follow standard patterns due to external variation (Bakka et al. 2019), or if the terrain does not allow for a lattice-like layout of plots. Another possible use of the Matérn model could be to jointly model neighbouring trials. In this case, the Matérn model can accommodate any layout of the plots across the trials, while the AR1 $\otimes$ AR1 model would require that plots from the neighbouring trials follow a common layout to all trials. Other applications of the Matérn model and the SPDE approach could be in conservation and utilization of genetic resources in forestry, particularly in natural or semi-natural stands not planted in a formal layout, and for identification of trees in the wild for collection of seed for cultivation or for reforestation. The approach can also make use of area observations (Lindgren et al. 2011; Bakka et al. 2018) to model total yield per area with varying area between plots. These flexibilities could enable design of new field trials or an advanced analysis of existing trials that do not follow the common lattice-like layout.

In this study, we focused on phenotypes that can be modelled with a Gaussian distribution only. However, the R package INLA enables seamless modelling of other distributions such as binomial, Poisson and others. Breeder's scores and other types of field trial data frequently follow these types of distributions. For most models, the only code change required is a switch of the distribution family; for example, to change the model with a continuous Gaussian distribution to a discrete Poisson distribution we simply change `inla(..., family = "Gaussian")` to `inla(..., family = "Poisson")`. Krainski et al. (2018) or Blangiardo and Cameletti (2015) provide further details on this. While the code change is simple, we have to note that the change of phenotype model impacts the interpretation of parameters. To this end, the R package INLA enables sampling from posterior distributions and these samples can be used to calculate parameters of interest. De Villemereuil et al. (2016) provide an excellent overview of this topic.

Limitations of the R package INLA

While the R package INLA enables flexible modelling of data from multiple trials and years, this might usually require increasing the model complexity by accounting for trial-specific residual variance or trial-specific spatial parameters—by increasing the number of hyper-parameters, that is, parameters controlling the likelihood and latent field, for example variance parameters. We have performed such an analysis with the real wheat data, where spatial variation in 2011 and 2012 trials differed substantially in both dependency with distance and variance. While this can be accommodated with the R package INLA, we highlight that the INLA method is best when it is based on a relative small number of non-Gaussian hyper-parameters, typically less than ten, and not exceeding 20. This limitation is due to the numerical integration of multidimensional posterior distribution of hyper-parameters in INLA (Rue et al. 2017). Since there is limited information to estimate hyper-parameters from a single trial, a parsimonious solution would be to group similar trials together and estimate hyper-parameters per group instead of per trial. This is what we did for the 2011 and 2012 trials with the real wheat data.

The main drawback with using R package INLA for analysing modern agricultural trials is that genome-wide marker data are highly dimensional, which leads to dense systems of equations. INLA is based on numerical approximations

and numerical methods for sparse matrices, and even though INLA can fit genomic models either via the genomic relationship matrix or via marker effects (Strandén and Garrick 2009), there is substantial computational overhead to handle such models, which is not the case for the pedigree model which has a sparse precision matrix (Steinsland and Jensen 2010; Henderson et al. 1984). This is why we chose to fit the genome-wide markers directly via the principal component approach, which is similar to the proposal of Ødegård et al. (2018). Another option would be to fit a model with individual genetic effects following VanRaden (2008), but with a genomic relationship matrix that uses dense–sparse partitioning into core and non-core individuals (Misztal 2016). More research is required in this area to increase the usefulness of the R package INLA for the modern breeding applications.

Finally, since the INLA method implements a full Bayesian analysis, prior distributions have to be set for all parameters of the model. The marker variance estimates in the models for Chilean wheat data were quite small, and we expected this to be larger. Testing the same models using the informative penalized complexity priors (Simpson et al. 2017) increased the mean marker variance. However, we have used the default prior distributions in the R package INLA for simplicity. It should be emphasized that using default priors is a choice as much as using any other prior or even using a specific distribution for the phenotype observations. Setting a prior based on the knowledge about the process is likely to improve the inference. Choosing a prior distribution for parameters in the model is not always straightforward, and more work is being done in the statistics community to improve this (Fuglstad et al. 2019).

Conclusion

This study showed how to fit established spatial models for analysing agricultural field trials using the open-source R package INLA. The results from the simulation study showed higher accuracy when spatial dependency was modelled and the highest increase in accuracy was reached using the discrete autoregressive (AR1 ⊗ AR1) model and the continuous Gaussian random field (Matérn) model. Both models can be seamlessly fitted with the R package INLA, including joint modelling of multiple trials. The Matérn model and SPDE approach provide a flexibility with respect to field design that is not obviously available elsewhere and are particularly suitable for agricultural field trials that do not have a standard lattice-like structure such as the Nelder wheel design used in tree breeding. This flexibility opens opportunities for new field trial designs. It is freely available and yields interpretable parameters for the estimated spatial effects.

Acknowledgements We would like to thank Kevin Wright for pointing to different field trials.

Author Contributions statement MLS, GG and IS conceived and designed the analysis. MLS and GG contributed the simulated data, and MLS performed the analysis. MLS wrote the manuscript, GG edited the manuscript, and IS and JH commented on the simulation design and the manuscript structure and content. All authors have read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Archbold D, Brown G, Cornelius P (1987) Rootstock and in-row spacing effects on growth and yield of spur-type delicious and golden delicious apple. J Am Soc Hortic Sci (USA) 112:219–222

Bakka H, Rue H, Fuglstad GA, Riebler A, Bolin D, Illian J, Krainski E, Simpson D, Lindgren F (2018) Spatial modeling with R-INLA: a review. Wiley Interdiscip Rev Comput Stat 10(6):e1443. https://doi.org/10.1002/wics.1443

Bakka H, Vanhatalo J, Illian JB, Simpson D, Rue H (2019) Non-stationary Gaussian models with physical barriers. Spat Stat 29:268–288

Bernal-Vasquez AM, Möhring J, Schmidt M, Schönleben M, Schön CC, Piepho HP (2014) The importance of phenotypic data analysis for genomic prediction: a case study comparing different spatial models in rye. BMC Genom 15(1):646. https://doi.org/10.1186/1471-2164-15-646

Besag J, Higdon D (1999) Bayesian analysis of agricultural field experiments. J R Stat Soc Ser B (Stat Methodol) 61(4):691–746

Blangiardo M, Cameletti M (2015) Spatial and spatio-temporal Bayesian models with R-INLA. Wiley, Hoboken

Borges A, González-Reymundez A, Ernst O, Cadenazzi M, Terra J, Gutiérrez L (2019) Can spatial modeling substitute for experimental design in agricultural experiments? Crop Sci 59(1):44–53. https://doi.org/10.2135/cropsci2018.03.0177

Butler D, Cullis BR, Gilmour A, Gogel B (2009) ASReml-R reference manual. The state of Queensland. Department of Primary Industries and Fisheries, Brisbane

Cullis B, Gleeson A (1991) Spatial analysis of field experiments-an extension to two dimensions. Biometrics 47(4):1449–1460

De Villemereuil P, Schielzeth H, Nakagawa S, Morrissey M (2016) General methods for evolutionary quantitative genetic inference from generalized mixed models. Genetics 204(3):1281–1294

Elias AA, Rabbi I, Kulakow P, Jannink JL (2018) Improving genomic prediction in Cassava field experiments using spatial analysis. G3 Genes Genom Genet 8(1):53–62

Faux AM, Gorjanc G, Gaynor RC, Battagin M, Edwards SM, Wilson DL, Hearne SJ, Gonen S, Hickey JM (2016) AlphaSim: software for breeding program simulation. Plant Genom 9(3):1–14

Fuglstad GA, Hem IG, Knight A, Rue H, Riebler A (2019) Intuitive principle-based priors for attributing variance in additive model structures. arXiv e-prints arXiv:1902.00242, arXiv:1902.00242

Gaynor RC, Gorjanc G, Bentley AR, Ober ES, Howell P, Jackson R, Mackay IJ, Hickey JM (2017) A two-part strategy for using genomic selection to develop inbred lines. Crop Sci 57(5):2372–2386

Gaynor RC, Gorjanc G, Wilson D, Money D, Hickey JM (2019) AlphaSimR: breeding program simulations. https://CRAN.R-project.org/package=AlphaSimR, r package version 0.9.0

Gelman A, Hill J (2006) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, Cambridge

Gianola D, Foulley JL, Fernando R (1986) Prediction of breeding values when variances are not known. Genet Sel Evol 18(4):485

Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. J Agric Biol Environ Stat 2(3):269–293

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

González-Barrios P, Díaz-García L, Gutiérrez L (2019) Mega-environmental design: using genotype × environment interaction to optimize resources for cultivar testing. Crop Sci. https://doi.org/10.2135/cropsci2018.11.0692

Gorjanc G, Gaynor RC, Hickey JM (2018) Optimal cross selection for long-term genetic gain in two-part programs with rapid recurrent genomic selection. Theor Appl Genet 131(9):1953–1966. https://doi.org/10.1007/s00122-018-3125-3

Hadfield JD et al (2010) MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J Stat Softw 33(2):1–22

Henderson CR et al (1984) Applications of linear models in animal breeding, vol 462. University of Guelph, Guelph

Holand AM, Steinsland I, Martino S, Jensen H (2013) Animal models and integrated nested Laplace approximations. G3 Genes Genom Genet 3(8):1241–1251

Huang J, Malone BP, Minasny B, McBratney AB, Triantafilis J (2017) Evaluating a Bayesian modelling approach (INLA-SPDE) for environmental mapping. Sci Total Environ 609:621–632

Jolliffe IT (1986) Principal component analysis and factor analysis. In: Principal component analysis. Springer series in statistics. Springer, New York, pp 115–128

Krainski ET, Gómez-Rubio V, Bakka H, Lenzi A, Castro-Camilo D, Simpson D, Lindgren F, Rue H (2018) Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall, London

Lado B, Matus I, Rodríguez A, Inostroza L, Poland J, Belzile F, del Pozo A, Quincke M, Castro M, von Zitzewitz J (2013) Increased genomic prediction accuracy in wheat breeding through spatial adjustment of field trial data. G3 Genes Genom Genet 3(12):2105–2114

Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J R Stat Soc Ser B (Stat Methodol) 73(4):423–498

Mao X, Dutta S, Wong RKW, Nettleton D (2019) Adjusting for spatial effects in genomic prediction. arXiv e-prints arXiv:1907.11581, arXiv:1907.11581

Martins TG, Simpson D, Lindgren F, Rue H (2013) Bayesian computing with INLA: new features. Comput Stat Data Anal 67:68–83

Mathew B, Léon J, Sillanpää MJ (2015) Integrated nested Laplace approximation inference and cross-validation to tune variance components in estimation of breeding value. Mol Breed 35(3):99

Misztal I (2016) Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics 202(2):401–409. https://doi.org/10.1534/genetics.115.182089

Ødegård J, Indahl U, Strandén I, Meuwissen TH (2018) Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genet Sel Evol 50(1):6

Parrott DL, Brinks JS, Lhotka JM (2012) Designing Nelder wheel plots for tree density experiments. New For 43(2):245–254

Piepho HP, Richter C, Williams E (2008) Nearest neighbour adjustment and linear variance models in plant breeding trials. Biom J 50(2):164–189

R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Rodríguez-Álvarez MX, Boer MP, van Eeuwijk FA, Eilers PH (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spat Stat 23:52–71

Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman and Hall, London

Rue H, Martino S (2007) Approximate Bayesian inference for hierarchical Gaussian Markov random field models. J Stat Plan Inference 137(10):3177–3192

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser b (Stat Methodol) 71(2):319–392

Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP, Lindgren FK (2017) Bayesian computing with INLA: a review. Annu Rev Stat Appl 4:395–421

Schabenberger O, Gotway CA (2017) Statistical methods for spatial data analysis. Chapman and Hall, London

Schulz-Streeck T, Ogutu JO, Piepho HP (2013) Comparisons of single-stage and two-stage approaches to genomic selection. Theor Appl Genet 126(1):69–82

Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH et al (2017) Penalising model component complexity: A principled, practical approach to constructing priors. Stat Sci 32(1):1–28

Sorensen D, Gianola D (2007) Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer, New York

Stein ML (2012) Interpolation of spatial data: some theory for kriging. Springer, New York

Steinsland I, Jensen H (2010) Utilizing Gaussian Markov random field properties of Bayesian animal models. Biometrics 66(3):763–771

Strandén I, Garrick D (2009) Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci 92(6):2971–2975

VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423

Velazco JG, Rodríguez-Álvarez MX, Boer MP, Jordan DR, Eilers PH, Malosetti M, van Eeuwijk FA (2017) Modelling spatial trends in sorghum breeding field trials using a two-dimensional P-spline mixed model. Theor Appl Genet 130(7):1375–1392

Paper II

Modeling environmental variation in genetic evaluations for smallholder breeding programs

Selle, M. L., Steinsland, I., Powell, O., Hickey, J. M., and Gorjanc, G. (2020)

Modeling environmental variation in genetic evaluations for smallholder breeding programs

Maria L. Selle [1] Ingelin Steinsland [1] Owen Powell [2]
John M. Hickey [2] Gregor Gorjanc [2]

March 27, 2020

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[2] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK

Abstract

Statistical models are used to separate the genetic and environmental effects in genetic evaluation for animal breeding programs. Herds or management groups are usually estimated using fixed or random effects to enhance the separation of the genetic and environmental effects. In smallholder dairy production systems of many low to middle income countries (LMIC) the genetic and environmental parts of the phenotype are often confounded due to small herd sizes and weak genetic connectedness across herds. Our working hypothesis was that estimating a spatially correlated random herd effect can enhance the separation of the genetic and environmental effects for smallholder dairy production systems, and thus improve genetic evaluation beyond the models assuming fixed or random independent herd effects. The objective of this study was therefore to use simulations and real data to quantify the power of a spatially correlated herd effect to improve genetic evaluation in smallholder breeding systems. The most important results showed that (i) including a spatial model improved the estimation and prediction of breeding values, (ii) spatial covariates did not improve estimates remarkably when a spatial model was included, (iii) the models without spatial effects were not able to separate genetic and spatial components, and (iv) the benefit of including a spatial model was largest when

the genetic and environmental components were hard to separate. We have demonstrated the potential of spatial modeling to improve genetic evaluation in LMIC smallholder dairy production systems. The improvement gained by the proposed models is driven by enhanced separation of the genetic and environmental effects. However, there are infrastructural and technological challenges that need to be solved before the LMIC smallholder breeding systems can benefit from this modeling.

1 Introduction

Over the past century genetic selection of dairy cattle has had a big impact on the increase in milk production in developed countries (Weigel et al., 2017). For example, the average milk production of US Holstein cows has almost doubled between 1960 and 2000, and more than half of this is due to improved genetics (Dekkers and Hospital, 2002). However, the same improvement in livestock productivity has not been achieved in low to middle income countries (LMIC), for example in the countries in East Africa. While large-scale farmers usually reach milk yields of 17–19 litres per cow per day, milk yields of smallholder producers in Kenya are about 5–8 litres per cow per day (Rademaker et al., 2016).

LMIC smallholders are constrained by both technological and infrastructural difficulties not present for farms in developed countries (Philipsson et al., 2011; Majiwa et al., 2017). Whereas large commercial farms in developed countries keep records of performance, pedigree, and can measure phenotypes accurately (Powell et al., 2019), the smallholders usually do not keep records (Ojango et al., 2019), and the absence of automated phenotyping systems leads to less accurate phenotypic measurements.

To get accurate genetic evaluations from a breeding program, a sufficient amount of data is needed, and the data should be properly structured (Jorjani et al., 2001). In developed countries, commercial farms usually have large herds, and there is widespread use of artificial insemination (AI), causing strong genetic connectedness between herds. In many LMIC breeding systems on the other hand, the smallholder farms contribute significantly to the dairy industry, and there is low genetic connectedness due low usage of AI. For example, smallholder milk-producing households in Kenya, who own one to three cows, own approximately 80% of the national dairy herds (Rademaker et al., 2016), and 87% of Kenyan farmers

asked in a survey used natural bull services rather than AI, even though 54% reported they preferred AI (Lawrence et al., 2015). Similar values for the proportion of natural bull services and AI usage was reported by Bebe et al. (2003) and Baltenweck et al. (2004).

Small herd sizes and low genetic connectedness across herds, lead to confounding of the genetic and environmental effects, which makes it hard to accurately estimate the genetic component of the phenotype. When the herd sizes are small, for example if a herd consists of only one cow, it is not possible to separate the genetic and environmental effects on the phenotype. When the genetic connectedness is low, the genetic relationship between animals in different herds is low, which also makes it hard to separate genetic and environmental components. Since most farmers use bull services, or in some cases their own or neighbor's bull, it is reasonable to assume that most farmers close in distance, for example farmers belonging to the same village, use the same bulls. This creates genetic connectedness across herds close in distance, for example across herds belonging to the same village, even though the overall genetic connecteness is low.

In the statistical models for genetic evaluations, the genetic effect is modeled using expected or realized relationship between animals, derived either from respectively a pedigree or genomic markers. A herd effect, or a herd-year-season effect is often included as the main environmental effect (Visscher and Goddard, 1993; Ojango et al., 2019; Pereira et al., 2019), to separate the genetic component of the phenotype from the environmental component. When herd sizes are small the herds are treated as random, as this has been found to give higher accuracy than treating them as fixed (Visscher and Goddard, 1993; Frey et al., 1997; Schaeffer, 2018; Powell et al., 2019). In addition, including other covariates in the statistical models is a way of including information in the model that can further enhance the separation of genetic and environmental effects.

Environmental effects can be on management level (herd level), or on a larger scale, possibly shared by herds close in distance. Examples of environmental effects on management level are education, age, experience, land size, cost of bull service and the use of AI. Some of these can be similar for herds close in distance. The practice and education level will probably be high for farmers that live in proximity to education facilities, the quality of feeding used in farms is likely similar in farms belonging to the same villages, and vaccination in farms is likely correlated with local,

regional or national government policies. These can therefore be similar between herds close in distance, for example herds that belong to the same village. Examples of large scale environmental effects are climate effects, proximity to roads, markets, and towns, and government policies. Many of the environmental effects, both the ones on management level and the large scale effects, can therefore be assumed to be spatially correlated. We will refer to environmental effects on management level as herd effects, and large scale environmental effects as spatial effects.

Genotype-by-environment interaction effects have been shown to exist in dairy cattle productive traits (Strandberg et al., 2009; Hayes et al., 2009), and there exists several studies on how to model these interactions. Genotype-by-environment interaction on larger geographic regions yielding environment-specific genomic parameters was modeled by Yao et al. (2017), and Schultz and Weigel (2019) have incorporated herdmate data to model genotype-by-herd interactions for prediction of within-herd performance. Although many aim to model the genotype-by-environment effect, it is not common to model the spatial dependency in the environmental effect between herd locations. One example is Tiezzi et al. (2017), who used geographical location and weather data in addition to herd summaries to describe environmental conditions in genetic evaluations, with and without genotype-by-environment interaction. They concluded that the farming environment explained variation in the data, as well as the genotype-by-environment component. An other example is Sæbø and Frigessi (2004), who proposed to model veterinary district as an environmental effect with prior spatial smoothing.

In many applications the collected data are geographically referenced, meaning their location in space is known, and these are called spatial data. The data collected from each herd are point-referenced data, with herd locations $\boldsymbol{s}$ typically two-dimensional containing latitude and longitude. The random outcome at the specific locations and the spatial indices can vary continuously in a fixed domain. A common model for spatial processes is a Gaussian random field (GRF) which for each set of locations $(\boldsymbol{s}_1, ..., \boldsymbol{s}_n)$, the vector $(y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))$ has a multivariate Normal distribution with mean $\boldsymbol{\mu}$ and a spatially structured covariance matrix $\boldsymbol{\Sigma}$ (Rue and Held, 2005).

Modeling with continuously indexed GRFs is computationally challenging because they give rise to dense precision matrices that are numerically

expensive to factorize (Rue and Held, 2005). Gaussian Markov random fields (GMRFs) do not incur this penalty because they have a sparse precision matrix due to their Markov property. Lindgren et al. (2011) showed how to construct an explicit link between (some) GRFs and GMRFs. This computationally effective approach is known as the stochastic partial differential equation (SPDE) approach, and allows implementation of computationally efficient numerical methods for spatial data.

The aim of this study was to determine whether modeling spatially dependent environmental effects in addition to independent herd effects could improve genetic evaluation and prediction in LMIC breeding systems, and to determine if the impact was dependent on the genetic connectedness across the herds, and the use of pedigree or genomic markers for modeling genetic relationships. In addition we wanted to test whether adding spatial covariates was necessary when we had a spatial model.

A simulation study was performed to evaluate the importance of including a spatially correlated effect in genetic evaluation. The design was developed to resemble the LMIC smallholder farming system commonly observed in East Africa with small herd sizes, and several breeding strategies with different genetic connectedness across herds. The results showed that including a spatial model improved genetic evaluations, especially with low genetic connectedness. We also performed a case study with cattle data from Slovenia, where the observations were sampled to resemble the LMIC smallholder farming system, and the results indicated that the models separated the genetic and environmental components in different ways.

2 Material and methods

We first introduce the data used in the analyses; a simulated smallholder dairy cattle data set, and a Slovenian Brown-Swiss cattle data set. Then we present the statistical models used for genetic evaluation, and how the models were evaluated.

2.1 Simulation models

We simulated data to evaluate the importance of including a spatial model to improve genetic evaluation. The design was developed to resemble the LMIC smallholder farming system structure commonly observed in

East Africa with small herds clustered around villages, and three different breeding strategies. The three breeding strategies controlled the genetic connectedness, from low genetic connectedness between herds from different villages, to strong genetic connectedness across all herds regardless of village. For each breeding strategy, we generated 60 independent data sets according to the following model for observation y_i

$$y_i = g_i + h_i + \xi_i + e_i, \tag{1}$$

where g_i was the genetic effect of individual i, $h_i \sim \mathcal{N}(0, \sigma_h^2)$ was the herd effect with $\sigma_h^2 = 0.25$, ξ_i was the spatial effect, and $e_i \sim \mathcal{N}(0, \sigma_e^2)$ was an independent residual effect with $\sigma_e^2 = 0.25$. We now describe in detail how the genetic and spatial effects were generated.

2.1.1 Simulation of genetic founder effects

The genetic effects were simulated from a burn-in phase designed to mimic historical evolution. A genome consisting of 10 chromosome pairs was simulated for a species similar to cattle. The sequence data was generated using the Markovian Coalescent Simulator (Chen et al., 2009) and AlphaSimR (Faux et al., 2016; Gaynor et al., 2019). The simulated genome sequences were used to produce 5000 founder individuals, who served as the initial parents. For each chromosome, sites segregating in the founders' sequences were randomly selected to serve as 5000 single-nucleotide polymorphism (SNP) markers and 1000 quantitative trait loci (QTL) per chromosome, yielding in total 50000 SNPs and 10000 QTL.

A single underlying trait architecture was simulated for all individuals via QTL allele substitution effects (Lynch et al., 1998) that were sampled from a standard normal distribution. The true breeding value of each individual was calculated by summing the QTL allele substitution effects. The single underlying trait architecture was used to create two correlated traits with different heritabilities (Lynch et al., 1998) for cows ($h^2 = 0.3$) and bulls ($h^2 = 0.8$), respectively. These phenotypes were used for the initial assignment of bulls and their selection throughout the evaluation phase.

2.1.2 Breeding and herd simulation

We created 100 villages, each consisting of 20 herds, with herd sizes generated from a zero truncated Poisson distribution with $\lambda = 1.5$. The 110 best males from the founder individuals based on true genetic values were set aside as selection bulls, 100 of them as natural selection bulls, and 10 of them as AI bulls. The remaining founders were considered as cows, and were randomly placed in the herds, until the herds were full. Since the herd sizes were sampled, we did not have the same number of individuals in each independent replicate. On average there were 3860 cows in total, and the cows not assigned to a herd were discarded.

The spatial coordinates in north-south direction and in east-west direction for each of the 100 villages were sampled from a uniform distribution on $(-1, 1)$. The spatial coordinates $\boldsymbol{s} \in \mathbb{R}^2$ of the 2000 herds were then sampled from a bi-variate normal distribution with mean from the corresponding village location, and variance $3.5 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$. This made the herds clustered around each village.

We created three different breeding strategies controlling the genetic connectedness in the population. In breeding strategy A, each village had their own bull, meaning that the cows were strongly related within the village and unrelated across villages. In breeding strategy B, each village had their own bull for mating in 75% of the herds, while mating in the remaining herds was performed with AI, using one of the 10 AI bulls at random, meaning that cows were still strongly related within villages, and somewhat related across villages. In breeding strategy C, the 100 natural selection bulls were randomly mated to cows across all herds and villages, meaning that cows were equally related within and across villages.

This population was then simulated over twelve discrete generations of breeding. Within each farm, old cows were replaced by new cows. The cows who had male calves were not replaced, and their calves were evaluated on phenotype for suitability as natural selection bulls if they came from a farm using natural selection bulls, or as AI bulls if they came from a farm using AI.

The true breeding values for cows in the 11th generation were scaled to have mean zero and variance $\sigma_g^2 = 0.1$, and were used as genetic effects in the model for observation y_i (1) with 3860 records on average. In addition, the true breeding values for new cows in the 12th generation were stored for prediction purposes. The estimated genetic effects of individuals in

generation 11 were used to predict the breeding values of non-phenotyped individuals in generation 12. For prediction using pedigree we had on average 1930 such records. To ease the computations with the genomic marker based model, we predicted breeding values for only 200 of the new cows in generation 12, which were chosen randomly.

2.1.3 Simulation of spatial effects

The spatial effects were simulated from multiple Gaussian processes to mimic several sources of environmental effects, both on spatial (large scale) and on management level (small scale). We imagined that these different sources could be temperature, precipitation, elevation, land size, proximity to markets and towns, education level of the farmer, vaccine use in farms, local or regional policies. We summed different Gaussian processes $\boldsymbol{v}$, where some were weighted to have different importance for the total spatial effect, according to

$$\boldsymbol{\xi} = \sum_{i=1}^{3} \boldsymbol{v}_i + \sum_{i=4}^{6} \boldsymbol{v}_i(1+\alpha_i) + \sum_{i=7}^{8} \boldsymbol{v}_i(1+\alpha_i+\beta_i)$$

where the weights $\alpha, \beta \sim$Uniform$(-0.5, 0.5)$, and the different sources $\boldsymbol{v}$ were distributed as GRFs with mean zero and Matérn covariance function (Matérn, 1960). The Matérn covariance function between locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^d$ is

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa \|\boldsymbol{s}_j - \boldsymbol{s}_i\|\right)^{\nu} K_{\nu} \left(\kappa \|\boldsymbol{s}_j - \boldsymbol{s}_i\|\right), \tag{2}$$

where K_ν is the modified Bessel function of the second kind and the order $\nu > 0$ determines the mean-square differentiability of the field. The parameter κ can be expressed as $\kappa = \sqrt{8\nu}/\rho$, where $\rho > 0$ is the range parameter describing the distance where correlation between two points is near 0.1, and σ^2 is the marginal variance. The range parameter ρ for each of the fields $\boldsymbol{v}$ was sampled from a uniform distribution on $(0.1, 0.5)$, the marginal variance σ^2 was either 0.2 or 0.3 with equal probability, and the parameter ν was fixed to 1. The final field $\boldsymbol{\xi}$ was scaled to have mean zero and variance $\sigma_\xi^2 = 0.4$.

2.1.4 Creating spatial covariates

The simulated Gaussian processes making up the spatial effects were used to create spatial covariates describing the different sources of spatial and environmental effects. In addition we sampled two GRFs with mean zero and a Matérn covariance function which were used as covariates not affecting the phenotype (dummy covariates).

For the processes $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3$ we assumed that we could observe the spatial covariates perfectly without any error, which could be reasonable for some covariate effects like average temperature and precipitation.

For the processes $\boldsymbol{v}_4, \boldsymbol{v}_5, \boldsymbol{v}_6$ we assumed that they were difficult to obtain accurately, so we added normal distributed error terms with mean zero and variance equal to 10% of the marginal variance of the weighted fields. This could be reasonable for some covariates that are difficult to measure or that vary with time. It could for example be difficult to quantify the area where cows are allowed to graze or the amount of different types of feeding used.

For the processes $\boldsymbol{v}_7, \boldsymbol{v}_8$ we assumed that we could only observe categorical values of the continuous effects, for example distance to markets and towns could be categorized as either a rural or urban area. For the process $\boldsymbol{v}_7$ we created a two-level categorical covariate by sampling a threshold from a uniform distribution between one standard deviation from the mean of $\boldsymbol{v}_7$ in both negative and positive direction. Values of $\boldsymbol{v}_7$ above the threshold were assigned one level, and values below were assigned the other level. For the process $\boldsymbol{v}_8$ we created a three-level categorical covariate by sampling two thresholds. The lower threshold was sampled from a uniform distribution between two standard deviations below the mean of $\boldsymbol{v}_8$ and the mean of $\boldsymbol{v}_8$. The upper threshold was sampled from a uniform distribution between the mean of $\boldsymbol{v}_8$ and two standard deviations above the mean of $\boldsymbol{v}_8$. The values of $\boldsymbol{v}_8$ were then assigned one of three categorical levels depending on which thresholds they were between. In this way we had ten spatial covariates, where the eight of them were continuous and two were categorical, and two of the continuous covariates did not affect the phenotype.

2.1.5 Changing the proportion of spatial variance and herd clustering

To evaluate how the models performed when there was no or little spatial effect on the phenotype, we created scenarios with different proportions of spatial variance relative to the sum of herd effect variance and spatial variance. We kept $\sigma^2_\xi + \sigma^2_h = 0.65$, and let $\sigma^2_\xi/(\sigma^2_\xi + \sigma^2_h) = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. This was repeated for 30 of the data sets.

We also wanted to evaluate the importance of how closely the herds were clustered around each village. To do this we changed the variance of the bi-variate distribution for the spatial coordinates $\boldsymbol{s} \in \mathbb{R}^2$ of the 2000 herds to $1 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$ which made the herds more closely clustered around the villages, and to $9 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$ which made the herds less clustered around the villages. This was repeated for each of the 60 data sets.

2.2 Case study: Brown-Swiss cattle data

We had phenotypic data for 30314 Brown-Swiss cattle from Slovenia collected between 2004 and 2019, from 2012 different herds. The data included a trait describing a body confirmation measure, year and scorer of the data, cattle age, stage of lactation, year and month of calving, and herd location coordinates. In addition the data contained a pedigree for 56465 animals including the cows with phenotypes. We analyzed the trait, which was centered and scaled by subtracting the phenotypic mean and dividing by the phenotypic standard deviation.

The average herd size was approximately 15 cows per herd, and most cows belonged to herds consisting of more than five animals. To imitate data similar to the typical LMIC design, with few individuals per herd, a subset of the full data was used. We sampled 3800 individuals from the full data without replacement, with sampling probability equal to the inverse herd size, meaning that larger herds had fewer records in the data subset. The subset contained cows from 1838 different herds, and the average herd size was about 2 cows per herd. The 1838 herds were scattered over most of Slovenia, and their locations are shown in Figure 1. The axes show the coordinates in kilometers from the Transverse Mercator coordinate system using datum WGS84.

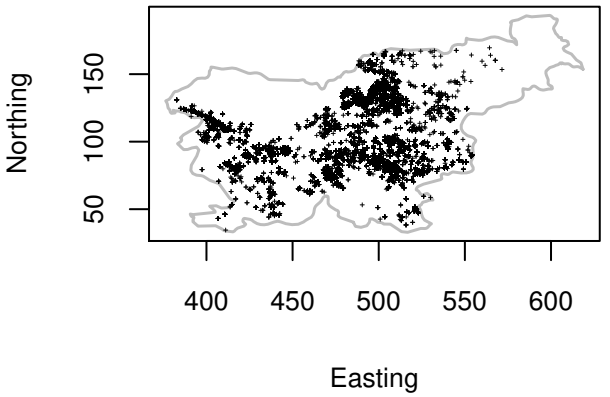


Figure 1: The location of the herds in the case study data shown with black points, and the border of Slovenia in grey. The axis units are in km

2.3 Statistical models

The following model was fitted to the observed phenotype y_i of individual $i = 1, ..., n$,

$$y_i = \boldsymbol{w}_i^T \boldsymbol{\beta} + a_i + h_i + x(\boldsymbol{s}_i) + e_i \tag{3}$$

where $\boldsymbol{\beta}$ is a vector of covariate effects, including a common intercept, with known covariate vector $\boldsymbol{w}_i$ and $\beta \sim \mathcal{N}(0, \sigma_\beta^2)$, a_i is the additive genetic effect, h_i is the herd effect with $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \sigma_h^2 \boldsymbol{I})$, $x(\boldsymbol{s}_i)$ is the spatial effect for the herd in plot coordinates $\boldsymbol{s}_i \in \mathbb{R}^2$ modeled as a GRF with $\boldsymbol{\mu} = \boldsymbol{0}$ and Matérn covariance function as given in (2) with variance σ_s^2 and range ρ, and e_i is a residual effect with $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$.

We assumed that the genetic effect could be explained by the additive genetic effect (breeding value), which was estimated using relationship matrix either based on pedigree or genome-wide markers. For the pedigree based model we assumed the breeding values were distributed as $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \sigma_a^2 \boldsymbol{A})$, where $\boldsymbol{A}$ was the relationship matrix derived from the pedigree (Lynch et al., 1998). For the genomic marker based model we assumed the breeding values were distributed as $\boldsymbol{a} \sim \mathcal{N}(\boldsymbol{0}, \sigma_a^2 \boldsymbol{G})$, where $\boldsymbol{G}$ was a relationship matrix calculated from $\boldsymbol{G} = \boldsymbol{Z}\boldsymbol{Z}^T/k$, $\boldsymbol{Z}$ was a column-centered SNP marker matrix, and $k = 2\Sigma_l q_l(1 - q_l)$, with q_l the allele frequency of marker l (VanRaden, 2008).

2.3.1 Prior distributions for hyper-parameters

We used a full Bayesian estimation approach which requires prior distributions for all parameters. For the intercept and fixed effects we assumed $\sigma_\beta^2 = 1000$, and for the remaining variance parameters and the spatial range we assumed penalized complexity (PC) priors (Simpson et al., 2017), which are proper priors that penalize model complexity to avoid over-fitting. The PC prior for variance parameters can be specified through a quantile u and a probability α which satisfy $\text{Prob}(\sigma > u) = \alpha$, and the PC prior for the spatial range parameter through a quantile u and a probability α which satisfy $\text{Prob}(\rho < u) = \alpha$. We state the parameters u and α below.

2.3.2 Fitted models for the simulation study

We fitted five different models to the simulated data; G, H, S, HS and HSC. All models had an intercept β_0, a genetic effect a_i, and a residual effect e_i. G had no additional model components, H had a herd effect h_i in addition, S had a spatial effect $x(\boldsymbol{s}_i)$ in addition, HS had both a herd effect and a spatial effect in addition, and HSC had a herd effect, a spatial effect and the spatial covariates $\boldsymbol{w}_i$ in addition. The models are summarized as

$$\begin{aligned}
\text{G: } y_i &= \beta_0 + a_i + e_i, \\
\text{H: } y_i &= \beta_0 + a_i + h_i + e_i, \\
\text{S: } y_i &= \beta_0 + a_i + x(\boldsymbol{s}_i) + e_i, \\
\text{HS: } y_i &= \beta_0 + a_i + h_i + x(\boldsymbol{s}_i) + e_i, \\
\text{HSC: } y_i &= \beta_0 + a_i + h_i + x(\boldsymbol{s}_i) + \boldsymbol{w}_i^T \boldsymbol{\beta} + e_i,
\end{aligned}$$

where $\boldsymbol{w}_i$ is the vector of spatial covariates for individual i and $\boldsymbol{\beta}$ is a vector of spatial covariate effects. The other effects were modeled as described in (3). The genetic effect was estimated using either pedigree or genomic markers. We used pedigree information for the phenotyped individuals, their offspring, and three previous generations. For the variances and spatial range we assumed PC prior distributions with quantiles u and probabilities α, shown in Table 1. Model HSC had the same quantiles and probabilities as model HS.

Table 1: Parameters u and α in the penalized complexity priors for variance parameters and the spatial range, for the models applied to the simulated data and case study data

	u_e, α_e	u_a, α_a	u_h, α_h	u_s, α_s	u_ρ, α_ρ [1]	u_ρ, α_ρ [2]
G	0.3, 0.5	0.1, 0.5	-	-	-	-
H	0.15, 0.5	0.1, 0.5	0.25, 0.5	-	-	-
S	0.15, 0.5	0.1, 0.5	-	0.25, 0.5	0.6, 0.95	50, 0.8
HS	0.15, 0.5	0.1, 0.5	0.15, 0.5	0.1, 0.5	0.6, 0.95	50, 0.8

[1] Simulation study
[2] Case study

2.3.3 Model evaluation with simulated data

For the simulated data, we evaluated the predictive performance of the models using correlation between the true breeding values and mean posterior breeding values, and the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007), comparing both the mean and standard deviation of the posterior breeding values with the true breeding values. The CRPS takes into account the whole posterior predictive distribution, meaning it compares the estimated mean posterior value with the true value while taking into account the standard deviation of the posterior distribution. The CRPS is negatively oriented, which means that lower CRPS values indicates a better estimate of breeding value.

2.3.4 Fitted models for the case study

To the case study we fitted four different models, G, H, S, and HS. All models had an intercept β_0, three categorical fixed effects (one describing the year and scorer of the data, one describing cattle age and stage of lactation, and one describing year and month of calving), a genetic effect a_i, and a residual effect e_i. G had no additional model components, H had a herd effect h_i in addition, S had a spatial effect $x(\boldsymbol{s}_i)$ in addition, and HS had both a herd effect and a spatial effect in addition. The genetic effect was estimated using the full pedigree. For the variances and spatial range we assumed PC prior distributions with quantiles u and probabilities α, shown in Table 1.

We used the deviance information criterion (DIC) (Spiegelhalter et al., 2002) to compare the model fit between the models fitted to the case study data. The DIC is widely used to compare model fit between different hierarchical Bayesian models while also assessing the model complexity. Lower values of the DIC indicate a better model fit.

2.4 Inference

For inference, we used the Bayesian numerical approximation procedure known as the Integrated nested Laplace approximations (INLA) introduced by Rue et al. (2009), with further developments described in Martins et al. (2013). INLA is suited for the class of latent Gaussian models, which includes for example generalized linear (mixed) models, generalized additive (mixed) models, spline smoothing methods, and the models used in this study. INLA calculates marginal posterior distributions for all model parameters (fixed and random effects, and hyper-parameters) and linear combinations of effects without using sampling-based methods such as Markov chain Monte Carlo (MCMC). For an in-depth description of INLA, useful sources are Rue et al. (2009), Martins et al. (2013) and the recent review Rue et al. (2017).

3 Results

In this section, we present the results from fitting the models to the simulated data and the Brown-Swiss cattle case study. We will refer to the mean posterior genetic effect for phenotyped individuals as the estimated breeding value (EBV), and the mean posterior genetic effect for non-phenotyped individuals as the predicted breeding value (PBV).

In the results from the simulation study, we compare average correlation between true breeding values and EBVs or PBVs between the tested models, and we compare the average CRPS between true breeding values and EBVs or PBVs with standard error between the tested models. In the results from the case study, we present the posterior variances from the tested models, the DIC, the posterior spatial effects, and show how the EBVs differ between two of the models (H and HS).

3.1 Simulation study

This section presents the results from the simulation study, where the models G, H, S, HS and HSC presented in Section 2.3.2 were fitted to data generated using different breeding strategies, A, B and C, where the genetic and environmental effects had different degrees of confounding as presented in Section 2.1.

Overall, the results showed that in a LMIC context (i) including a spatial model improved the estimation and prediction of breeding values, (ii) the spatial covariates did not improve the results remarkably when a spatial model was included, (iii) the models without spatial effects were not able to separate genetic and spatial components, (iv) the benefit of including a spatial model was largest when the genetic and environmental components were most confounded, (v) including a spatial model to the random herd effect even when there was no spatial effects did not decrease the prediction accuracy, and (vi) when spatial and genetic effects were confounded the estimation accuracy improved when herds were weakly clustered rather than closely clustered. We go through each of these findings in detail below.

3.1.1 Improving estimated and predicted breeding values via spatial modeling

Including a spatial model improved the estimation and prediction of breeding values. Table 2 presents the average correlations between true breeding values and EBVs or PBVs for all models, and breeding strategies. Across all metrics, model HS gave the highest correlations, when we do not consider model HSC. The second best was S, third was H, and the poorest was G. We also note that using genomic data improved the correlation compared to using pedigree, and that the EBVs had overall higher correlation than the PBVs. With breeding strategy A, the correlations for PBVs were comparable to the correlations for the EBVs, and the models using pedigree had almost as high correlation as the models using genomic markers. This is reasonable since the individuals from strategy A were strongly related within the villages and by the pedigree.

Table 3 presents the average CRPS. The trends in the CRPS were the same as for the correlation, with HS having the lowest CRPS. Again, we note that using genomic data improved the CRPS compared to using

pedigree, and in most cases average CRPS for the EBVs were lower than for the PBVs.

Table 2: Average correlation over 60 independent replications for the different breeding scenarios, using pedigree or genomic markers, and for both estimated breeding values (EBV) and predicted breeding values (PBV). The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.33	0.28	0.32	0.18	0.32	0.20
H	0.36	0.29	0.41	0.22	0.42	0.25
S	0.52	0.50	0.56	0.34	0.55	0.35
HS	0.54	0.52	0.58	0.36	0.57	0.37
HSC	0.57	0.55	0.59	0.36	0.58	0.37
Genomic markers						
G	0.33	0.32	0.40	0.29	0.42	0.32
H	0.36	0.33	0.51	0.38	0.59	0.46
S	0.58	0.56	0.70	0.54	0.72	0.57
HS	0.63	0.60	0.74	0.57	0.75	0.60
HSC	0.64	0.62	0.74	0.58	0.75	0.60

3.1.2 Including spatial covariates

The spatial covariates did not improve the results remarkably when a spatial model was already included. In the correlation results in Table 2 and the CRPS results in Table 3, results are included for model HSC, the model including spatial covariates to the herd and spatial effects. Both the correlation and CRPS were only marginally better for the HSC model compared to the HS model in some cases, and in the remaining cases they were comparable. Because of this we have focused on the "cheaper" models and not included model HSC in the remaining results from the simulation study. Some additional results with model HSC are presented in the additional results (see Section 6.1).

Table 3: Average CRPS over 60 independent replications for the different breeding scenarios, using pedigree or genomic markers, and for both estimated breeding values (EBV) and predicted breeding values (PBV). The standard error for all values had order of magnitude 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.54	0.43	0.65	0.40	0.70	0.37
H	0.41	0.37	0.34	0.28	0.33	0.25
S	0.17	0.17	0.17	0.18	0.18	0.18
HS	0.16	0.16	0.17	0.18	0.18	0.18
HSC	0.16	0.16	0.16	0.18	0.17	0.18
Genomic markers						
G	0.39	0.39	0.32	0.30	0.30	0.26
H	0.36	0.37	0.22	0.22	0.18	0.18
S	0.15	0.15	0.13	0.15	0.13	0.15
HS	0.14	0.15	0.12	0.15	0.12	0.14
HSC	0.14	0.14	0.12	0.15	0.12	0.14

3.1.3 Separating genetic and spatial components

The models without spatial effects (G,H) were not able to separate the genetic and spatial components. In Table 4 we present the average correlation between the EBVs from all models and the true spatial effects of herd locations in each breeding strategy. This shows that the EBVs from models G and H were correlated with the spatial effects, and suggests that the genetic effect in G and H captured parts of the spatial components of the simulated phenotype. The correlations from models S and HS were closer to zero, suggesting that these models were better able to separate genetic and spatial effects. This, together with the correlation results in Table 2 and CRPS results in Table 3, suggests that the herd effect alone was not sufficient to account for all environmental effects in LMIC breeding systems, and that the EBVs from models G and H have captured parts of the spatially dependent effects in the genetic effect.

Table 4: Average correlation between EBVs and true spatial effects for the different breeding strategies, using pedigree or genomic markers. The standard error for all values had order of magnitude 10^{-3}

Strategy	A	B	C
Pedigree			
G	0.68	0.64	0.64
H	0.70	0.60	0.58
S	0.11	0.06	0.06
HS	0.12	0.06	0.06
Genomic markers			
G	0.84	0.74	0.69
H	0.83	0.63	0.50
S	0.16	0.05	0.04
HS	0.21	0.05	0.04

3.1.4 Comparing breeding strategies and genetic models

The benefit of including a spatial model was largest when the spatial and genetic effects were hard to separate. In Figure 2 we have plotted the relative improvement in average correlation and CRPS between true breeding values and EBVs/PBVs from model H to model HS, for the different breeding strategies. With both the genomic marker based and the pedigree based models, the improvement was largest with strategy A (about 50% to 80%), second largest with strategy B (about 35% to 65%), and third largest with strategy C (about 20% to 45%). These strategies correspond to strongly confounded genetic and spatial effects, to separable genetic and spatial effects. With breeding strategy A there was not much difference in improvement between models using genomic markers or pedigree, whereas there was a tendency in breeding strategy B and C that the improvement was largest with the pedigree based models. This is because with genomic markers, genetic relationship between individuals not related by pedigree is captured and helps to separate genetic and environmental effects.

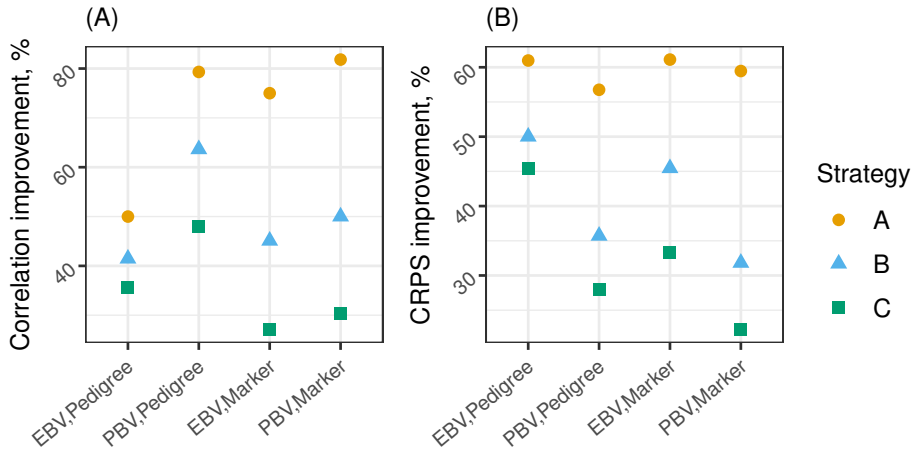


Figure 2: Percentage relative improvement in EBV correlation (A) and CRPS (B) going from model H to HS for the different breeding strategies, using pedigree or genomic markers

3.1.5 Changing proportion of spatial variance

Including a spatial model to the random herd effect, even when there was no spatial variation in the phenotype, did not decrease the prediction accuracy. We varied the proportion of variance in the phenotype due to herd and spatial effects, and in Figure 3 we present the average correlation and CRPS for EBVs estimated with genomic markers in breeding strategy B. The x-axis goes from all herd effect variance to all spatial effect variance relative to the total herd and spatial variance. For models G and H, the average correlation and CRPS became worse as the proportion of spatial variance increased, whereas for models S and HS the average correlation and CRPS became better. Overall, model HS had the best correlation and CRPS for all spatial variance proportions. It was equal to model H when there was no spatial variation, and equal to model S when there was no herd effect variation.

From the results so far we have seen that model S had better correlation and CRPS than model H. However, this is not always the case. When most of the environmental variation was caused by herd effects rather than spatial effects, model H gave better estimates than model S.

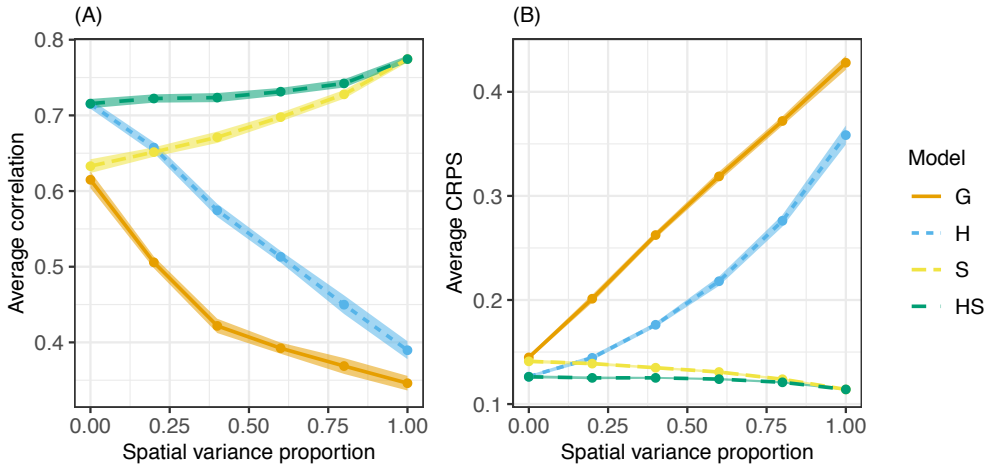


Figure 3: Average correlation (A) and CRPS (B) with 95% confidence intervals for EBVs with breeding strategy B for genetic models based on genomic markers with varying spatial variance proportion

The same tendencies were seen for the PBVs for both genomic marker based models and pedigree based models, and in the other breeding strategies, which can be seen in tables presented in the additional results (see Section 6.1).

3.1.6 Changing the herd clustering

When spatial and genetic effects were difficult to separate the prediction accuracy improved when herds were weakly clustered rather than closely clustered. When simulating the data we varied the distribution of herd locations, from more closely clustered around each village (with herd location variance $1 \cdot 10^{-4}$) to less closely clustered around each village (with herd location variance $9 \cdot 10^{-4}$). In Figure 4 we present the average correlation and CRPS for EBVs estimated using genomic markers with breeding strategy A for the three intensities of clustering. The figure shows that as herds were less clustered, the correlation increased and the CRPS decreased across all models.

The same trend appeared for the PBVs and from the models using pedigree, but not with breeding strategy B and C, where the genetic and

spatial effects were less confounded. Tables showing the correlation and CRPS between true breeding values and EBVs or PBVs and correlation between EBVs and true spatial effects for all breeding strategies and herd clustering are given in the additional results (see Section 6.1).

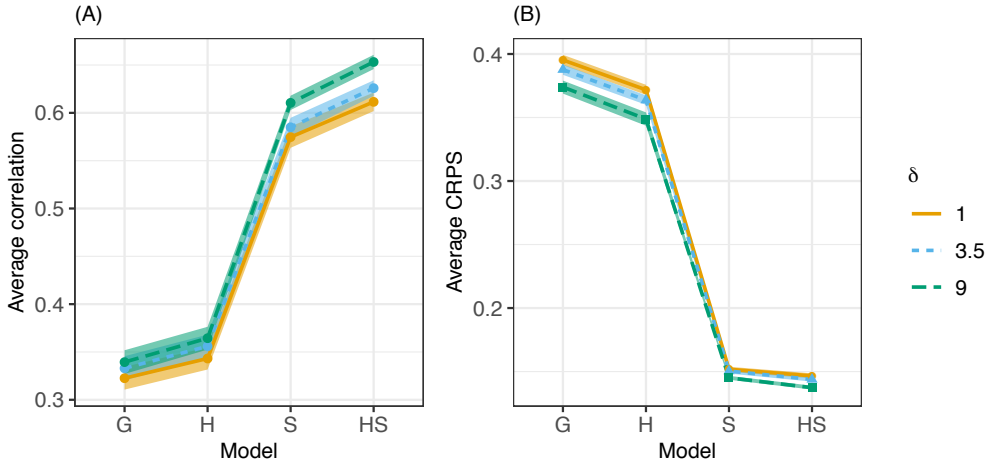


Figure 4: Average correlation (A) and CRPS (B) with 95% confidence intervals for different models, with different values of herd location variance $\delta \cdot 10^{-4}$, using breeding strategy A for EBVs and genomic marker models

3.2 Case study: Brown-Swiss cattle

In this section we present the results from fitting the models to the subset of the Brown-Swiss cattle data with 3800 individuals. We present the posterior distributions of the hyper-parameters, the DIC, the estimated spatial field from model HS, and compare the EBVs from models H and HS. The corresponding results for the full Brown-Swiss cattle data set are presented in the additional results (see Section 6.2).

We had four models, where model G included a common intercept, a genetic effect estimated with pedigree, additional fixed categorical effects and a residual effect. Model H included a herd effect in addition, model S included a spatial effect in addition, and model HS included both a herd effect and a spatial effect in addition.

In general the results showed that (i) models H and HS explained most

of the variation in the data and had the best fit according to the DIC, (ii) the data had a spatially dependent structure captured by models S and HS, and (iii) the two models with the best fit according to the DIC, models H and HS, separated the genetic and environmental effects differently. We go through each of these points in detail below.

3.2.1 Explained variation and model fit

Models H and HS explained most of the variation in the data and had the best fit according to the DIC. In Figure 5 the posterior distributions for the hyperparameters in the models are presented. The figure has five panels showing the additive genetic variance σ_a^2, the residual variance σ_e^2, the herd effect variance σ_h^2, the spatial variance σ_s^2, and the spatial range ρ in kilometers.

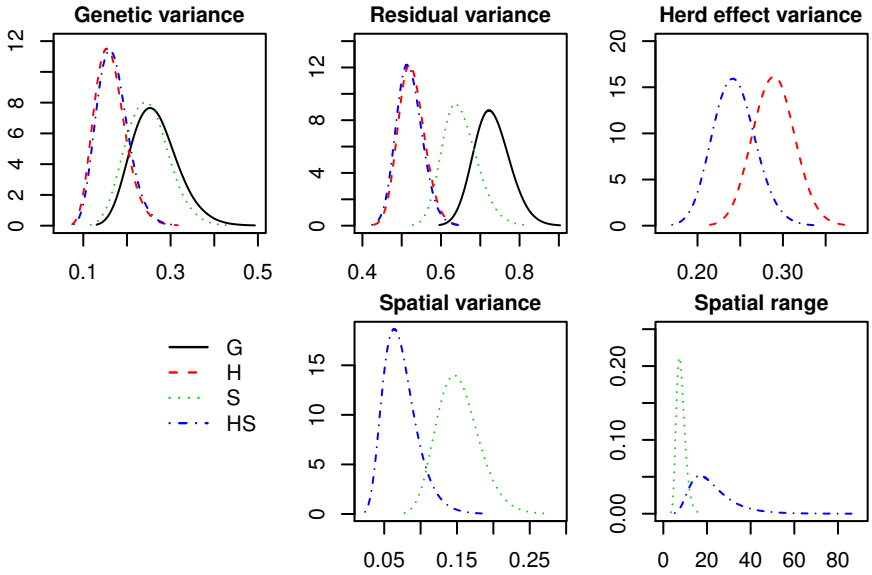


Figure 5: Posterior distributions of hyper-parameters from models G, H, S and HS applied to the case study data

The posterior additive genetic variance was similar between models H and HS, higher in model S, and even higher in model G. The same tendency was seen for the posterior residual variance. The posterior herd effect variance was lower in model HS than model H, which was reasonable

since the herd effect in model H captured the spatial component in the phenotype, which model HS assigned to the spatial effect. The posterior spatial variance in model S was higher than in model HS since model S captured herd effects. Finally, the posterior spatial range was lower in model S than in model HS, since model S captured herd effects in the spatial effects which means shorter range of dependency between spatial locations. The mean posterior range from model HS indicated that herds more than 22 km apart had independent (large scale) environments.

Since model G cannot explain variation due to herd or other environmental effects, it was reasonable to assume that some of the posterior genetic effect in G was actually due to confounding with other effects arising from few individuals per herd. This explains the high posterior additive genetic variance from model G. A similar reasoning could be used for model S, which had to assign variation due to herd effects, either to the genetic effects, the residual effects or the spatial effects. From Figure 5 it can seem that the variation from herd effects was distributed to all other effects, which explains why the posterior additive genetic variance and posterior residual variance was higher in model S than models H and HS, and why the posterior spatial variance was higher than in model HS. It seems that models H and HS distributed variation similarly except for the herd effect which is expected to be higher in H than in HS.

In Table 5 the DIC for each of the models are presented. The table indicates that model HS had the best fit, followed by model H, then model S and finally model G. These numbers are in line with the posterior hyperparameters, where we saw that model H and HS could explain most of the variation in the phenotype. Although model S has the potential to explain much variation as well, it is forced to assign herd effects either to genetic effects, spatially dependent effects, or residual effects. We saw from the results with the simulated data that model S gave worse predictions that model H when most of the environmental variation was due to herd effects. This seems to be the case here considering the low posterior spatial variance. Finally, model G was not able to separate the genetic and environmental effects, which lead to a poor model fit. A rule of thumb, is that a complex model should be preferred over a less complex model if the DIC is reduced with more than ten units. When it comes to choosing between models H and HS, model HS should be preferred, as its DIC was 36 units smaller.

Table 5: DIC for models G, H, S and HS applied to the case study

Model	DIC
G	10494
H	9795
S	10233
HS	9759

3.2.2 The estimated spatial effects

The data had a spatially dependent structure captured by models S and HS. The posterior spatial field from model HS is shown in Figure 6. The figure shows both the mean, in panel (A), and the standard deviation, in panel (B), of the posterior distribution. The axes show coordinates in the Transverse Mercator system in kilometers.

In the western part of Slovenia model HS suggested two environmental regions with mean different from zero, one with positive effect, and one with negative effect. In the central part of Slovenia, there were several smaller regions with either positive or negative environmental effect. In the north east there were not many observations, so there was only a small region of positive effect, and zero effects otherwise. The standard deviation was lowest where we had observations, ranging between 0.1 and 0.2 in these areas, and was highest where there were no observations, ranging between 0.2 and 0.3 in these areas.

3.2.3 Comparing breeding values from models H and HS

The two models with the best fit, models H and HS, separated the genetic and environmental effects differently. The DIC in Table 5 and the posterior hyperparameters in Figure 5, indicated that models H and HS had the best model fit and a similar decomposition of the genetic and environmental effects. Furthermore, the EBVs from models H and HS were highly correlated, with a correlation of about 0.995.

To evaluate how well the models separated the genetic and environmental effects, we computed the correlation of the EBVs from both model H and model HS with the mean posterior spatial effects from model HS. For model H this was about 0.14, whereas for model HS it was about 0.07.

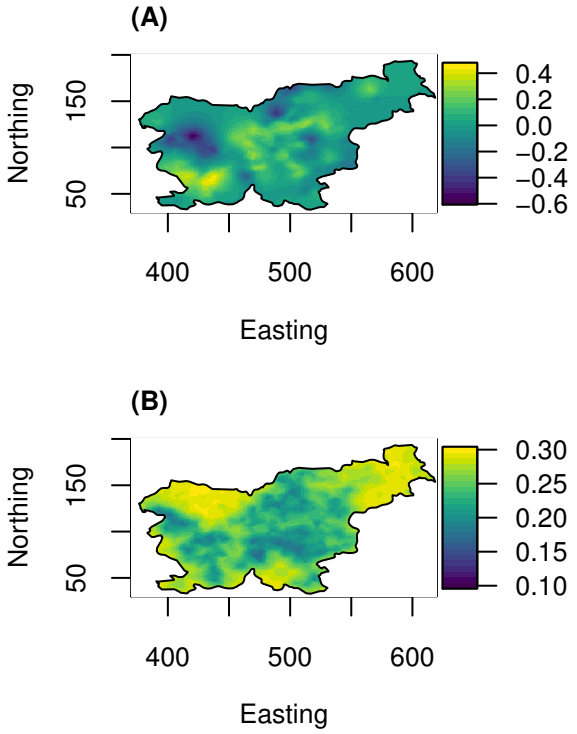


Figure 6: Mean (A) and standard deviation (B) of the posterior spatial effect from model HS. The axis units are in km

This could suggest that there were some effects that were assigned to be spatial in model HS, but assigned to be genetic in model H.

In Figure 7 we present the difference in EBVs between models H and HS as boxplots according to the mean posterior spatial effects from model HS. This shows that the difference in EBVs was correlated with the spatial effect from HS. When the posterior spatial effects were negative, the posterior genetic effect from model H was smaller than in model HS, and when the posterior spatial effect was positive the genetic effect from model H was larger than from model HS. The figure also shows how many cows were used for each boxplot, and shows that for many of the cows, living in areas not strongly affected by spatial effects, the difference in EBV was not large.

The correlation between the EBV difference and the posterior spatial

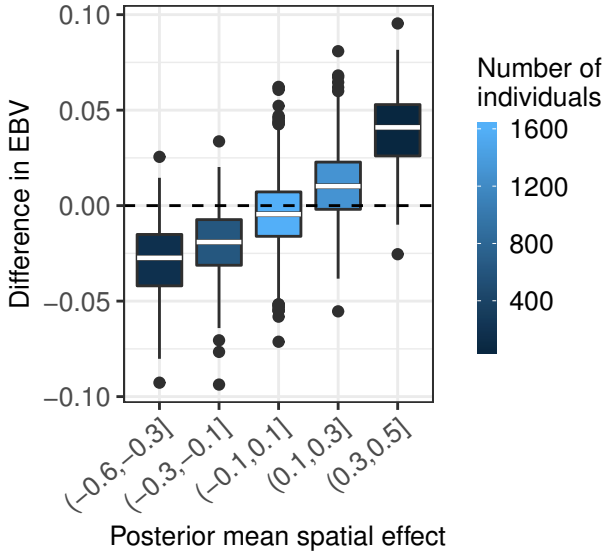


Figure 7: The difference in EBV between models H and HS against the mean posterior spatial effect from model HS

effect from model HS was about 0.62. This is in line with what was seen from the simulation results and suggests that although the two models had highly correlated EBVs, there were differences between the EBVs due to model H not separating the environmental and genetic effects as well as model HS in the LMIC context.

4 Discussion

Our results highlight three main points for discussion, specifically: the improvement in EBVs and PBVs from modeling spatial variation, the limitations of the study, and the future possibilities for improving genetic evaluation in LMIC animal breeding programs.

4.1 The improvement from modeling spatial variation

Our simulations showed that including a spatial component in the models for genetic evaluation can improve the correlation and CRPS between

true breeding values and EBVs or PBVs. The improvement in correlation and CRPS was largest when the genetic and environmental effects were strongly confounded, a pedigree was used to model genetic relationships rather than genomic markers, and when most of the environmental variation was spatially correlated. We also saw that the EBVs from models without spatial effects were correlated with the true spatial effects, whereas the EBVs from models with spatial effects were much less correlated with the true spatial effects. This suggested that the models without spatial effect did a poorer separation of genetic and environmental effects than the models with spatial effect.

From changing the proportion of variation due to spatial and herd effects, we saw that including a spatial effect to the herd effect did not give worse correlation and CRPS than only having a herd effect, even if there was no spatial variation in the observations. However, the model with only spatial effect as environmental effect (S) did worse than the model with only herd effect as environmental effect (H), when only a small part of the environmental variation was due to spatial variation. This means that excluding the random herd effect in favor of a spatial effect is not recommended, but including both a herd effect and a spatial effect is the recommended model choice.

The model that included spatial covariates (HSC) did not improve the EBVs or PBVs remarkably compared to the model with herd and spatial effects (HS). This was because the spatial covariates all explained spatial variation, which the spatial model was able to capture very well.

Among the three breeding strategies, A, B and C, the most realistic was breeding strategy B, which assumed that most of the mating was performed using the same bull for all herds belonging to the same village, and 25% of farmers randomly chosen used AI. This is similar to what has been reported in surveys. For example, in a survey for Kenyan farmers, 87% reported that they used bull services, and the remaining 13% used AI (Lawrence et al., 2015).

From the case study we saw that the four models assigned variation in the observations to the model components differently, because they had different model components. However, they all had the genetic component in common. The two models without herd effect (G, S) assigned higher genetic variance than the two models with herd effect (H, HS), indicating that environmental effects were not separated from the genetic effects.

We also saw that even though the two models with herd effect (H, HS) had highly correlated EBVs, there were differences between the EBVs and these differences were correlated with the mean posterior spatial effects. Following the results from the simulation study, it is reasonable to assume that the model with both herd and spatial effect (HS) had the best separation of the genetic and environmental effects also in the case study.

4.2 The limitations of the study

The chosen model for simulating the data did not include all factors that would emerge in a LMIC breeding program. Among the simplifications that were made were: the absence of non-additive genetic effects in the true breeding values, absence of geneotype-by-environment interaction, absence of errors in the pedigree and genotype observations, and considering only a single trait and breed. The animals were initially distributed to herds by random, and the farms using AI were also chosen by random.

The simplifications are likely to yield better correlation and CRPS results than what could be expected in a real LMIC breeding program, but the case study largely corroborates the main conclusions from the simulation study. Future studies could for example consider a non-random distribution of animals to herds and have farms using AI chosen non-randomly. These non-random associations are realistic since well-resourced farmers are more likely to use AI than farmers constrained by infrastructural challenges (Schaeffer, 2018).

We tested how the models responded to changing the variation caused by spatial and herd effects, but we did not test how the models responded to increasing or decreasing the genetic variance or the residual variance. However, we used reasonable values that were based on conversations with geneticists.

The case study was sampled from a larger data set to imitate the data structure of LMIC smallholder breeding systems with few individuals per herd. The results would be slightly different had a different subset been used, but the conclusions would likely be the same since it is the data structure which makes it hard to separate genetic and environmental effects.

Genotype-by-environment interactions have been modeled in several studies (Strandberg et al., 2009; Hayes et al., 2009; Tiezzi et al., 2017; Yao et al., 2017; Schultz and Weigel, 2019), but was not considered here, since

it would likely lead to over-fitting the models when the herd sizes were as small as in this case.

Finally, since the INLA method implements full Bayesian analysis, prior distributions had to be assigned to all model parameters, which is not always straightforward. However, setting a prior based on the knowledge about the process is likely to improve the inference. We used penalized complexity priors (Simpson et al., 2017) since these penalize model complexity to avoid over-fitting, with parameters chosen based on prior knowledge about the relative importance of the different effects in the models.

4.3 Future possibilities for improving genetic evaluation in LMIC animal breeding

In this study we have shown that including a spatial effect in the models for genomic evaluation in LMIC smallholder breeding systems can improve the EBVs and PBVs. We have also shown that spatial covariates do not make a remarkable improvement in the EBVs and PBVs when a spatial model is included. This result could be important for people considering to collect spatial covariates or not, and resources could be spent elsewhere when a spatial model is sufficient to capture the environmental effects.

In order for genetic evaluation to be carried out, targeted phenotyping and genotyping of animals in the smallholder farms is necessary. The use of genomic markers over pedigree will yield higher accuracy than pedigree (Powell et al., 2019), can be acquired for animals faster than building a pedigree from scratch, and can be made cost efficient by the usage of genotype imputation (Aliloo et al., 2018). More widespread use of AI, can improve genetic connectedness between herds across large distances, but is still hindered by infrastructural challenges and higher costs than natural bull services (Lawrence et al., 2015). Finally, when chosing herds to be part of a larger cattle breeding program, farms should be chosen from different areas, not clustered in the same area to make the separation of genetic and environmental effects as good as possible.

5 Conclusions

With this study we have demonstrated the potential of spatial modeling to improve genetic evaluation in LMIC smallholder dairy production systems by enhancing the separation of the genetic and environmental effects beyond using a fixed or random independent herd effect. This has been shown for three different breeding strategies, and for different proportions of spatial variation and clustering of herds, with both pedigree and genomic marker based genetic relationship matrices.

The inclusion of a spatial model in addition to a random herd effect did not perform worse than a model with only a random herd effect even when there was no spatial effect in the observed phenotype. Further, the inclusion of spatial covariates did not improve results remarkably when a spatial model was included.

Bibliography

Aliloo, H., Mrode, R., Okeyo, A., Ni, G., Goddard, M., and Gibson, J. (2018). The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of east africa. *Journal of dairy science*, 101(10):9108–9127.

Baltenweck, I., Ouma, R., Anunda, F., Okeyo Mwai, A., and Romney, D. (2004). Artificial or natural insemination: The demand for breeding services by smallholders. *Proceedings of 9th KARI Biennial Scientific Conference and Research week. 8th to 12th November 2004, Nairobi Kenya*.

Bebe, B. O., Udo, H. M., Rowlands, G. J., and Thorpe, W. (2003). Smallholder dairy systems in the Kenya highlands: Breed preferences and breeding practices. *Livestock Production Science*, 82(2-3):117–127.

Chen, G. K., Marjoram, P., and Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. *Genome research*, 19(1):136–142.

Dekkers, J. C. and Hospital, F. (2002). Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 3(1):22.

Faux, A.-M., Gorjanc, G., Gaynor, R. C., Battagin, M., Edwards, S. M., Wilson, D. L., Hearne, S. J., Gonen, S., and Hickey, J. M. (2016). AlphaSim: Software for breeding program simulation. *The plant genome*, 9(3).

Frey, M., Hofer, A., and Künzi, N. (1997). Comparison of models with a fixed or a random contemporary group effect for the genetic evaluation for litter size in pigs. *Livestock Production Science*, 48(2):135–141.

Gaynor, R. C., Gorjanc, G., Wilson, D., Money, D., and Hickey, J. M. (2019). *AlphaSimR: Breeding Program Simulations*. R package version 0.9.0.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Savin, K., Van Tassell, C. P., Sonstegard, T. S., and Goddard, M. E. (2009). A validated genome wide association study to breed cattle adapted to an environment altered by climate change. *PloS one*, 4(8):e6676.

Jorjani, H., Philipsson, J., and Mocquot, J. (2001). Interbull guidelines for national and international genetic evaluation systems in dairy cattle with focus on production traits. *Interbull Bulletin*, 28.

Lawrence, F., Mutembei, H., Lagat, J., Mburu, J., Amimo, J., Okeyo, A., et al. (2015). Constraints to use of breeding services in Kenya. *International Journal of Veterinary Science*, 4(4):211–215.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Lynch, M., Walsh, B., et al. (1998). *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, MA.

Majiwa, E., Murage, H., and Kavoi, M. (2017). Smallholder dairying in Kenya: The assessment of the technical efficiency using the stochastic production frontier model. 14(2).

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83.

Matérn, B. (1960). Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsintitut*, 49.

Ojango, J. M., Mrode, R., Rege, J., Mujibi, D., Strucken, E., Gibson, J., and Mwai, O. (2019). Genetic evaluation of test-day milk yields from smallholder dairy production systems in Kenya using genomic relationships. *Journal of dairy science*, 102(6):5266–5278.

Pereira, R., Schenkel, F., Ventura, R., Ayres, D., El Faro, L., Machado, C., and Albuquerque, L. (2019). Contemporary group alternatives for genetic evaluation of milk yield in small populations of dairy cattle. *Animal Production Science*, 59(6):1022–1030.

Philipsson, J., Zonabend, E., Bett, R., and Okeyo, A. (2011). Global perspectives on animal genetic resources for sustainable agriculture and food production in the tropics. Technical report, Animal Genetics Training Resource, version 3, International Livestock Research Institute, Nairobi, Kenya, and Swedish University of Agricultural Sciences, Uppsala, Sweden.

Powell, O., Mrode, R., Gaynor, R., Johnsson, M., Gorjanc, G., and Hickey, J. (2019). Genomic data enables genetic evaluation using data recorded on LMIC smallholder dairy farms. Preprint. https://doi.org/10.1101/827956, accessed Nov 5th 2019.

Rademaker, C. J., Bebe, B. O., van der Lee, J., Kilelu, C., and Tonui, C. (2016). Sustainable growth of the Kenyan dairy sector: A quick scan of robustness, reliability and resilience. Technical report, Wageningen University & Research, Wageningen Livestock Research.

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(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4:395–421.

Rue, H. v. and Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. Chapman and Hall/CRC.

Sæbø, S. and Frigessi, A. (2004). A genetic and spatial Bayesian analysis of mastitis resistance. *Genetics Selection Evolution*, 36(5):527.

Schaeffer, L. (2018). Necessary changes to improve animal models. *Journal of Animal Breeding and Genetics*, 135(2):124–131.

Schultz, N. and Weigel, K. (2019). Inclusion of herdmate data improves genomic prediction for milk-production and feed-efficiency traits within North American dairy herds. *Journal of dairy science*, 102(12):11081–11091.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Strandberg, E., Brotherstone, S., Wall, E., and Coffey, M. (2009). Genotype by environment interaction for first-lactation female fertility traits in UK dairy cattle. *Journal of dairy science*, 92(7):3437–3446.

Tiezzi, F., de Los Campos, G., Gaddis, K. P., and Maltecca, C. (2017). Genotype by environment (climate) interaction improves genomic prediction for production traits in US Holstein cattle. *Journal of dairy science*, 100(3):2042–2056.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91(11):4414–4423.

Visscher, P. and Goddard, M. (1993). Fixed and random contemporary groups. *Journal of dairy science*, 76(5):1444–1454.

Weigel, K., VanRaden, P., Norman, H., and Grosu, H. (2017). A 100-year review: Methods and impact of genetic selection in dairy cattle—from daughter-dam comparisons to deep learning algorithms. *Journal of dairy science*, 100(12):10234–10250.

Yao, C., De Los Campos, G., VandeHaar, M., Spurlock, D., Armentano, L., Coffey, M., De Haas, Y., Veerkamp, R., Staples, C., Connor, E., et al. (2017). Use of genotype× environment interaction model to accommodate genetic heterogeneity for residual feed intake, dry matter intake, net energy in milk, and metabolic body weight in dairy cattle. *Journal of dairy science*, 100(3):2007–2016.

6 Additional results

6.1 Simulation study

Changing proportion of spatial variance

Here we show the average correlation and CRPS between true breeding value and EBV or PBV in all breeding strategies, using both pedigree and genomic markers, when varying the proportion of spatial variance relative to the sum of spatial variance and herd effect variance. The herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $3.5 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$.

Table 6 and Table 7 show the correlation and CRPS respectively for breeding strategy A. Table 8 and Table 9 show the correlation and CRPS respectively for breeding strategy B. Table 10 and Table 11 show the correlation and CRPS respectively for breeding strategy C.

Changing the herd clustering

Table 12 and Table 13 show average correlation and CRPS respectively between true breeding value and EBV/PBV in all breeding strategies, using both pedigree and genomic markers, when the herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $1 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$.

Table 14 and Table 15 show average correlation and CRPS respectively between true breeding value and EBV/PBV in all breeding strategies, using both pedigree and genomic markers, when the herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $9 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$.

Correlation between true spatial effect and EBV with changing herd clustering

Table 16 shows the average correlation between the EBV and the true spatial effect in all breeding strategies, using both pedigree and genomic markers, when the herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $1 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$.

Table 17 shows the average correlation between the EBV and the true spatial effect in all breeding strategies, using both pedigree and genomic markers, when the herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $3.5 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$. This is an extended table from the main results.

Table 18 shows the average correlation between the EBV and the true spatial effect in all breeding strategies, using both pedigree and genomic markers, when the herd locations were simulated from a bivariate normal distribution with mean equal to the village locations, and variance $9 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$.

Table 6: Average correlation for EBV and PBV in breeding strategy A, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.53	0.40	0.35	0.34	0.32	0.31	0.54	0.38	0.32	0.29	0.26	0.23
H	0.57	0.46	0.39	0.37	0.34	0.32	0.57	0.42	0.34	0.31	0.27	0.24
S	0.51	0.48	0.49	0.52	0.56	0.60	0.51	0.46	0.47	0.50	0.53	0.58
HS	0.57	0.53	0.53	0.54	0.56	0.60	0.56	0.51	0.51	0.52	0.53	0.58
Genomic markers												
G	0.54	0.40	0.36	0.34	0.31	0.32	0.55	0.38	0.34	0.31	0.27	0.30
H	0.64	0.47	0.40	0.36	0.33	0.33	0.62	0.43	0.37	0.33	0.28	0.30
S	0.53	0.51	0.54	0.57	0.63	0.70	0.53	0.50	0.53	0.55	0.60	0.67
HS	0.64	0.60	0.61	0.62	0.64	0.70	0.62	0.58	0.58	0.58	0.61	0.67

Table 7: Average CRPS for EBV and PBV in breeding strategy A, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for all values had order of magnitude 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.164	0.271	0.395	0.552	0.688	0.730	0.164	0.255	0.347	0.429	0.489	0.542
H	0.155	0.237	0.332	0.406	0.492	0.618	0.155	0.213	0.289	0.359	0.431	0.519
S	0.161	0.170	0.171	0.170	0.168	0.165	0.162	0.155	0.155	0.152	0.149	0.142
HS	0.155	0.164	0.167	0.166	0.166	0.164	0.155	0.146	0.148	0.147	0.147	0.142
Genomic markers												
G	0.158	0.242	0.317	0.385	0.452	0.514	0.158	0.237	0.313	0.384	0.450	0.524
H	0.139	0.214	0.287	0.359	0.436	0.508	0.142	0.209	0.284	0.357	0.433	0.517
S	0.154	0.160	0.157	0.153	0.144	0.129	0.155	0.141	0.142	0.138	0.131	0.116
HS	0.139	0.148	0.147	0.146	0.140	0.129	0.142	0.134	0.135	0.133	0.127	0.116

Table 8: Average correlation for EBV and PBV in breeding strategy B, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.51	0.40	0.35	0.33	0.32	0.32	0.29	0.22	0.18	0.17	0.17	0.16
H	0.61	0.51	0.44	0.39	0.36	0.33	0.36	0.29	0.24	0.21	0.18	0.16
S	0.53	0.53	0.55	0.59	0.61	0.65	0.31	0.32	0.33	0.35	0.37	0.41
HS	0.61	0.60	0.60	0.62	0.62	0.65	0.37	0.37	0.37	0.38	0.38	0.41
Genomic markers												
G	0.61	0.51	0.42	0.39	0.37	0.35	0.46	0.40	0.33	0.26	0.26	0.26
H	0.72	0.66	0.57	0.51	0.45	0.39	0.56	0.52	0.44	0.36	0.33	0.28
S	0.63	0.65	0.67	0.70	0.73	0.77	0.47	0.52	0.54	0.54	0.57	0.60
HS	0.72	0.72	0.72	0.73	0.74	0.77	0.56	0.58	0.58	0.57	0.58	0.60

Table 9: Average CRPS for EBV and PBV in breeding strategy B, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for all values had order of magnitude 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.165	0.268	0.396	0.536	0.685	0.709	0.181	0.233	0.306	0.364	0.411	0.431
H	0.153	0.211	0.296	0.370	0.455	0.578	0.174	0.189	0.239	0.284	0.335	0.398
S	0.160	0.167	0.164	0.159	0.159	0.153	0.178	0.160	0.158	0.156	0.154	0.150
HS	0.153	0.161	0.159	0.157	0.158	0.153	0.173	0.156	0.154	0.154	0.153	0.150
Genomic markers												
G	0.145	0.201	0.262	0.319	0.372	0.428	0.162	0.187	0.235	0.285	0.322	0.364
H	0.126	0.144	0.176	0.218	0.276	0.358	0.148	0.141	0.165	0.203	0.244	0.311
S	0.141	0.139	0.135	0.131	0.124	0.114	0.159	0.134	0.129	0.130	0.125	0.120
HS	0.126	0.125	0.125	0.124	0.121	0.114	0.148	0.125	0.123	0.125	0.122	0.120

Table 10: Average correlation for EBV and PBV in breeding strategy C, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.49	0.38	0.32	0.32	0.31	0.32	0.30	0.24	0.20	0.19	0.18	0.17
H	0.57	0.50	0.44	0.41	0.38	0.35	0.36	0.32	0.27	0.25	0.22	0.19
S	0.51	0.52	0.53	0.55	0.57	0.61	0.31	0.33	0.34	0.35	0.36	0.39
HS	0.57	0.57	0.57	0.57	0.58	0.61	0.36	0.37	0.37	0.37	0.37	0.39
Genomic markers												
G	0.65	0.54	0.49	0.44	0.40	0.38	0.52	0.41	0.38	0.35	0.27	0.29
H	0.74	0.69	0.66	0.60	0.55	0.49	0.60	0.54	0.52	0.48	0.39	0.38
S	0.67	0.67	0.70	0.72	0.75	0.79	0.53	0.53	0.56	0.58	0.59	0.63
HS	0.74	0.74	0.75	0.75	0.76	0.79	0.60	0.59	0.61	0.60	0.60	0.63

Table 11: Average CRPS for EBV and PBV in breeding strategy C, using pedigree or genomic markers, with varying proportion of spatial variance. The standard error for all values had order of magnitude 10^{-3}

	EBV						PBV					
$\sigma_s^2/(\sigma_s^2+\sigma_h^2)$	0	0.2	0.4	0.6	0.8	1	0	0.2	0.4	0.6	0.8	1
Pedigree												
G	0.180	0.329	0.555	0.701	0.720	0.731	0.187	0.247	0.331	0.363	0.375	0.389
H	0.171	0.234	0.288	0.336	0.383	0.499	0.179	0.189	0.217	0.239	0.267	0.324
S	0.174	0.179	0.178	0.180	0.182	0.180	0.184	0.165	0.164	0.165	0.164	0.159
HS	0.170	0.178	0.175	0.177	0.181	0.179	0.179	0.162	0.160	0.161	0.162	0.159
Genomic markers												
G	0.139	0.188	0.237	0.289	0.337	0.394	0.155	0.173	0.211	0.243	0.289	0.313
H	0.122	0.134	0.149	0.176	0.205	0.264	0.143	0.132	0.142	0.160	0.188	0.221
S	0.136	0.135	0.130	0.126	0.120	0.111	0.154	0.130	0.127	0.125	0.122	0.115
HS	0.122	0.122	0.120	0.120	0.117	0.111	0.143	0.122	0.120	0.122	0.120	0.115

Table 12: Average correlation for the different breeding strategies for EBV and PBV, using pedigree or genomic markers, and the herd locations simulated using variance $1 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$. The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.32	0.27	0.32	0.18	0.32	0.19
H	0.35	0.29	0.41	0.22	0.41	0.25
S	0.51	0.48	0.56	0.34	0.55	0.35
HS	0.53	0.50	0.58	0.36	0.57	0.37
HSC	0.56	0.54	0.59	0.37	0.58	0.38
Genomic markers						
G	0.32	0.30	0.40	0.29	0.42	0.32
H	0.34	0.32	0.51	0.38	0.58	0.46
S	0.57	0.55	0.70	0.54	0.72	0.57
HS	0.61	0.58	0.73	0.58	0.75	0.60
HSC	0.63	0.60	0.74	0.59	0.75	0.61

Table 13: Average CRPS for the different breeding strategies for EBV and PBV, using pedigree or genomic markers, and the herd locations simulated using variance $1 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$. The standard error for all values had order of magnitude 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.559	0.438	0.667	0.406	0.706	0.371
H	0.419	0.374	0.343	0.281	0.335	0.252
S	0.168	0.168	0.166	0.180	0.180	0.183
HS	0.165	0.164	0.166	0.178	0.179	0.181
HSC	0.160	0.159	0.163	0.176	0.176	0.179
Genomic markers						
G	0.395	0.402	0.325	0.302	0.299	0.264
H	0.372	0.378	0.225	0.222	0.180	0.181
S	0.152	0.156	0.130	0.151	0.126	0.146
HS	0.147	0.151	0.124	0.146	0.120	0.142
HSC	0.143	0.147	0.123	0.145	0.120	0.142

Table 14: Average correlation for the different breeding strategies for EBV and PBV, using pedigree or genomic markers, and the herd locations simulated using variance $9 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$. The standard error for some values had order of magnitude 10^{-2}, and most had 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.33	0.29	0.32	0.17	0.32	0.19
H	0.37	0.31	0.41	0.22	0.42	0.25
S	0.55	0.54	0.56	0.34	0.55	0.35
HS	0.56	0.55	0.59	0.37	0.57	0.37
HSC	0.58	0.57	0.59	0.37	0.58	0.37
Genomic markers						
G	0.34	0.31	0.40	0.27	0.44	0.33
H	0.36	0.33	0.53	0.38	0.61	0.47
S	0.61	0.60	0.70	0.54	0.72	0.57
HS	0.65	0.63	0.74	0.57	0.75	0.59
HSC	0.67	0.65	0.74	0.58	0.75	0.60

Table 15: Average CRPS for the different breeding strategies for EBV and PBV, using pedigree or genomic markers, and the herd locations simulated using variance $9 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$. The standard error for all values had order of magnitude 10^{-3}

	Strategy A		Strategy B		Strategy C	
	EBV	PBV	EBV	PBV	EBV	PBV
Pedigree						
G	0.500	0.410	0.615	0.392	0.688	0.370
H	0.393	0.348	0.326	0.270	0.325	0.248
S	0.164	0.163	0.165	0.179	0.180	0.184
HS	0.160	0.158	0.163	0.176	0.178	0.181
HSC	0.156	0.155	0.161	0.175	0.177	0.180
Genomic markers						
G	0.374	0.387	0.308	0.289	0.282	0.253
H	0.349	0.362	0.209	0.211	0.169	0.174
S	0.145	0.147	0.130	0.152	0.126	0.147
HS	0.137	0.141	0.123	0.146	0.119	0.143
HSC	0.135	0.138	0.122	0.146	0.119	0.143

Table 16: Average correlation between EBV and true spatial effect in all breeding strategies, using pedigree or genomic markers, and the herd locations simulated using variance $1 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$. The standard error for all values had order of magnitude 10^{-3}

Strategy	A	B	C
Pedigree			
G	0.87	0.76	0.70
H	0.86	0.65	0.51
S	0.23	0.06	0.03
HS	0.27	0.07	0.03
HSC	0.21	0.06	0.03
Genomic markers			
G	0.67	0.64	0.63
H	0.71	0.62	0.60
S	0.12	0.07	0.06
HS	0.13	0.06	0.06
HSC	0.10	0.05	0.05

Table 17: Average correlation between EBV and true spatial effect in all breeding strategies, using pedigree or genomic markers, and the herd locations simulated using variance $3.5 \cdot 10^{-4}\boldsymbol{I}_{2\times 2}$. The standard error for all values had order of magnitude 10^{-3}

Strategy	A	B	C
Pedigree			
G	0.68	0.64	0.64
H	0.70	0.60	0.58
S	0.11	0.06	0.06
HS	0.12	0.06	0.06
HSC	0.10	0.06	0.06
Genomic markers			
G	0.84	0.74	0.69
H	0.83	0.63	0.50
S	0.16	0.05	0.04
HS	0.21	0.05	0.04
HSC	0.18	0.05	0.03

Table 18: Average correlation between EBV and true spatial effect in all breeding strategies, using pedigree or genomic markers, and the herd locations simulated using variance $9 \cdot 10^{-4} \boldsymbol{I}_{2\times 2}$. The standard error for all values had order of magnitude 10^{-3}

Strategy	A	B	C
Pedigree			
G	0.83	0.72	0.67
H	0.81	0.59	0.47
S	0.12	0.04	0.03
HS	0.15	0.04	0.03
HSC	0.14	0.04	0.02
Genomic markers			
G	0.69	0.65	0.63
H	0.67	0.57	0.56
S	0.09	0.05	0.05
HS	0.09	0.05	0.05
HSC	0.08	0.04	0.04

6.2 Case study: Full Brown-Swiss cattle data

For the models G, H, S and HS applied to the full case study data set for Brown-Swiss cattle we present the posterior hyperparameters in Figure 8, the DIC in Table 19, the mean and standard deviation of the posterior spatial effects from model HS in Figure 9, and the difference in EBV between models H and HS plotted against the mean posterior spatial effect from model HS in Figure 10.

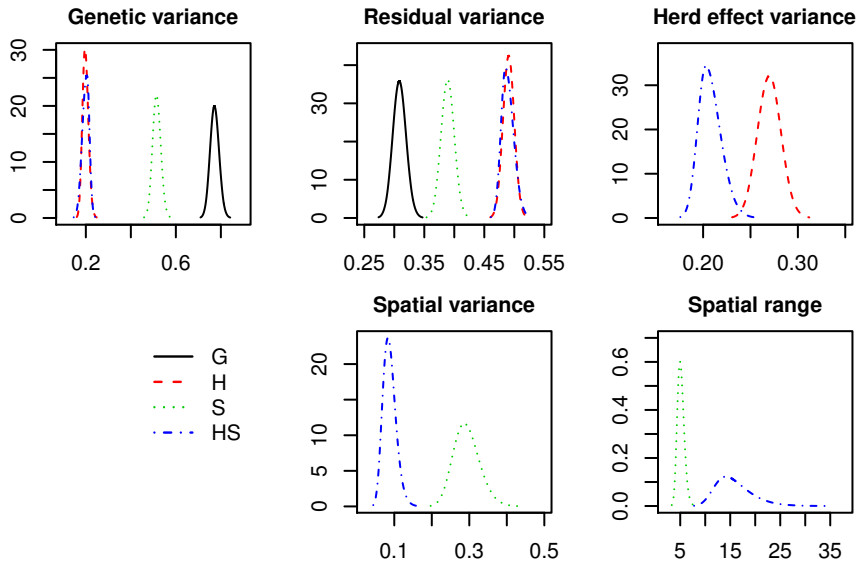


Figure 8: Posterior distributions of hyper-parameters from models G, H, S and HS applied to the full case study data

Table 19: DIC for models G, H, S and HS applied to the full case study data

Model	DIC
G	67329
H	70964
S	70096
HS	70929

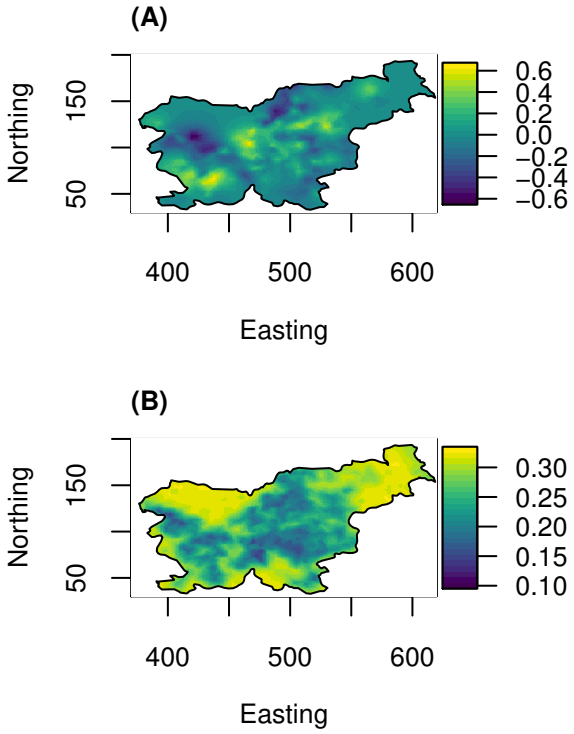


Figure 9: Mean (A) and standard deviation (B) of the posterior spatial effect from HS for the full case study data. The axis units are in km

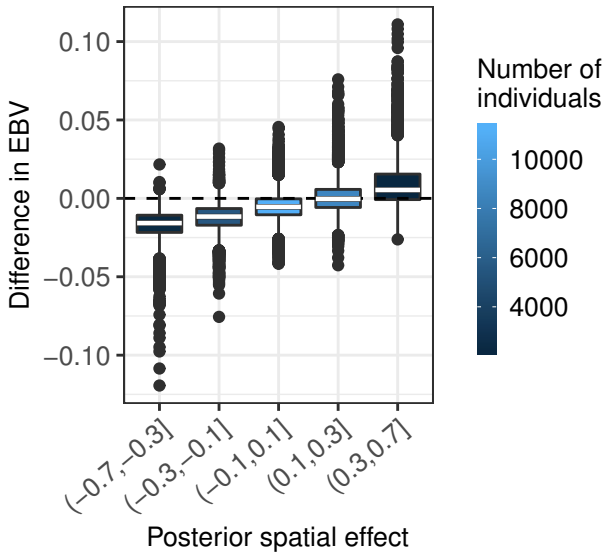


Figure 10: The difference in EBV between H and HS against the mean posterior spatial effect from HS for full case study data

Paper III

Genomic Prediction Including SNP-Specific Variance Predictors

Mouresan, E.F, Selle, M., and Rönnegård, L. (2019) published in *G3: Genes, Genomes, Genetics*

G3 Genes | Genomes | Genetics

GENOMIC PREDICTION

Genomic Prediction Including SNP-Specific Variance Predictors

Elena Flavia Mouresan,*,[1] Maria Selle,[†] and Lars Rönnegård*,[‡]

*Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden, 75007, [†]Department of Mathematical Sciences, Norwegian University of Science and Technology, Norway, 7491, and [‡]School of Technology and Business Studies, Dalarna University, Sweden, 79188

ORCID IDs: 0000-0002-1335-7610 (E.F.M.); 0000-0002-2062-3235 (M.S.); 0000-0002-1057-5401 (L.R.)

ABSTRACT The increasing amount of available biological information on the markers can be used to inform the models applied for genomic selection to improve predictions. The objective of this study was to propose a general model for genomic selection using a link function approach within the hierarchical generalized linear model framework (hglm) that can include external information on the markers. These models can be fitted using the well-established hglm package in R. We also present an R package (CodataGS) to fit these models, which is significantly faster than the hglm package. Simulated data were used to validate the proposed model. We tested categorical, continuous and combination models where the external information on the markers was related to 1) the location of the QTL on the genome with varying degree of uncertainty, 2) the relationship of the markers with the QTL calculated as the LD between them, and 3) a combination of both. The proposed models showed improved accuracies from 3.8% up to 23.2% compared to the SNP-BLUP method in a simulated population derived from a base population with 100 individuals. Moreover, the proposed categorical model was tested on a dairy cattle dataset for two traits (Milk Yield and Fat Percentage). These results also showed improved accuracy compared to SNP-BLUP, especially for the Fat% trait. The performance of the proposed models depended on the genetic architecture of the trait, as traits that deviate from the infinitesimal model benefited more from the external information. Also, the gain in accuracy depended on the degree of uncertainty of the external information provided to the model. The usefulness of these type of models is expected to increase with time as more accurate information on the markers becomes available.

KEYWORDS
BLUP
hglm
CodataGS
external information
Genomic Prediction
GenPred
Shared Data Resources

The identification of a large number of Single Nucleotide Polymorphisms (SNPs) along the genome, as a by-product of the sequencing efforts (*e.g.*, Daetwyler *et al.* 2014) and the development of SNP-chip genotyping technology (Gunderson *et al.* 2005) have made genotyping of thousands of markers affordable at low cost. Meuwissen *et al.* (2001) foresaw these breakthroughs in technology and proposed a new method of selection in animal breeding denoted as Genomic Selection (GS). This method has been tested through simulation studies (Meuwissen *et al.* 2001; Muir 2007) and cross validation with real data in different species such as mice (Legarra *et al.* 2008), dairy cattle (Luan *et al.* 2009; VanRaden *et al.* 2009), aquaculture (Sonesson and Meuwissen 2009) and poultry (González-Recio *et al.* 2009). Nowadays, GS has become part of the routine breeding schemes in dairy cattle (Hayes *et al.* 2009) and other species including pigs (Ostersen *et al.* 2011; Hidalgo *et al.* 2015; Tusell *et al.* 2016) and poultry (Wolc *et al.* 2015).

Several statistical models have been proposed for genomic prediction using whole-genome markers. The most popular method provides best linear unbiased predictions (BLUP) of marker effects (Meuwissen *et al.* 2001) by assuming that the marker effects come from a Gaussian distribution with constant variance and every marker can have an effect on the analyzed trait. This method is referred to either as GBLUP or SNP-BLUP depending on the implementation. Biologically, it seems

doi: https://doi.org/10.1534/g3.119.400381
Manuscript received May 30, 2019; accepted for publication August 9, 2019; published Early Online August 29, 2019.

Supplemental material available at FigShare: https://doi.org/10.25387/g3.9247832.
[1]Corresponding author: Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Ulls väg 26, Box 7023, 75007 Uppsala, Sweden. E-mail: elena.flavia.mouresan@slu.se

more reasonable to assume that some of the markers are in linkage disequilibrium (LD) with a causative gene or a quantitative trait locus (QTL) and therefore can capture their effect on the studied trait, whereas some markers are not in LD with any gene and should therefore not capture any effect. To achieve this idea, several methods have been developed to incorporate different prior assumptions on the genetic architecture of the trait. For this family of methods, often referred to as the Bayesian Alphabet (Gianola 2013), it is assumed that the genetic effects of the SNPs follow alternative distributions like a t-distribution (Bayes A) (Meuwissen *et al.* 2001), a double exponential distribution (Bayes LASSO) (de los Campos *et al.* 2009; Usai *et al.* 2009) or a mixture of distributions (*i.e.*, Bayes B, Bayes Cπ, Bayes R) (Meuwissen *et al.* 2001; Habier *et al.* 2011; Erbe *et al.* 2012). The prior assumptions of these methods are rather arbitrary and their performance relies heavily on the model assumptions capturing accurately the true genetic architecture of the trait of interest (Daetwyler *et al.* 2010; Hayes *et al.* 2010).

Whole-genome sequencing of individuals has facilitated the detection of genetic variants that can be used for GS. Currently, in *Bos Taurus* cattle ∼28 million genetic variants have been reported (Daetwyler *et al.* 2013). This large number of polymorphic markers comes with a major challenge in terms of computational speed and memory. One way to deal with this problem is to make use of the biological information available on the markers, *e.g.*, to annotate the markers in classes based on genome location or functionality and prioritize those classes that show a higher probability of containing trait associated markers. Koufariotis *et al.* (2014) showed that protein coding regions explain significantly more variation than similar number of randomly chosen markers across many traits in cattle. Moreover, in a study by Schork *et al.* (2013), the upstream and downstream classes showed significant enrichment in trait associated variants suggesting that these classes can potentially have important regulatory functions. In the same line, Yang *et al.* (2011) stated that genic regions contributed more additive genetic variance than non-genic regions for human traits. However, Do *et al.* (2015) found that the contribution to total genomic variance per SNP among the annotated classes was similar for all regions in a feed efficiency study in pigs.

Several authors have also investigated the predictive ability of models based on annotation classes. Using kernel methods, Morota *et al.* (2014) and Abdollahi-Arpanahi *et al.* (2016) showed that a whole-genome approach provided better predictive ability than that obtained from classes of genomic regions considered separately. Likewise, Do *et al.* (2015) using GBLUP and Bayesian methods (Bayes A, B and Cπ) found that classification of SNPs by genomic annotation had little impact on the accuracy of prediction for feed efficiency traits in pigs.

Apart from genome annotation information, other biological information is available on the SNPs. QTL databases are available for most livestock species (Hu *et al.* 2013) and Genome-Wide Association Studies (GWAS) (Bush and Moore 2012) have identified a great number of trait-associated markers. Moreover, metabolic and signaling pathways (Kanehisa *et al.* 2008; Croft *et al.* 2011; Caspi *et al.* 2012) and gene regulatory networks (Lee *et al.* 2002; Shalgi *et al.* 2007; Hecker *et al.* 2009) can also provide valuable insight to the underlying biology of the traits of interest (Snelling *et al.* 2013). A rather new tool that has been developed to incorporate existing knowledge of the genetic architecture of complex traits into a GS model is BLUP|GA, *i.e.*, "BLUP approach given the Genetic Architecture" (Zhang *et al.* 2014). This tool uses publicly available GWAS results and showed improved prediction accuracies compared to traditional GBLUP and Bayes B methods. Also, a similar approach was developed by Kadarmideen (2014) (system genomic BLUP, - sgBLUP-) where SNPs with known biological role were explicitly modeled in addition to conventional random SNP effects in SNP-BLUP or GBLUP methods. Along with the BLUP approaches, several Bayesian methods were also developed. Bayes Bπ (Gao *et al.* 2015) is a modified version of Bayes B (Meuwissen *et al.* 2001) able to utilize locus-specific priors. In their study, the authors obtained locus-specific priors from variance analysis (ANOVA) based on information from each single marker separately and the results showed improved accuracy and decreased bias compared to Bayes B and Bayes Cπ. In a similar way, MacLeod *et al.* (2016) proposed a modification to the BayesR method (Erbe *et al.* 2012) that incorporates prior biological knowledge. This method provides a flexible approach to improve the accuracy of genomic prediction and QTL discovery taking advantage of available biological knowledge. The basic idea of previously developed methods is to group SNPs into those having a biological function and those with an unknown function. Both the BLUP|GA and BayesBπ methods, also include continuous weights for all, or a subset of markers. For the BLUP|GA method, weights computed using trait-specific GWAS results are used to construct the genomic relationship matrix, whereas in BayesBπ the weights are computed from single-SNP ANOVA analyses.

Although a large number of methods have been developed already for GS, a general BLUP method to include explanatory variables for SNP-specific variances that allow both continuous and class variables seems to be missing. Here we propose a general model using a link function approach within the hierarchical generalized linear model framework (Lee *et al.* 2006). The algorithm proposed by Lee and Nelder (1996) is used, where the hierarchical generalized linear model is fitted by iterating between augmented generalized linear models. With this approach, rather complex models can be fitted using a single deterministic fitting algorithm (see Rönnegård *et al.* 2010a, 2010b).

The aim of the paper is to assess the accuracy for such models including predictors for SNP variances, with special emphasis on the effect of the trait's genetic architecture and LD structure on estimation accuracy. We present a family of models where the SNP variances can be modeled using both, categorical and continuous predictors, or a combination of the two. The computation time of these models is also studied and a new, faster R package (CodataGS) to fit these models is presented.

MATERIALS AND METHODS

Data simulation

Data were simulated to evaluate the models. Four different scenarios for QTL variance distribution were simulated under three different genetic architectures in which the number of QTL per chromosome was 10, 20 or 100. For each combination of scenario and genetic architecture, 100 simulation replicates were produced. This section describes the simulations in detail.

A base population was simulated of 100 individuals that evolved under random mating for 400 non-overlapping generations (generation -399 to 0) maintaining the population size constant. After the 400 historical generations, two more generations were simulated, still under random mating and expanding the population size from 100 to 200 individuals per generation. Generation 1 was used as training set and generation 2 as validation set. The genome comprised of two chromosomes of 1 Morgan each with 8,800 loci, evenly distributed across the genome. In the base population alleles were coded as 0 or 1 with equal probability resulting in intermediate average allele frequencies. In the first generation, 1,000 loci per chromosome were selected randomly

among those loci with a Minor allele frequency (MAF) higher than 0.05 to simulate the SNP marker panel. The same loci were used for validation in generation 2.

To simulate phenotypes in generation 1 (training set), N_{QTL} loci were selected randomly excluding loci that were on the edge of the chromosome and those with a MAF lower than 0.05. In order to simulate different scenarios of genetic architecture underlying the trait, the number of QTL (N_{QTL}) varied between 10, 20 and 100 per chromosome. Moreover, the QTL effects, u_j, $j = 1, \ldots, N_{QTL}$, were assumed to be normally distributed with mean 0 and varying variance assigned in one of the following ways:

Scenario 0 (Sc0): $u_j \sim N(0, \sigma_j^2)$, where $\sigma_j^2 = e^1$

Scenario 1 (Sc1): $u_j \sim N(0, \sigma_j^2)$, where $\sigma_j^2 = e^1$, with probability 0.5, and $\sigma_j^2 = e^3$, with probability 0.5

Scenario 2 (Sc2): $u_j \sim N(0, \sigma_j^2)$, where $\sigma_j^2 = e^1$, if u_j belonged to chromosome 1, and $\sigma_j^2 = e^3$, if u_j belonged to chromosome 2

Here, e is the natural number and therefore the variance can take values between $e^1 = 2.7$ and $e^3 = 20.1$. The difference between the scenarios Sc1 and Sc2 is that in Sc1 heterogeneous QTL effects are allowed on the same chromosome and may be in linkage disequilibrium with each other. On the other hand, in Sc2 the two different types of QTL are located on different chromosomes to ensure low LD between them.

Scenario 3 (Sc3): $u_j \sim N(0, \sigma_j^2)$, where $\sigma_j^2 = e^{3f(s_j)}$, s_j is the position of QTL j and f is a function of relative distance to the chromosome edge. Consequently, σ_j^2 take values between e^1 and e^3. This scenario is motivated by the finding that fitness genes tend to be located closer to the center of the chromosomes (see *e.g.*, Carneiro *et al.* (2009) and references therein).

For each scenario, the three separate genetic architectures were simulated, *i.e.*, with 10, 20 or 100 QTL per chromosome. In order for the results from the different scenarios and genetic architectures to be comparable, the total genetic variance was scaled to 1.0. In this way, the obtained traits were either controlled by a small number of QTL with medium-large effects or by a large number of QTL with small effects.

In generation 1 (training set) phenotypes were simulated for all 200 individuals as:

$$y_i = \mu + \sum_{j=1}^{N_{QTL}} Z_{QTL,1ij} u_j + e_i,$$

where y_i is the phenotype of individual i, μ is a fixed effect which was set equal to 0, $Z_{QTL,1ij}$ is the genotype for the j^{th} QTL coded as 0, 1 or 2 for the homozygote, heterozygote and the alternative homozygote respectively for individual i in generation 1, u_j is the simulated normally distributed j^{th} QTL effect as described above, and e_i is the residual effect of the i^{th} individual normally distributed with mean 0 and the appropriate variance σ_e^2 in order to create a trait with heritability of 0.2.

Generation 2 was used as validation set where true genomic breeding values (TBVs) were computed as:

$$TBVs_i = \sum_{j=1}^{N_{QTL}} Z_{QTL,2ij} u_j,$$

where $Z_{QTL,2ij}$ is the QTL genotype for QTL j and individual i for this generation.

Genomic evaluation

To estimate the SNP effects, the marker panel of 1,000 SNPs per chromosome mentioned above was used and the following model was assumed:

$$y_i = \mu + \sum_{j=1}^{p} Z_{ij} \upsilon_j + \epsilon_i, \quad (1)$$

where y_i is the phenotype of individual i, μ is a fixed effect, p is the total number of SNPs, Z_{ij} is the genotype of the SNP j for individual i coded as 0, 1 or 2, $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is the residual effect, and

$$\upsilon_j \sim N(0, \tau_j^2) \quad (2)$$

is the j^{th} SNP effect normally distributed with mean 0 and variance

$$\tau_j^2 = e^{\alpha + \beta x_j}, \quad (3)$$

where $\alpha + \beta x_j$ is a linear predictor for the SNP-specific variance the components of which are explained in the following section.

Evaluation models

The linear predictor for variance ($\alpha + \beta x_j$) allows to incorporate any type of external information about the SNP variance, making it possible to assign the same variance for all SNPs, a subgroup of SNPs or assign a unique variance for each SNP. We used this linear predictor for variance to introduce external information on the SNPs into the models and the predictive performance of different prior assumptions was tested. The log link ensures a positive variance (Aitkin 1987; Lee and Nelder 1998) and due to its computational robustness is a common choice of link function in variance modeling (Jaffrezic *et al.* 2000; Sorensen and Waagepetersen 2003; Rönnegård *et al.* 2010a). By using a Gamma generalized linear model with a log link, the score function for this model is equivalent to the score function of the REML likelihood in a linear mixed model (Lee and Nelder 1996, Lee *et al.* 2017 page 91) and therefore produces REML estimates of the variance components. Furthermore, especially for variances close to zero the likelihood will be more symmetric on a logarithmic scale than on an untransformed scale, and thereby gives better standard errors for the fitted variance components.

The models tested in this study were:

1. **SNP-BLUP**: In the traditional model the variance of the markers is assumed to be equal for all markers and therefore $x_j = 0$ in the linear predictor for the variance for all markers.
2. Categorical models (**W10**, **W20** and **W40**): For these models the genome was divided into non-overlapping windows of 10, 20 or 40 SNPs. Then, all the SNPs within a given window were given the value $x_j = 1$ if they contained a QTL and $x_j = 0$ if they did not. Hence, a study with known regions harboring the QTL was mimicked, where these regions were known with varying degree of uncertainty.
3. Continuous model (**LD**): For this model, following Yang and Tempelman (2012) and Rönnegård and Lee (2010), the linkage disequilibrium (LD) between a SNP and a QTL was calculated as $r^2 = D^2/(p_S p_s p_Q p_q)$, where $D = f_{SQ} f_{sq} - f_{Sq} f_{sQ}$ (Falconer and Mackay 1996), p_S, p_s, p_Q and p_q are the allele frequencies of the SNP and QTL, f_{SQ}, f_{sq} are the homozygous haplotype frequencies and f_{Sq}, f_{sQ} are the heterozygous haplotype frequencies. Then, each SNP was assigned the value of $x_j = \sum_{k=1}^{N_{QTL}} r_{jk}^2$. The relationship between SNPs and QTL was modeled in such way that markers in

Table 1 SUMMARY OF MODELS TESTED FOR EACH SCENARIO OF GENETIC ARCHITECTURE SIMULATED

Models[a] Scenario[b]	Sc0	Sc1	Sc2	Sc3
SNP-BLUP	+	+	+	+
W10	+	+	+	+
W20	+	+	+	+
W40	+	+	+	+
LD	+	+	+	+
W10-LD	+	+	+	+
W20-LD	+	+	+	+
W40-LD	+	+	+	+

[a]W10= categorical model with window of 10 SNPs, W20= categorical model with window of 20 SNPS, W40= categorical model with window of 40 SNPS, LD= continuous model with LD estimates, W10-LD= combined model with window of 10 SNPs and LD estimates, W20-LD= combined model with window of 20 SNPs and LD estimates, W40-LD= combined model with window of 40 SNPs and LD estimates.

[b]Sc0= simulation scenario 0, Sc1= simulation scenario 1, Sc2= simulation scenario 2, Sc3= simulation scenario 3.

higher LD with one or more QTL would be given more importance in the model compared to other markers not in LD with any QTL.

4. Combination of categorical and continuous models (**W10-LD**, **W20-LD** and **W40-LD**): In these models the genome was divided into windows as in the previous categorical models but the SNPs located within a window that harbored a QTL were given the value of the LD with the QTL instead of 1. The model could, therefore, differentiate between SNPs not only based on location but also based on the relationship with the real QTL.

Table 1 gives an overview of all simulated scenarios and models tested. Each scenario was simulated with 10, 20 and 100 QTL per chromosome as described previously.

5. Additional models (**W10-2var**, **W20-2var**, **W40-2var**, **Dis**, **W10-Dis**, **W20-Dis** and **W40-Dis**): The previously described models include external information on the physical location of the QTL relative to the SNPs or/and the relationship of the SNPs with the QTL but they do not include any information about the QTL variance. Therefore, a few additional models were created based on the particular parameters used for the simulation of each genetic architecture scenario. These models are defined as follows.
   a. For the scenarios where the QTL effects came from distributions with two different variances (Sc1 and Sc2) we assumed this information was known and we expanded the linear predictor to $\alpha + \beta x_{j1} + \gamma x_{j2}$ in order to accommodate for more variances (in the models W10-2var, W20-2var, and W40-2var). The genome was divided in non-overlapping windows as before and SNPs associated with a QTL with variance $\sigma_j^2 = e^1$ was assigned $x_{j1} = 1$ and $x_{j2} = 0$, while if it was associated with a QTL with variance $\sigma_j^2 = e^3$ it was assigned $x_{j1} = 0$ and $x_{j2} = 1$. If a SNP was located within a window with no QTL then both x_{j1} and x_{j2} had a value of 0.
   b. For Sc3, we used the distance of the markers from the edge of the chromosome as external information either as a continuous variable (Dis) or within windows (W10-Dis, W20-Dis and W40-Dis), since the QTL variances were simulated in the same way.

German Holstein population data

To demonstrate the model on real data, we used a German Holstein genomic prediction population consisting of 5024 bulls (Zhang *et al.* 2015). Three traits were measured, where the first two had highly significant QTL from a GWAS. Including this information as explanatory variables for the SNP-specific variances was expected to improve genomic selection. We were also able to compare our results with Zhang *et al.* (2015), who have developed the algorithm BLUP|GA that includes information about genetic architecture by building trait-specific genomic covariance matrices.

All bulls had been genotyped and we used the 42,373 SNPs with minor allele frequency above 0.01. For the three traits, which were milk yield, milk fat percentage and somatic cell score, Zhang *et al.* (2015) provide highly reliable estimated breeding values (EBVs) for all bulls from previous studies (Hu *et al.* 2013; Zhang *et al.* 2014). The EBVs for milk yield and milk fat percentage were used as phenotypes.

We chose to fit the model 1) **SNP-BLUP** and models 2) **W11** and **W41**, with windows of size 11 and 41 SNPs centered around candidate QTL peaks. To find candidate QTL, we performed GWAS, correcting for genomic relationship using estimated residual and additive genetic variance from GBLUP. All SNPs from GWAS with p-value less than 10^5 were considered a candidate QTL. For milk yield we identified 6 candidate QTL peaks and for the fat percentage we identified 5 candidate QTL peaks, which were used as the center of the windows.

Hglm method and CodataGS

The estimation of the SNP effects was performed by fitting the model described by equations 1-3 that allows both continuous and categorical predictors for the SNP-specific variances, or a combination of continuous and categorical predictors. We tested a few examples of external information on the SNPs and these models were fitted using the **hglm** package in R (Rönnegård *et al.* 2010b). In the **hglm** package the linear predictor for variance $\alpha + \beta x_j$ is specified using the *X.rand.disp* option in the hglm function and the function estimates SNP effects (example of the command line to call the hglm function with the option *X.rand.disp* can be found in the Supplementary File S1 line 156).

When the number of markers largely exceeds the number of individuals, the computational speed and memory requirements can be improved by fitting individual effects (*i.e.*, EGBVs) in an equivalent model instead of SNP effects (Strandén and Garrick 2009; Shen *et al.* 2013). This equivalent model, which uses the external information on each SNP in the same way as in the **hglm** package, was implemented in the R package **CodataGS** and is available on CRAN (https://cran.r-project.org/web/packages/CodataGS). The theory is explained in the Supplementary File S3. The CodataGS R package was used for the analysis of the German Holstein population data.

Accuracy

The predictive ability of all models was evaluated as the correlation of the estimated genomic breeding values (EGBVs) and the true genomic breeding values (TGBVs) for the validation set (Generation 2). For each simulation setup, 100 replicates were generated. The convergence of the models varied from 71 to 100% and results are presented for those replicates where all models converged. For the German Holstein population, we performed a fivefold cross-validation with all bulls randomly separated in four groups of 1005 and one group of 1004 with both model 1) and 2). Each group served as a test set while the rest of the groups were used to estimate the SNP effects. The predictive ability was measured as the correlation between the EBVs and the phenotypes of the testing individuals.

Data availability

Simulation of the data that support the findings is possible through the attached simulation code in File S1 and File S2 (Functions for the

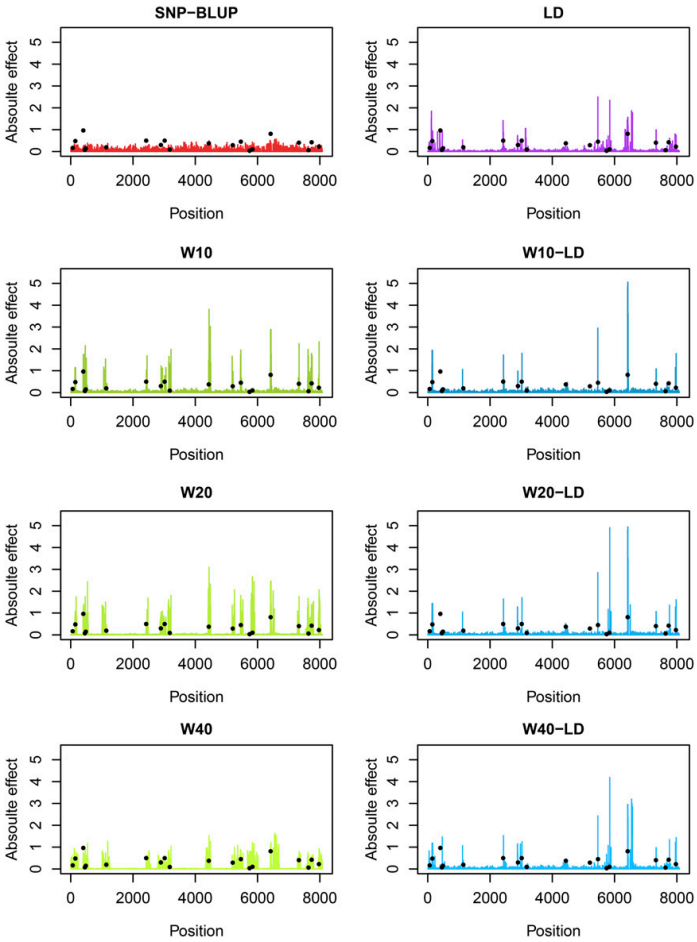


Figure 1 Simulated QTL effects (black dots) and fitted SNP effects under SNP-BLUP and 7 alternative models (Categorical: W10, W20 and W40, Continuous: LD, Combination: W10-LD, W20-LD and W40-LD) for one simulation replicate under simulation scenario Sc0 with 10 QTL per chromosome underlying the trait.

simulation) deposited at figshare. The simulation code and the methodology described previously are sufficient to reproduce the results of this study. The analysis program CodataGS used to apply the alternative models on the Holstein dataset is available at https://cran.r-project.org/web/packages/CodataGS. Supplemental material available at FigShare: https://doi.org/10.25387/g3.9247832.

RESULTS

Table 1 contains different versions of the model tested. The fitted SNP effects obtained from **hglm** for one simulation replicate under scenario Sc0 with 10 QTL per chromosome are presented in Figure 1. The R code to reproduce Figure 1 is found in Supplementary File S1 (along with File S2). The results show how the fitted SNP effects may change between model specifications. For example, it can be observed that with increasing window size the estimated effects tend to be spread between more SNPs.

Model performance

Table 2 shows the accuracies of the predicted EGBVs in the validation set (generation 2) for scenario 0 (Sc0) with 10 QTL per chromosome underlying the trait. In general, the alternative models performed better than SNP-BLUP. The categorical models yielded higher accuracies compared to the SNP-BLUP model by 14.3% (0.670 ± 0.013), 11.9% (0.656 ± 0.012) and 8.4% (0.635 ± 0.012) for the models W10, W20 and W40, respectively. Nonetheless, we observe that the advantage of the categorical models over the SNP-BLUP decreased with increasing window sizes. Moreover, the continuous model (LD) resulted in higher accuracy than the SNP-BLUP or the categorical models with an increase of 22.4% (0.717 ± 0.011) in accuracy with respect to the SNP-BLUP. Similarly, the combination models performed 20.6% (W10-LD, 0.707 ± 0.013) 21.8% (W20-LD, 0.714 ± 0.013) and 23.2% (W40-LD, 0.722 ±0.013) better than the SNP-BLUP model. Contrary to the categorical models, the combination models maintained the gain in accuracy with increasing window size. The alternative models provided unbiased predictions while the SNP-BLUP showed upward bias (Table 2). Finally, the mean squared error of prediction (MSEP) in the validation set improved with the alternative models compared to the SNP-BLUP, indicating that predictions are closer to the true breeding values in the alternative models compared with the SNP-BLUP.

Table 2 ACCURACY AND BIAS OF THE PREDICTED EGBVS IN THE VALIDATION SET (GENERATION 2) FOR THE SCENARIO 0 (SC0) WITH 10 QTLS PER CHROMOSOME UNDERLYING THE TRAIT

Models[a]	Accuracy (r)	Bias (b)
SNP-BLUP	0.586 (0.010)	1.213 (0.089)
W10	0.670 (0.013)	1.003 (0.044)
W20	0.656 (0.012)	1.014 (0.048)
W40	0.635 (0.012)	1.030 (0.045)
LD	0.717 (0.011)	1.024 (0.041)
W10-LD	0.707 (0.013)	1.050 (0.053)
W20-LD	0.714 (0.013)	1.044 (0.050)
W40-LD	0.722 (0.013)	1.028 (0.042)

[a]W10= categorical model with window of 10 SNPs, W20= categorical model with window of 20 SNPS, W40= categorical model with window of 40 SNPS, LD= continuous model with LD estimates, W10-LD= combined model with window of 10 SNPs and LD estimates, W20-LD= combined model with window of 20 SNPs and LD estimates, W40-LD= combined model with window of 40 SNPs and LD estimates.

Effect of number of simulated QTL

In order to investigate the performance of the alternative models for traits with different genetic architectures we simulated a trait controlled by an increasing number of QTL with each having a decreasing effect. As an overview, the accuracies of the different models in Sc0 with 20 and 100 QTL per chromosome are visualized in Figure 2 together with the results from 10 QTL per chromosome. The advantage of the alternative models over the SNP-BLUP model decreased with increasing number of QTL controlling the trait. When the number of QTL underlying the trait is 20 QTL per chromosome, the accuracies obtained were 9.6%, 7.5% and 3.8% better than the SNP-BLUP for the W10, W20 and W40 models, respectively. The continuous model resulted in a gain of 12.9% in accuracy while the combination models performed slightly better than all the alternative models yielding gains in accuracy of 14%, 14.2% and 13.5% for the W10-LD, W20-LD and W40-LD models, respectively. Finally, in the case of 100 QTL per chromosome, all models performed roughly the same as SNP-BLUP, yielding accuracies between 0.583 ± 0.012 (W40) and 0.599 ± 0.011 (W10-LD).

Effect of variance of the QTL effects

The genetic architecture of a trait does not only depend on the number of QTL that affect the trait. For example, mutations can affect protein coding regions or regulatory regions and these mutations can have a bigger or smaller effect on the trait. Therefore we can assume that their effects come from a mixture of distributions with varying variance over the genome. For this purpose we simulated several scenarios where the QTL effects were drawn from a mixture of distributions (see Sc1 – Sc3 in Materials and Methods). We compared the performance of all models under all scenarios of QTL effect variances and all cases of number of QTL affecting the trait (Figure 3). In general the models performed similarly under Sc1, Sc2 and Sc3 as in Sc0. Small differences were observed in the case of 10 QTL per chromosome where all models performed slightly better in Sc0 and Sc2 (QTL effects from a low variance distribution on chromosome 1 and from high variance distribution on chromosome 2) compared with the results from Sc1 and Sc3. Nonetheless, this minimum difference disappeared quickly with increasing number of QTL per chromosomes. The external information included in the alternative models was related to the position of the QTL on the genome and/or the relationship of the SNPs with the QTL (LD), but no information about the distribution of the variance itself was included. Therefore, we fitted additional models that considered the way the QTL were simulated (see linear predictor 5: Additional models Material and Methods, and Supplementary file Table S1). For Sc1 and Sc2 we extended the linear predictor ($\alpha + \beta x_{j1} + \gamma x_{j2}$) to accommodate for two types of variances for the SNPs in windows that harbored a QTL assuming that we knew beforehand the distribution variance of the effect of that QTL and, as before, we tested 3 different window sizes (10, 20 and 40 SNPs per window). The results showed that these additional models performed similarly as the categorical models (W10, W20 and W40) under all cases of genetic architecture simulated. The only exception to these results was for the Sc2 with 100 QTL per chromosome where additional models showed a small increase in accuracy compared to all other models (Supplementary files, Figure S1). For the Sc3 we used the distance of the SNP from the edge of the chromosome as external information, either as a continuous variable or within windows. Similarly as before, the additional models that included information on the simulated distribution variance of the QTL did not perform better than the alternative models. The combined models (W10-Dis, W20-Dis and W40-Dis) performed the same as the categorical models while the continuous model (Dis) showed no benefit compared to the alternative models or the SNP-BLUP model under any simulation scenario of genetic architecture.

Computation time

When the number of markers exceeds the number of individuals, the computational speed and memory requirements can be an important drawback for the use of such models. A solution to this problem is to fit individual effects (*i.e.*, EGBVs) in an equivalent model instead of SNP effects. In this study all evaluations were performed using the **hglm** R package that fits SNP effects. For a larger number of SNPs the computations would be unfeasible and an equivalent model which uses the external information on each SNP in the same way as in the **hglm** package was implemented in the R package **CodataGS**

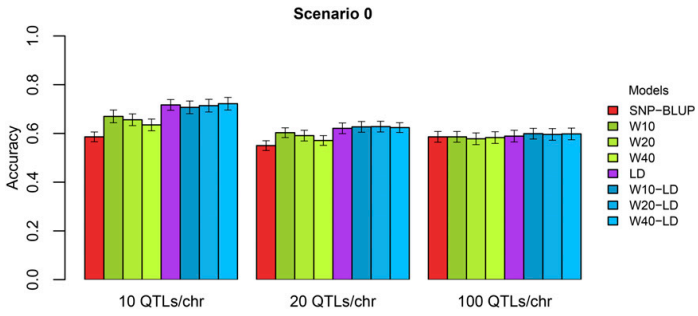


Figure 2 Accuracies obtained under different cases of genetic architecture of the trait for SNP-BLUP and the alternative models.

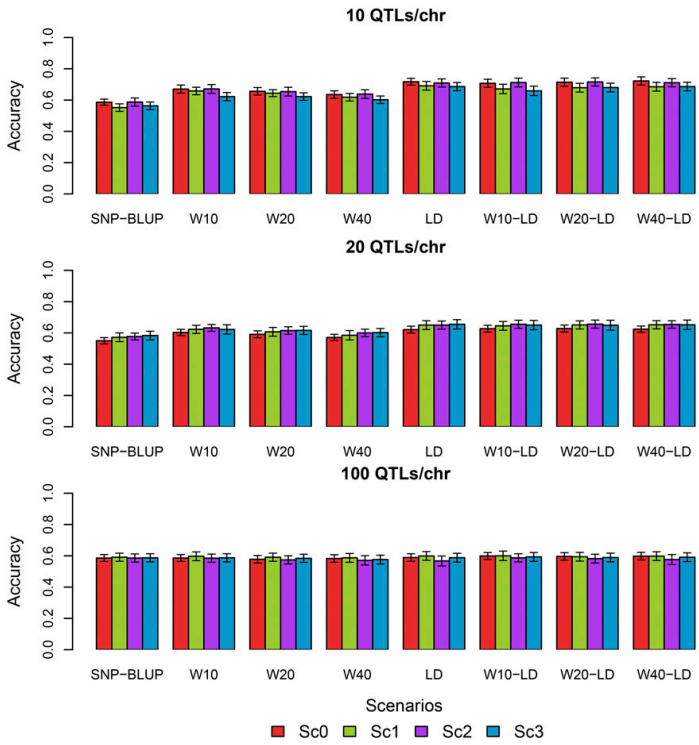


Figure 3 Accuracies obtained from SNP-BLUP model and alternative models under all simulated scenarios and genetic architectures.

(https://cran.r-project.org/web/packages/CodataGS). The theory is explained in the Supplementary File S3. Fitting individual effects instead of SNP effects resulted in largely improved run time of all models. For a training population of 200 individuals with 2,000 SNP markers, fitting SNP effects (**hglm**) required on average 9.35 sec per iteration while fitting individual effects (**CodataGS**) required only 0.46 sec per iteration (Figure 4). The improved speed and memory requirements of the equivalent model can be considerably beneficial since the usual size of the training sets is much larger than the one used here (thousands of individuals with tens of thousands of SNPs). Nonetheless, the speed performance of the equivalent model depends heavily on the number of individuals and the relationship between time and number of individuals is not linear but rather exponential (Supplementary Figure S2).

German Holstein population results

To demonstrate the model on real data, we used a German Holstein population consisting of 5024 bulls (Zhang *et al.* 2015). We chose to fit the model 1) **SNP-BLUP** and models 2) **W11** and **W41** with windows of sizes 11 and 41 SNPs centered around candidate QTL peaks. We obtained the candidate QTL peaks after performing a GWAS, correcting for genomic relationship using estimated residual and additive genetic variance from GBLUP. All SNPs from the GWAS with p-value less than 10^5 were considered a candidate QTL. For milk yield (MY) we identified 6 candidate QTL peaks and for the fat percentage (Fat%) we identified 5 candidate QTL peaks, which were used as the center of the windows.

Table 3 shows the average accuracies obtained from the SNP-BLUP and W41 models for two traits (MY and Fat%) in the

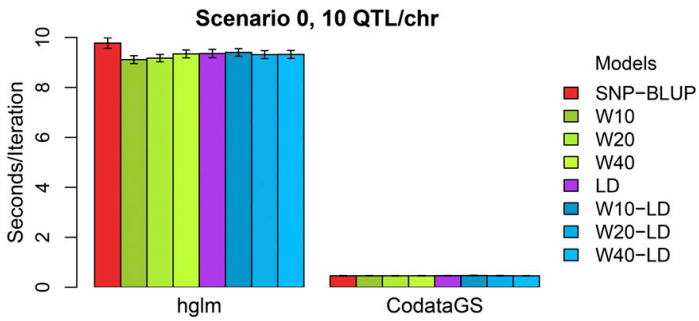


Figure 4 Time of execution (seconds per iteration) of SNP-BLUP and alternative models from hglm package and CodataGS package.

Table 3 MEAN ACCURACY (STANDARD ERROR) OF THE PREDICTED EGBVS IN A 5-FOLD CROSS VALIDATION ANALYSIS USING THE GERMAN HOLSTEIN DATA FOR TWO TRAITS

Models[a]	MY	Fat%
SNP-BLUP	0.771 (0.002)	0.811 (0.004)
W41	0.785 (0.002)	0.862 (0.003)

[a]W41= categorical model with window of 40 SNPs around the top SNP for the trait detected on a GWAS study. MY: Milk Yield, Fat%: Fat percentage.

fivefold-cross-validation analysis. We present only the results from the W41 model as the model W11 yielded very similar accuracies as the W41 model. For both traits the W41 model yielded higher accuracies than the SNP-BLUP. The W41 model showed a higher advantage in predictive ability for the trait Fat% yielding an accuracy of 0.862 compared to the 0.811 obtained from the SNP-BLUP model. The results for the MY trait were similar but the predictive advantage of the W41 model was lower compared to the Fat% trait (accuracy of 0.785 from the W41 model over 0.771 from the SNP-BLUP model).

DISCUSSION

The knowledge on the genetic architecture of different traits, and SNP-specific biological information, is increasing rapidly and several authors have proposed methods for genomic selection that can make use of this available biological information to improve selection accuracy (Zhang *et al.* 2010; Zhang *et al.* 2014; Su *et al.* 2014). In this line, this study proposes a general model using a link function approach within the hierarchical generalized linear model framework (Lee *et al.* 2006) to include biological external information into the model. Following Zhang *et al.* (2010), we used a base population of 100 individuals in our simulation study. This is a rather small population size and the results should therefore be extrapolated to larger effective population sizes with caution.

All the results in the current study use the same general model (described by equations 1 – 3) for predicting breeding values. The alternative models in Table 1, including SNP-BLUP, are fitted within this single framework and in the results the accuracies of the alternative models are compared. There are numerous Bayesian models not included within this framework that may be of interest to compare with. However, we use SNP-BLUP as a basic model to compare the results to and study the accuracies of models that make use of external information on the SNPs.

A very attractive feature of the method proposed in this study is that it provides a flexible way to model the SNP variances using a linear predictor (equation 3). Any type of existing knowledge on the SNP markers can be utilized and potentially increase the predictive ability of the model. In this study we investigated the performance of external information related to the position of the QTL on the genome and the relationship of the SNP markers with the QTL and we showed that the inclusion of such information can improve the predicting ability of genomic selection. From our results we identified two main factors that influence the performance of such models, the genetic architecture of the trait and the quality/accuracy of the external information.

In the models W10, W20 and W40, the causative effect is assumed to be within a window and does not assume that the exact position of the causative mutation is known. This model should be suitable for genomic prediction where external information from QTL studies is included. For the LD model and the combined models (W10-LD, W20-LD and W40-LD) it is assumed that the position of the causative SNP is known. Especially in plant breeding, there is a need to include major genes, whose positions are accurately known, in genomic prediction. For such cases the models including LD information combines marker assisted selection and genomic selection in a dynamic way.

We investigated models with three different window sizes that were suitable for our simulated data. For applications on real data the optimal number of markers to be included in each window, in terms of prediction accuracy, will depend on marker density and the genetic architecture. In our application on the dairy cattle data the optimal number of markers within a window was not assessed statistically, but since the marker map was much denser than in the simulated data we chose the model with the largest number of markers, *i.e.*, window size 40.

Genetic architecture of the trait

The performance of several alternative models in our study was better compared to the SNP-BLUP method when the trait was controlled by a small number of QTL with medium-large effects. The advantage of these models was reduced with increasing number of QTL with smaller effects. However, the alternative models did not result in lower accuracies compared to the SNP-BLUP model. The reason is that as the estimated effect of the external information on the SNP variances approaches zero the model reduces to a SNP-BLUP model. Furthermore, as the number of QTL that control the trait increases, the external information on SNPs becomes more similar among the SNPs. For example, for the categorical models, a QTL is located within most or all defined windows and as a result all SNPs get the same weight in the model. Moreover, most or all SNPs are in LD with a QTL at similar levels. Consequently, the alternative models turn into a SNP-BLUP model. These results are in agreement with the findings of Zhang *et al.* (2010). In their simulation study they investigated the performance of a BLUP model with weighted G matrix and showed that for traits controlled by high number of QTL the traditional GBLUP and their method performed similarly. This effect has also been observed in studies on real data (Zhang *et al.* 2014). Analyzing three dairy cattle traits (Milk Yield (MY), Fat percentage (FP) and Somatic Cell Count (SCC)) these authors found that traits controlled by a few QTL with large effects (MY and FP) perform better under models with external information on the SNPs while the SCC trait, that is controlled by many QTL evenly distributed along the genome, performed better under the standard GBLUP model.

In our simulation study we created different genetic architectures for the trait with respect not only to the number of the QTL affecting the trait but also to the distribution of the QTL effects and their variances (see Material and Methods). Our results showed that this aspect did not affect the performance of the alternative models. Moreover, the additional models that included information on the variance distribution across the genome were not able to provide any benefit, contrary to methods that assume mixtures of distributions for the SNP markers like Bayesian methods (Erbe *et al.* 2012).

External information

In this study we investigated the performance of models that include information on the location of the QTL on the genome (categorical models) and thereby tried to mimic the external information available on the QTL databases and the different window sizes resemble the degree of uncertainty of a QTL region. Our results indicate that this type of external information has the potential to improve the accuracy of genomic selection and that the degree of improvement is inversely related to the degree of uncertainty on the QTL region. The usefulness of the QTL database information has been demonstrated by Zhang *et al.* (2014). In their study these authors searched for reported QTL on the traits under consideration (Fat percentage, milk yield and somatic cell score for dairy cattle and several traits for rice) and after a quality

control to avoid the possible false positive reports they included this information into a GBLUP model. For most of the examined traits an increase in accuracy was observed, especially for the traits that showed a characteristic genetic architecture. The discovery of new QTL or the causative mutations is expected to increase in the future with the use of whole genome sequence and the development of new methods for analysis and as a consequence the information available will become more accurate.

The external information that proved to be more valuable in this study was the LD estimates between the SNPs and the QTL. In the standard GBLUP method, markers in linkage equilibrium (LE) to the causative QTL tend to capture effects due to family relationship, whereas mainly markers in LD capture the QTL effects themselves (Habier *et al.* 2007, de los Campos *et al.* 2015). In the BayesB model (Meuwissen *et al.* 2001), the prior for the SNP variances is a mixture of two distributions that tends to group markers into two classes: those in LD and those in LE with the QTL. By modeling the two classes of markers better predictions for unrelated individuals can be obtained. In other studies, LD information has been incorporated in a model for the marker variances, which smooths the effects between markers in close LD (*e.g.*, the Bayesian antedependence model by Yang and Tempelman 2012, and the double hierarchical generalized linear model by Rönnegård and Lee 2010), and thereby captures the QTL effects rather than family information. These models give better predictions than GBLUP when individuals are unrelated and the total number of QTL is small. This is in line with our findings where the models including LD between markers and QTL resulted in improved prediction accuracies, especially when the number of simulated QTL was small. Finally, the results obtained from the combined models indicate that information on the real relationship between markers and QTL can compensate for the loss of information due to the uncertainty of the QTL report.

The prior of BayesB is rather general because it does not use any external information on the SNPs, whereas the model we propose gives more specific information about each SNP. Since the information on each SNP is more specific in our model its performance compared to GBPLUP and BayesB is expected to improve as the number of individuals in the training set decreases, in line with the results of Zhang *et al.* (2015, Supplementary Table 1) .

The model applied in Zhang *et al.* (2015) is BLUP|GA and was developed in Zhang *et al.* (2014). It includes external data on SNPs in the model and has similarities to our model since both methods fit trait-specific genomic relationship matrices. In the BLUP|GA method SNPs are divided into two groups by the user. In the first group there is a single genetic variance for all SNPs and in the second group SNP-specific variances are modeled as proportional to user-specific weights. Furthermore, the ratio between the variances for the two groups is also user-specified. This is indeed similar to our proposed method, but with some significant differences. The method that we propose uses a regression approach where covariates are specified by the user, whereas all model parameters are estimated. The covariates can include negative values in our method but the SNP variances will still be positive because the genetic variances are modeled using a logarithmic link function. By specifying covariates rather than weights for the SNP variances, hopefully, our proposed method will also be user friendly and the implementation in the CodataGS package (https://cran.r-project.org/web/packages/CodataGS) fits rather well with the regression framework in R.

CONCLUSIONS

In this study we investigated the potential benefit of external information on improving the accuracy of genomic selection. In conclusion, using external information to model SNP-specific variances can provide gains in accuracy compared to the traditional SNP-BLUP. Nonetheless, the level of gain depends on the genetic architecture of the trait of interest and the quality of the external information on the SNP markers. The usefulness of these type of models is expected to increase with time as more accurate information on the SNPs becomes available. Finally, our analysis on real data indicated that the proposed method has potential but further studies are required to confirm the advantage of this approach.

ACKNOWLEDGMENTS

This project was supported by the Mistra Biotech project, a research program financed by Mistra – the Swedish foundation for strategic environmental research, and the Swedish University of Agricultural Sciences, SLU. M. Selle acknowledges the financial support given by the Research Council of Norway, grant number 250362.

LITERATURE CITED

Abdollahi-Arpanahi, R., G. Morota, B. D. Valente, A. Kranis, G. J. M. Rosa *et al.*, 2016 Differential contribution of genomic regions to marked genetic variation and prediction of quantitative traits in broiler chickens. Genet Sel Evol: GSE 48: 10. https://doi.org/10.1186/s12711-016-0187-z

Aitkin, M., 1987 Modelling variance heterogeneity in normal regression using GLIM. J. R. Stat. Soc. Ser. C Appl. Stat. 36: 332–339.

Bush, W. S., and J. H. Moore, 2012 Chapter 11: Genome-Wide Association Studies. PLOS Comput. Biol. 8: e1002822. https://doi.org/10.1371/journal.pcbi.1002822

Carneiro, M., N. Ferrand, and M. W. Nachman, 2009 Recombination and Speciation: Loci near Centromeres Are More Differentiated than Loci near Telomeres between Subspecies of the European Rabbit (Oryctolagus Cuniculus). Genetics 181: 593–606. https://doi.org/10.1534/genetics.108.096826

Caspi, R., T. Altman, K. Dreher, C. A. Fulcher, P. Subhraveti *et al.*, 2012 The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 40: D742–D753. https://doi.org/10.1093/nar/gkr1014

Croft, D., G. O'Kelly, G. Wu, R. Haw, M. Gillespie, *et al.*, 2011 Reactome: A Database of Reactions, Pathways and Biological Processes. Nucleic Acids Res. 39 (SUPPL. 1).

Daetwyler, H. D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen *et al.*, 2014 Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46: 858–865. https://doi.org/10.1038/ng.3034

Daetwyler, H. D., M. P. L. Calus, R. Pong-Wong, G. de los Campos, and J. M. Hickey, 2013 Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. Genetics 193: 347 LP-365.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams, 2010 The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021–1031. https://doi.org/10.1534/genetics.110.116855

de los Campos, G., D. Sorensen, and D. Gianola, 2015 Genomic heritability: What is it? PLoS Genet. 11: e1005048. https://doi.org/10.1371/journal.pgen.1005048

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385. https://doi.org/10.1534/genetics.109.101501

Do, D. N., L. L. G. Janss, J. Jensen, and H. N. Kadarmideen, 2015 SNP Annotation-Based Whole Genomic Prediction and Selection : An Application to Feed Efficiency and Its Component Traits in Pigs. J. Anim. Sci. 93: 2056–2063. https://doi.org/10.2527/jas.2014-8640

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Dairy Sci. 95: 4114–4129. https://doi.org/10.3168/jds.2011-5019

Falconer, D. S., and T. F. C. Mckay, 1996 Introduction to Quantitative Genetics. Ed. 4. Longmans Green, Harlow, Essex.

Gao, N., J. Li, J. He, G. Xiao, Y. Luo *et al.*, 2015 Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. BMC Genet. 16: 120. https://doi.org/10.1186/s12863-015-0278-9

Gianola, D., 2013 Priors in whole-genome regression: The Bayesian alphabet returns. Genetics 194: 573–596. https://doi.org/10.1534/genetics.113.151753

González-Recio, O., D. Gianola, G. J. M. Rosa, K. A. Weigel, and A. Kranis, 2009 Genome-Assisted Prediction of a Quantitative Trait Measured in Parents and Progeny: Application to Food Conversion Rate in Chickens. Genet Sel Evol: GSE 41: 3. https://doi.org/10.1186/1297-9686-41-3

Gunderson, K. L., F. J. Steemers, G. Lee, L. G. Mendoza, and M. S. Chee, 2005 A genome-wide scalable SNP genotyping assay using microarray technology. Nat. Genet. 37: 549–554. https://doi.org/10.1038/ng1547

Habier, D., R. L. Fernando, and J. C. M. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. Genetics 177: 2389–2397. https://doi.org/10.1534/genetics.107.081190

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186. https://doi.org/10.1186/1471-2105-12-186

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard, 2010 Genetic Architecture of Complex Traits and Accuracy of Genomic Prediction: Coat Colour, Milk-Fat Percentage, and Type in Holstein Cattle as Contrasting Model Traits. PLoS Genet. 6: e1001139. https://doi.org/10.1371/journal.pgen.1001139

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard, 2009 Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443. https://doi.org/10.3168/jds.2008-1646

Hecker, M., S. Lambeck, S. Toepfer, E. van Someren, and R. Guthke, 2009 Gene regulatory network inference: Data integration in dynamic models-a review. Biosystems 96: 86–103. https://doi.org/10.1016/j.biosystems.2008.12.004

Hidalgo, A. M., J. W. M. Bastiaansen, M. S. Lopes, B. Harlizius, M. A. M. Groenen, *et al.*, 2015 Accuracy of predicted genomic breeding values in purebred and crossbred pigs. G3 (Bethesda) 5: 1575–1583.

Hu, Z. L., C. A. Park, X. L. Wu, and J. M. Reecy, 2013 Animal QTLdb: An improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41: D871–D879. https://doi.org/10.1093/nar/gks1150

Jaffrezic, F., I. M. S. White, R. Thompson, and W. G. Hill, 2000 A Link Function Approach to Model Heterogeneity of Residual Variances Over Time in Lactation Curve Analyses. J. Dairy Sci. 83: 1089–1093. https://doi.org/10.3168/jds.S0022-0302(00)74973-3

Kadarmideen, H., 2014 Genomics to systems biology in animal and veterinary sciences: Progress, lessons and opportunities. Livest. Sci. 166: 232–248. https://doi.org/10.1016/j.livsci.2014.04.028

Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, *et al.*, 2008 KEGG for Linking Genomes to Life and the Environment. Nucleic Acids Res. 36 (SUPPL. 1).

Koufariotis, L., Y. P. Chen, S. Bolormaa, B. J. Hayes, A. J. Schork, *et al.*, 2014 Regulatory and Coding Genome Regions Are Enriched for Trait Associated Variants in Dairy and Beef Cattle. BMC Genomics 15 (1). BioMed Central: 436. https://doi.org/10.1186/1471-2164-15-436

Lee, T. I., N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph *et al.*, 2002 Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298: 799–804. https://doi.org/10.1126/science.1075090

Lee, Y., and J. A. Nelder, 1996 Hierarchical Generalized Linear Models. J. R. Stat. Soc. B 58: 619–678.

Lee, Y., and J. A. Nelder, 1998 Generalized Linear Models for the Analysis of Quality-Improvement Experiments. Can. J. Stat. 26: 95–105. https://doi.org/10.2307/3315676

Lee, Y., J. A. Nelder, and Y. Pawitan, 2006 *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman & Hall/CRC, Boca Raton. https://doi.org/10.1201/9781420011340

Lee, Y., L. Rönnegård, and M. Noh, 2017 *Data Analysis Using Hierarchical Generalized Linear Models with R*, Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781315211060

Legarra, A., C. Robert-Granié, E. Manfredi, and J. M. Elsen, 2008 Performance of Genomic Selection in Mice. Genetics 180: 611–618. https://doi.org/10.1534/genetics.108.088575

Luan, T., J. A. Woolliams, S. Lien, M. Kent, M. Svendsen *et al.*, 2009 The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics 183: 1119–1126. https://doi.org/10.1534/genetics.109.107391

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper *et al.*, 2016 Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genomics 17: 144. https://doi.org/10.1186/s12864-016-2443-6

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola, 2014 Genome-Enabled Prediction of Quantitative Traits in Chickens Using Genomic Annotation. BMC Genomics 15 (1). BioMed Central: 109. https://doi.org/10.1186/1471-2164-15-109

Muir, W. M., 2007 Comparison of Genomic and Traditional BLUP-Estimated Breeding Value Accuracy and Selection Response under Alternative Trait and Genomic Parameters. J. Anim. Breed. Genet. 124: 342–355. https://doi.org/10.1111/j.1439-0388.2007.00700.x

Ostersen, T., O. F. Christensen, M. Henryon, B. Nielsen, G. Su *et al.*, 2011 Deregressed EBV as the Response Variable Yield More Reliable Genomic Predictions than Traditional EBV in Pure-Bred Pigs. Genet Sel Evol: GSE 43: 38. https://doi.org/10.1186/1297-9686-43-38

Rönnegård, L., and Y. Lee, 2010 Hierarchical generalized linear models have a great potential in genetics and animal breeding. In proceedings: World Congress on Genetics Applied to Livestock Production, Leipzig, Germany.

Rönnegård, L., M. Felleki, F. Fikse, H. A. Mulder, and E. Strandberg, 2010a Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. Genet Sel Evol: GSE 42: 8. https://doi.org/10.1186/1297-9686-42-8

Rönnegård, L., X. Shen, and M. Alam, 2010b hglm: A Package for Fitting Hierarchical Generalized Linear Models. R J. 2: 20–28. https://doi.org/10.32614/RJ-2010-009

Schork, A. J., W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey *et al.*, 2013 All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. PLoS Genet. 9: e1003449. https://doi.org/10.1371/journal.pgen.1003449

Shalgi, R., D. Lieber, M. Oren, and Y. Pilpel, 2007 Global and local architecture of the mammalian microRNA-transcription factor regulatory network. PLOS Comput. Biol. 3: e131. https://doi.org/10.1371/journal.pcbi.0030131

Shen, X., M. Alam, F. Fikse, and L. Rönnegård, 2013 A Novel Generalized Ridge Regression Method for Quantitative Genetics. Genetics 193: 1255–1268. https://doi.org/10.1534/genetics.112.146720

Snelling, W. M., R. A. Cushman, J. W. Keele, C. Maltecca, M. G. Thomas *et al.*, 2013 BREEDING AND GENETICS SYMPOSIUM: Networks and pathways to guide genomic selection. J. Anim. Sci. 91: 537–552. https://doi.org/10.2527/jas.2012-5784

Sonesson, A. K., and T. H. E. Meuwissen, 2009 Testing Strategies for Genomic Selection in Aquaculture Breeding Programs. Genet Sel Evol: GSE 41: 37. https://doi.org/10.1186/1297-9686-41-37

Sorensen, D., and R. Waagepetersen, 2003 Normal linear models with genetically structured residual variance heterogeneity: a case study. Genet. Res. 82: 207–222. https://doi.org/10.1017/S0016672303006426

Strandén, I., and D. J. Garrick, 2009 Technical Note: Derivation of Equivalent Computing Algorithms for Genomic Predictions and Reliabilities of Animal Merit. J. Dairy Sci. 92: 2971–2975. https://doi.org/10.3168/jds.2008-1929

Su, G., O. F. Christensen, L. Janss, and M. S. Lund, 2014 Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus/specific variances. J. Dairy Sci. 97: 6547–6559. https://doi.org/10.3168/jds.2014-8210

Tusell, L., H. Gilbert, J. Riquet, M. J. Mercat, A. Legarra *et al.*, 2016 Pedigree and genomic evaluation of pigs using a terminal-cross model. Genet Sel Evol: GSE 48: 32. https://doi.org/10.1186/s12711-016-0211-3

Usai, M. G., M. E. Goddard, and B. J. Hayes, 2009 LASSO with cross-validation for genomic selection. Genet. Res. 91: 427–436. https://doi.org/10.1017/S0016672309990334

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel *et al.*, 2009 Invited Review: Reliability of Genomic Predictions for North American Holstein Bulls. J. Dairy Sci. 92: 16–24. https://doi.org/10.3168/jds.2008-1514

Wolc, A., H. H. Zhao, J. Arango, P. Settar, J. E. Fulton *et al.*, 2015 Response and inbreeding from a genomic selection experiment in layer chickens. Genet Sel Evol: GSE 47: 59. https://doi.org/10.1186/s12711-015-0133-5

Yang, J., T. A. Manolio, L. R. Pasquale, E. Boerwinkle, N. Caporaso *et al.*, 2011 Genome Partitioning of Genetic Variation for Complex Traits Using Common SNPs. Nat. Genet. 43: 519–525. https://doi.org/10.1038/ng.823

Yang, W., and R. J. Tempelman, 2012 A Bayesian antedependence model for whole genome prediction. Genetics 190: 1491–1501. https://doi.org/10.1534/genetics.111.131540

Zhang, Z., U. Ober, M. Erbe, H. Zhang, N. Gao *et al.*, 2014 Improving the accuracy of Whole Genome Prediction for Complex Traits using the results of Genome Wide Association Studies. PLoS One 9: e93017. https://doi.org/10.1371/journal.pone.0093017

Zhang, Z., M. Erbe, J. He, U. Ober, N. Gao *et al.*, 2015 Accuracy of whole-genome prediction using a genetic architecture-enhanced variance-covariance matrix. G3: Genes, Genomes. Genetics 5: 615–627.

Zhang, Z., J. F. Liu, X. D. Ding, P. Bijma, D. J. de Koning *et al.*, 2010 Best linear unbiased prediction of genomic breeding values using trait-specific marker-derived relationship matrix. PLoS One 5: e12648. https://doi.org/10.1371/journal.pone.0012648

Communicating editor: L. McIntyre

Paper IV

Hierarchical modeling of haplotype effects based on a phylogeny

Selle, M. L., Steinsland, I., Lindgren, F., Brajkovic, V., Cubric-Curik,V., and Gorjanc, G. (2020) submitted to *Frontiers in Genetics*

Hierarchical modeling of haplotype effects based on a phylogeny

Maria L. Selle [1] Ingelin Steinsland [1] Finn Lindgren [2]
Vladimir Brajkovic [3] Vlatka Cubric-Curik [3]
Gregor Gorjanc [4]

March 27, 2020

[1] Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[2] School of Mathematics, University of Edinburgh, Edinburgh, UK
[3] Department of Animal Science, Faculty of Agriculture, University of Zagreb, Zagreb, Croatia
[4] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Edinburgh, UK

Abstract

This paper introduces a hierarchical model to estimate haplotype effects based on phylogenetic relationships between haplotypes and their association with observed phenotypes. In a population there are usually many, but not all possible, distinct haplotypes and few observations per haplotype. Further, haplotype frequencies tend to vary substantially − few haplotypes have high frequency and many haplotypes have low frequency. Such data structure challenge estimation of haplotype effects. However, haplotypes often differ only due to few mutations and leveraging these similarities can improve the estimation of haplotype effects. There is extensive literature on this topic. Here we build on these observations and develop an autoregressive model of order one that hierarchically models haplotype effects by leveraging phylogenetic relationships between the haplotypes described with a directed acyclic graph. The phylogenetic relationships can be either in a form of a tree or a network, and we therefore refer to the model as the haplotype

network model. The haplotype network model can be included as a component in a phenotype model to estimate associations between haplotypes and phenotypes. The key contribution of this work is that by leveraging the haplotype network structure we obtain a sparse model, and by using hierarchical autoregression the flow of information between similar haplotypes is estimated from the data. We show with a simulation study that the hierarchical model can improve estimates of haplotype effects compared to an independent haplotype model, especially when there are few observations for a specific haplotype. We also compared our model to a mutation model and observed comparable performance, although the haplotype model has the potential to capture background specific effects. We demonstrate the model with a case study of modeling the effect of mitochondrial haplotypes on milk yield in cattle. We provide R code to fit the model with the R INLA package.

1 Introduction

This paper develops a hierarchical model to estimate haplotype effects based on phylogenetic relationships between haplotypes and their association with observed phenotypes. With current technology we can readily obtain genome-wide information about an individual, either through single-nucleotide polymorphism array genotyping or sequencing platform. Since the genome-wide information has become abundant, modeling this data has become the standard in animal and plant breeding as well as human genetics. The application of this modeling has been shown to improve genetic gains in breeding (Meuwissen et al., 2001; Hickey et al., 2017; Ibanez-Escriche and Simianer, 2016) and has potential for personalized prediction in human genetics and medicine (Begum, 2019; de los Campos et al., 2018; Lello et al., 2018; Maier et al., 2018).

Geneticists aim to infer which mutations are causing variation in phenotypes and what are their effects. This aim is nowadays approached with genome-wide association studies of regressing observed phenotypes on mutation genotypes (see the recent review of Morris and Cardon, 2019). However, mutations arise on specific haplotypes passed between generations, which limits accurate estimation due to low frequency of mutations, correlation with other mutations and limited ability to observe all mutations with a used genomic platform (e.g., Gibson, 2018; Simons et al., 2018; Uricchio, 2019). Further, most mutations do not affect a phenotype,

while some mutations have background (haplotype) specific effects (e.g., Chandler et al., 2017; Wojcik et al., 2019; Steyn et al., 2019).

Instead of focusing on mutation effects we here focus on haplotype effects and their differences to estimate the effect of mutations on specific haplotypes. There is extensive literature on estimating haplotype effects (Balding, 2006; Thompson, 2013; Morris and Cardon, 2019). One issue with estimating haplotype effects is that there is usually an uneven distribution of haplotypes in a population (Ewens, 1972, 2004; Walsh and Lynch, 2018), and estimating the effects of rare haplotypes is equally challenging as estimating the effect of rare mutations. However, the described genetic processes in the previous paragraph create a "network" of haplotypes (sometimes referred to as *genealogy* or *phylogeny*), which suggests that effects of similar haplotypes are similar. This observation inspired Templeton et al. (1987) to cluster phylogenetically similar haplotypes. Others have used similar approaches to utilize this data structure (Balding, 2006; Thompson, 2013; Morris and Cardon, 2019).

We here approach the problem of estimating haplotype effects by leveraging phylogenetic relationships between haplotypes described with a directed acyclic graph (DAG) (Koller and Friedman, 2009), and develop a hierarchical model of haplotype effects on this graph. We were inspired by recent advances in building phylogenies on large data sets (Kelleher et al., 2019), and aimed to develop a hierarchical model that could scale to a large number of haplotypes. Our work extends the phylogenetic mixed modeling of the whole genome (Lynch, 1991; Pagel, 1999; Housworth et al., 2004; Hadfield and Nakagawa, 2010) to a specific region. This region specific modeling could be applied either across species (macroevolution) or within a species (microevolution).

A potentially important modeling aspect with respect to across and within species modeling is that the phylogenetic mixed model assumes Brownian motion for evolution of phenotypes along a phylogeny (Felsenstein, 1988; Huey et al., 2019). Brownian motion is a continuous random-walk process with variance that grows over time (is non-stationary) (Blomberg et al., 2019; Gardiner, 2009), which makes it a plausible model of evolution due to mutation and drift. There are alternatives to Brownian motion, in particular the Ornstein-Uhlenbeck process that can accommodate various forms of selection (Lande, 1976; Hansen and Martins, 1996; Martins and Hansen, 1997; Paradis, 2014). The Ornstein-Uhlenbeck process is also a

continuous random-walk, but with an additional parameter that reverts the process to the mean (is a stationary process; e.g., Gardiner (2009); Blomberg et al. (2019)). Both of these models imply Gaussian distributions for the initial state and increments. The differences between the two processes might be important in the context of modeling haplotypes that likely manifest less variation than whole genomes, particularly when considering haplotypes within a species or even a specific population.

The aim of this paper is to develop a hierarchical model for haplotype effects by leveraging phylogenetic relationships between haplotypes. We assume that such relationships are encoded with a DAG and therefore call the model the haplotype network model. Since haplotypes differ due to a small number of mutations and very few mutations have an effect, we expect that phylogenetically similar haplotypes will have similar effects. Furthermore, the small discrete number of mutation differences suggest discrete-time analogues of Brownian and Ornstein-Uhlenbeck processes. Therefore, we have modeled the effect of a mutated haplotype given its parental haplotype with a stationary autoregressive model of order one following the phylogenetic structure encoded with a DAG. The results show that the haplotype network model improves the estimation of haplotype effects compared to an independent haplotype model due to sharing of information. The results also show that it is comparable to a mutation model, but as we discuss it has a potential to capture background specific effects.

2 Material and Methods

In this section we present the haplotype network model and show how to use it as a component in a phenotype model. We also describe simulations, a case study of modeling mitochondrial effects on milk yield in cattle, and the chosen method to perform inference and model evaluation.

2.1 The haplotype network model

Here, we present the haplotype network model, which is a hierarchical model for haplotype effects based on phylogenetic relationships between haplotypes encoded with a DAG. The phylogenetic relationships can be either in a form of a tree or a more general network. We also present two

generalizations of the model − first due to multiple parental haplotypes and second due to genetic recombination.

We assume throughout that the phylogeny between haplotypes is known and that it can be encoded with a DAG. The haplotype network model can in principle deal with different types of mutations, but for simplicity we focus only on biallelic mutations with the code 0 used for the ancestral/reference allele (commonly at a higher frequency in a population), and the code 1 used for the alternative allele that arose due to a mutation.

2.1.1 Motivating example

To motivate the haplotype network model, we use the example from Kelleher et al. (2019) that presents 5 haplotypes spanning 7 biallelic polymorphic sites (Table 1). Note that the 5 haplotypes are just a sample of the $2^7 = 128$ possible haplotypes over the 7 sites. An example of a phylogeny for the haplotypes is shown in Figure 1, where haplotypes are denoted as nodes (we also show their allele sequence), relationships between haplotypes are denoted as edges, and mutated sites are denoted with a number on edges. For example, the ancestral haplotype i has allele sequence `0000000`, and the haplotype g with sequence `1000100` differs from the ancestral haplotype due to mutations at the sites 5 and 1.

Assuming that similar haplotypes have similar effects, we model dependency between parent-progeny pairs of haplotypes with an autoregressive Gaussian process of order one. For the haplotypes in Table 1 and Figure 1, this model implies the following set of conditional dependencies

$$\begin{array}{lcl}
h_i & \sim & N\left(0, \sigma^2_{h_m}\right) \\
h_{g'}|h_i & \sim & N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_g|h_{g'} & \sim & N(\rho h_{g'}, \sigma^2_{h_c}) \\
h_a|h_g & \sim & N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_f, h_b, h_c, |h_g & \sim & N\left(\rho h_g, \sigma^2_{h_c}\right) \\
h_{h'}|h_i & \sim & N\left(\rho h_i, \sigma^2_{h_c}\right) \\
h_h, h_d|h_{h'} & \sim & N\left(\rho h_{h'}, \sigma^2_{h_c}\right) \\
h_e|h_h & \sim & N\left(\rho h_h, \sigma^2_{h_c}\right)
\end{array}$$

where $h_i, h_g, \ldots, h_a$ indicate the effect of haplotypes $i, g, \ldots, a$ and $h_{*'}$ indicates the effect of haplotypes that occur between haplotypes separated by multiple mutations, for example, $g^{'}$ is the additional haplotype between

the haplotypes i and g due to two mutations between i and g. We describe the other model parameters $\left(\rho, \sigma^2_{h_m}, \sigma^2_{h_c}\right)$ in the following.

Table 1: Example of 5 haplotypes spanning 7 mutations from Kelleher et al. (2019). The ancestral (reference) alleles are coded as 0 and alternative alleles are coded as 1

		Site						
		1	2	3	4	5	6	7
Haplotype	a	1	0	0	1	1	0	0
	b	1	0	0	0	1	1	0
	c	1	0	0	0	1	1	0
	d	0	1	0	0	0	0	1
	e	0	1	1	0	0	0	1

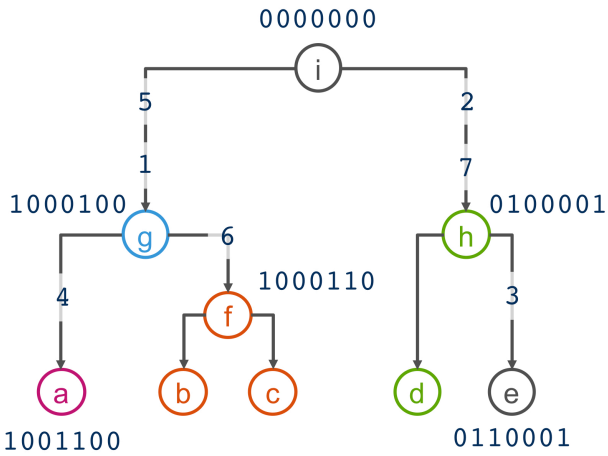


Figure 1: Phylogenetic relationship of haplotypes in Table 1

2.1.2 The model

Assume a known general phylogenetic network of haplotypes described with a DAG with haplotype effects as nodes and relationships between the haplotype effects as edges, such as in Figure 1. We model the effect of a

chosen "starting" (this could be either a central, ancestral, most common or some other haploptype) haplotype 0 with mean zero and marginal variance $\sigma^2_{h_m}$ as

$$h_0 \sim \mathcal{N}(0, \sigma^2_{h_m}), \tag{1}$$

and any other haplotype j in the phylogenetic network as a function of its one-mutation-removed parental haplotype $p\,(j)$ assuming the autoregressive Gaussian process of order one with the autocorrelation between haplotype effects of ρ ($|\rho| < 1$) and conditional variance of $\sigma^2_{h_c}$ as

$$h_j | h_{p(j)} \sim \mathcal{N}(\rho h_{p(j)}, \sigma^2_{h_c}) \tag{2}$$

We consider the autoregressive Gaussian process of order one that is stationary both in mean and variance, which is achieved by setting the marginal variance to $\sigma^2_{h_m} = \sigma^2_{h_c}/(1-\rho^2)$, so $\sigma^2_{h_c} = \sigma^2_{h_m}(1-\rho^2)$. This is the standard autoregressive model of order one used in time-series analysis (Rue and Held, 2005). The difference here is that we are applying the model onto a phylogenetic network described with a DAG or a tree (Basseville et al., 1992; Datta et al., 2019; Wu et al., 2020).

The set of distributions in (1) and (2) give a system of equations for all n haplotype effects $\boldsymbol{h} = (h_1, ..., h_n)^T$ as

$$\boldsymbol{h} = \boldsymbol{T}\,(\rho)\,\boldsymbol{\varepsilon}, \tag{3}$$

$$\boldsymbol{T}\,(\rho)^{-1}\,\boldsymbol{h} = \boldsymbol{\varepsilon}, \tag{4}$$

where the matrices $\boldsymbol{T}\,(\rho)$ and $\boldsymbol{T}\,(\rho)^{-1}$ respectively represent marginal and conditional phylogenetic regression between haplotype effects $\boldsymbol{h}$, and the vector $\boldsymbol{\varepsilon}$ represents haplotype effect deviations, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{D}\,(\rho)\,\sigma^2_{h_c})$. The expression $\boldsymbol{T}\,(\rho)$ indicates that the matrix $\boldsymbol{T}$ depends on the value of ρ. Since haplotype effect deviations are independent, the matrix $\boldsymbol{D}\,(\rho)$ is diagonal and has value $1/(1-\rho^2)$ for the "starting" haplotype and 1 for the other haplotypes. Following the assumed autoregressive process of order one (2), the non-zero elements of $\boldsymbol{T}\,(\rho)^{-1}$ are 1 along the diagonal and $-\rho$ between a haplotype effect (row index) and its parental haplotype effect (column index). This simple sparse lower-triangular structure of the matrix $\boldsymbol{T}\,(\rho)^{-1}$ arises from the Markov properties of the autoregressive process (Rue and Held, 2005).

From (3) covariance between haplotype effects are

$$\begin{aligned}\mathrm{Var}(\boldsymbol{h}) &= \mathrm{Var}(\boldsymbol{T}(\rho)\,\boldsymbol{\varepsilon}), &(5)\\ &= \boldsymbol{T}(\rho)\,\mathrm{Var}(\boldsymbol{\varepsilon})\boldsymbol{T}(\rho)^T = \boldsymbol{T}(\rho)\,\boldsymbol{D}(\rho)\,\boldsymbol{T}(\rho)^T\,\sigma^2_{h_c} &(6)\\ &= \boldsymbol{H}(\rho)\,\sigma^2_{h_c} = \boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right) &(7)\end{aligned}$$

The covariance expression (5) shows that haplotype covariances $\boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)$ depend on the autocorrelation and variance parameters, while the covariance coefficients $\boldsymbol{H}(\rho)$ depend only on the autocorrelation parameter. So, the variance parameter is capturing scale (spread) of effects and the autocorrelation parameter is capturing the dependency structure. Note that these two parameters are dependent by definition $\sigma^2_{h_c} = \sigma^2_{h_m}(1-\rho^2)$. When $\rho = 0$ there is no covariance between haplotype effects due to phylogenetic relationships, which suggests a model where haplotype effects are identically and independently distributed, $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{h_m})$. When $\rho \neq 0$ effects of phylogenetically related haplotypes co-vary due to shared mutations.

For completeness, the joint density of all n haplotype effects $\boldsymbol{h}$ is multivariate Gaussian

$$\boldsymbol{h}|\rho, \sigma^2_{h_c} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)), \tag{8}$$

with the probability density function

$$p(\boldsymbol{h}|\rho, \sigma^2_{h_c}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n}_{h_c}(1-\rho^2)^{1/2}\exp\left(-\frac{1}{2\sigma^2_{h_c}}\boldsymbol{h}^T\boldsymbol{H}(\rho)^{-1}\boldsymbol{h}\right) \tag{9}$$

The expression (9) involves inverse of the covariance coefficient (precision) matrix $\boldsymbol{H}(\rho)^{-1}$, which we can obtain without computationally expensive inverse of the $\boldsymbol{H}(\rho)$ (5). Following the definition (5), inverting both sides and using the described structure of $\boldsymbol{T}(\rho)^{-1}$ available from the DAG and $\boldsymbol{D}(\rho)$, we can efficiently get this inverse by

$$\boldsymbol{H}(\rho)^{-1} = \frac{1}{\sigma^2_{h_c}}\boldsymbol{T}(\rho)^{-1^T}\boldsymbol{D}(\rho)^{-1}\boldsymbol{T}(\rho)^{-1}. \tag{10}$$

Inspection of the structure of (10) shows that this is a very sparse matrix with a structure. We can compute the non-zero elements directly with the following simple algorithm where we loop over all haplotypes:

if the haplotype is the "starting" haplotype **then**
  add $1 - \rho^2$ to the diagonal element
else
  add 1 to the diagonal element
end if
if the haplotype has a parental haplotype **then**
  set off-diagonal element between the haplotype and its parental haplotype to $-\rho$
  add ρ^2 to the diagonal element of the parental haplotype
end if

To fully specify the model for $\boldsymbol{h}$ (8), prior distributions must be assigned to the autocorrelation parameter ρ, and the marginal variance $\sigma^2_{h_m}$ or the conditional variance $\sigma^2_{h_c}$. Because most mutations do not have an effect we can expect that most parent-progeny pairs of haplotypes will have similar effects, which suggests that the autocorrelation parameter will be close to 1. This knowledge can be incorporated in the prior distribution for ρ. For the variance parameters there may be some prior knowledge about the size of haplotype effects relative to other effects, which can also be taken into account when choosing the prior distribution. We specify prior distributions for these parameters in later sections.

2.1.3 Multiple parental haplotypes

Sometimes phylogenetic inference cannot resolve bifurcating trees with dichotomies (one parental haplotype and two progeny haplotypes), and outputs a multifurcating tree with polytomies (one parental haplotype and multiple progeny haplotypes) or even just a network (multiple parent haplotypes and multiple progeny haplotypes, e.g., Schliep et al. (2017); Uyeda et al. (2018)). The multiple progeny case works out of the box with the initial model, and we here present an extension of the model presented in Section 2.1.2 "The model" that can accommodate the multiple parent haplotypes and multiple progeny haplotypes case where the trees or networks can be described with a DAG.

We assume that the effects of all ancestral haplotypes, the haplotypes at the top of the network, are independent and come from the same Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2_{h_m})$. We further assume conditional independence between a haplotype and all previous haplotypes in the network given the parents of that haplotype. In the model where each haplotype had only

a single parent haplotype it was assumed that the haplotype effect was ρ times the parental haplotype effect plus some Gaussian noise. When a haplotype has multiple parents, we now assume that the effect is the average over each of these processes from each parental haplotype.

We illustrate this with a small example which implies the model construction used. Let haplotype d have parental haplotypes a, b, and c. We denote the contribution from each of these parents h_{d_a}, h_{d_b}, h_{d_c}, and assume

$$\begin{aligned} h_{d_a} =& \rho h_a + \varepsilon_{d_a} \\ h_{d_b} =& \rho h_b + \varepsilon_{d_b} \\ h_{d_c} =& \rho h_c + \varepsilon_{d_c} \end{aligned}$$

where $(\varepsilon_{d_a}, \varepsilon_{d_b}, \varepsilon_{d_c})^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2_{h_c})$. Further, we assume that the resulting effect of haplotype h_d is the average over all parent processes

$$h_d = \frac{\rho}{3}(h_a + h_b + h_c) + \frac{1}{3}(\varepsilon_{d_a} + \varepsilon_{d_b} + \varepsilon_{d_c})$$

The distribution of h_d conditional on h_a, h_b and h_c becomes

$$h_d | h_{d_a}, h_{d_b}, h_{d_c} \sim \mathcal{N}\left(\frac{\rho}{3}(h_a + h_b + h_c), \frac{\sigma^2_{h_c}}{3}\right)$$

In general this means that $h_i | h_1, ..., h_k \sim \mathcal{N}(\frac{\rho}{k}\sum_{j=1}^{k} h_j, \frac{\sigma^2_{h_c}}{k})$, for haplotype i with parental haplotypes $1, ..., k$. This model construction corresponds to a model where one first takes every path down through the DAG and assigns separate stationary autoregressive processes of order 1 to each such path, and then assumes conditionally independent but identical autoregressive processes of order 1, that is, the processes have the same parameters.

Multiple parental haplotypes change the structure of the $\boldsymbol{T}(\rho)^{-1}$ matrix to having $-\rho/k_i$ value between a haplotype effect (row index) and its parental haplotype effect (column index), and the $\boldsymbol{D}(\rho)^{-1}$ matrix diagonals for "non-starting" haplotypes to k_i, where k_i is the number of parental haplotypes of the haplotype i. The algorithm to setup the $\boldsymbol{H}(\rho)^{-1}$ matrix is then, looping over all haplotypes:

if the haplotype is a "starting" haplotype **then**
add to the diagonal element $1 - \rho^2$
else
add k_i to the diagonal element
end if
if the haplotype has parental haplotype(s) **then**
set off-diagonal element between the haplotype and its parental haplotype to $-\rho$
set off-diagonal elements between all parental haplotypes that share that progeny haplotype to ρ^2/k_i
add ρ^2/k_i to the diagonal element of the parental haplotype(s)
end if

The model presented in this section is a straightforward model, and is only one of many possible choices for a model accommodating multiple parental haplotypes. We are only presenting one option for allowing such graph structures in the model, other choices should also be explored.

2.1.4 Expanding to multiple regions due to recombination

The haplotype phylogeny can differ along genome regions due to recombination − the process of swapping genome regions between haplotypes during meiosis. We accommodate this in the haplotype network model by considering each haplotype region separately, but still within the framework of the same model. This means that the effect of haplotype h_i is modeled as the sum of effects for all haplotype regions. Consider haplotypes spanning three regions. The effect of haplotype i, is then assumed to be the sum of the effects of haplotype segments in each of the three regions

$$h_i = h_{1,i} + h_{2,i} + h_{3,i}$$

We assume the haplotype network model for each haplotype region, but with joint hyper-parameters $\left(\rho, \sigma^2_{h_c}\right)$. Let $\boldsymbol{h} = (h_{1,1}, \ldots, h_{1,n_1}, h_{2,1}, \ldots, h_{m,n_m})$ be the effect of all haplotypes in all regions, where m is the number of regions and n is the number of haplotypes in each region. For $\boldsymbol{h}$ we then assume

$$p(\boldsymbol{h}|\rho, \sigma^2_{h_c}) = \left(\frac{1}{\sqrt{2\pi}}\right)^{n_1+\ldots+n_m} \sigma_{h_c}^{-(n_1+\ldots+n_m)} (1-\rho^2)^{\frac{m}{2}} \exp\left(-\frac{\boldsymbol{h}^T \boldsymbol{H}\,(\rho)^{-1} \boldsymbol{h}}{2\sigma^2_{h_c}}\right),$$

with

$$\boldsymbol{H}(\rho)^{-1} = \begin{pmatrix} \boldsymbol{H}(\rho)_1^{-1} & & \\ & \ddots & \\ & & \boldsymbol{H}(\rho)_m^{-1} \end{pmatrix} \tag{11}$$

Although recombination is common, we have focused on the special case of no recombination in this study, where the haplotypes are connected in one phylogeny as presented in Section 2.1.2 "The model".

2.2 Phenotype model with haplotype effects

We now show how the haplotype effects can be included in a model for phenotypic observations. We also present a phenotype model that includes independent haplotype effects or mutation effects rather than the haplotypes.

Let $\boldsymbol{y}_{p\times 1}$ be phenotype observations of p individuals and let $\boldsymbol{h}_{n\times 1}$ be the effect of n haplotypes obtained from phasing genotypic data of the individuals. We assume the following model for the centered and scaled phenotypic observations

$$\boldsymbol{y}_{p\times 1} = \boldsymbol{X}_{p\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{f}^1_{p\times 1} + ... + \boldsymbol{f}^s_{p\times 1} + \boldsymbol{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{12}$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}1000)$ is a vector of r fixed effects with covariate matrix $\boldsymbol{X}$, $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2_f)$ are random effects, $\boldsymbol{h}$ are the haplotype effects with incidence matrix $\boldsymbol{Z}$ that maps haplotypes to individuals, and the residual effect is $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2_e)$. In the case of diploid individuals there will be two entries in every row of $\boldsymbol{Z}$, and a single entry for haploid individuals, male sex chromosome or mitogenome.

For the haplotype effects $\boldsymbol{h}$ we will assume three models. The first is a base model with independent haplotype effects (IH model), where $\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma^2_{h_m}\right)$. The second is the haplotype network model presented in Section 2.1.2 "The model" (HN model), where $\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)\right)$. The third is an alternative way of estimating haplotype effects using a linear combination of mutation effects (mutation model). Assume $\boldsymbol{h} = \boldsymbol{U}\boldsymbol{v}$ with $\boldsymbol{v} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma^2_v\right)$ being mutation effects, and $\boldsymbol{U}$ is the matrix containing the haplotype allele sequences with reference alleles coded as 0 and alternative alleles coded as 1.

The models do not have a common intercept because a common intercept and the mean level in the haplotype effects are not identifiable when ρ approaches 1. Instead the mean level in the observations is captured by the haplotype effects. A sum-to-zero constraint can be specified for the haplotype network part of the model if a common intercept is required, but changes the model interpretation if ρ is close to 1. This problem is not special for the haplotype network model, but occurs for all autoregressive models when they are used as part of a structured mixed effects model. When the goal is to make predictions about the haplotype effects, this model choice will not influence the prediction results.

2.2.1 Prior distributions

We assigned penalized complexity (PC) prior distributions to the variances of random effects and the autocorrelation parameter. PC priors are proper prior distributions developed by Simpson et al. (2017), that penalize increased complexity induced by deviation from a simpler base model to avoid over-fitting. For a random effect with a variance parameter the base models is a model where the variance of this random effect is zero. For the autoregressive model of order one we have assumed a base model with $\rho = 1$. We could have assumed a base model with $\rho = 0$, but it is more likely that phylogenetically similar haplotypes have similar effects. The penalized complexity prior can be specified through a quantile u and a probability α which satisfy $\text{Prob}(x > u_x) = \alpha_x$ for the parameter x.

Although the precision matrix is specified with the conditional variance (10), the prior is specified for the marginal variance since we often have a better intuition for the marginal variance than the conditional variance. Specifically, we specify the prior for the marginal standard deviation σ_{h_m}, and assume the conditions $u_{\sigma_{hm}} > 0$ and $0 < \alpha_{\sigma_{hm}} < 1$. For the autocorrelation parameter we use the PC prior developed for stationary autoregressive processes (Sørbye and Rue, 2017) with base model at $\rho = 1$, and parameters satisfying $-1 < u_\rho < 1$ and $\sqrt{(1 - u_\rho)/2} < \alpha_\rho < 1$. We highlight that the prior by Sørbye and Rue (2017) was developed for a stationary autoregressive process with different model assumptions than the models presented in this paper. Ideally, the prior for the autoregressive parameter would be tailored to the haplotype network model.

2.3 Inference and evaluation

In this section we describe the used method for statistical inference; the integrated nested Laplace approximations (INLA), and the used methods for evaluating model fit in the simulation study.

2.3.1 Inference

All models in this study fit in the framework of hierarchical latent Gaussian models, which makes INLA (Rue et al., 2009) a suitable choice to perform inference as implemented in the R (R Core Team, 2018) package `INLA` (available at `www.r-inla.org`). In this section we give a brief introduction to latent Gaussian models and how INLA are used to approximate the marginal posterior distributions in such models. For an in-depth description of INLA see Rue et al. (2009), Blangiardo and Cameletti (2015), and Rue et al. (2017).

The class of latent Gaussian models includes several models, for example generalized linear (mixed) models, generalized additive (mixed) models, spline smoothing methods, and the models presented in this article. Latent Gaussian models are hierarchical models where observations $\boldsymbol{y}$ are assumed to be conditionally independent given a latent Gaussian random field $\boldsymbol{x}$ and hyper-parameters $\boldsymbol{\theta}_1$, meaning $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1) \sim \Pi_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta}_1)$. The latent field $\boldsymbol{x}$ includes both fixed and random effects and is assumed to be Gaussian distributed given hyper-parameters $\boldsymbol{\theta}_2$, that is $p(\boldsymbol{x}|\boldsymbol{\theta}_2) \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{\Sigma}(\boldsymbol{\theta}_2))$. The parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ are known as hyper-parameters and control the Gaussian field and the likelihood for the data. These are usually variance parameters for simple models, but can also include other parameters, for example the ρ parameter in the HN model. We must also assign prior distributions to the hyper-parameters to completely specify the model.

The main aim of Bayesian inference is to estimate the marginal posterior distribution of the variables of interest, that is, $p(\theta_j|\boldsymbol{y})$ for hyper-parameters and $p(x_i|\boldsymbol{y})$ for the latent field. INLA computes approximations to these densities fast and with high accuracy. The INLA methodology is based on numerical integration of non-Gaussian hyper-parameters and utilizing Markov properties of the Gaussian parameters. Hence, for the computations to be both fast and accurate, the latent Gaussian models have to satisfy some assumptions. The number of non-Gaussian hyper-parameters $\boldsymbol{\theta}$ should be low, typically less than 10, and not exceeding 20. Further, the

latent field should not only be Gaussian, it should be a Gaussian Markov random field. The conditional independence property of a Gaussian Markov random field yields sparse precision matrices which makes computations in INLA fast due to the use of efficient algorithms for sparse matrices. Lastly, each observation y_i should depend on the latent Gaussian field only through one component x_i.

The R package INLA is run using the `inla()` function with three mandatory arguments: a data frame or stack object containing the data, a formula much like the formula for the standard `lm()` function in R, and a string indicating the likelihood family. Prior distributions for the hyper-parameters are specified through additional arguments. Several tools to manipulate models and likelihoods exist as described in tutorials at `www.r-inla.org`, and in the books by Blangiardo and Cameletti (2015), and Krainski et al. (2018). In the supplementary section, we have included a script showing how we simulated the data from the haplotype network model and how we fitted the model to the data.

2.3.2 Evaluation of model performance

We evaluated the predictive performance of the models using the continuous rank probability score (CRPS) (Gneiting and Raftery, 2007). The CRPS is a proper score which takes into account the whole posterior distribution, meaning that it compares the whole estimated posterior distribution for haplotype effects with the true value, and with this, accounts for the uncertainty in estimation. The CRPS is negatively oriented, so the smaller the CRPS the closer the posterior distribution is to the true value. The full Bayesian posterior output from `inla()` for these models are mixtures of Gaussians, for which there is no closed form expression for CRPS. The mixtures here are similar to plain Gaussians, so we approximate the exact CRPS with the Gaussian CRPS using only the posterior mean and variances provided in the results.

We calculated the CRPS for estimated haplotype effects with the IH, HN and mutation models. To ease the comparison we have then calculated a relative CRPS (RCRPS) score as the log of the ratio between the averages of the CRPS from the HN model and IH model, and correspondingly for

the mutation model relative to the IH model. The score is computed as

$$\log\left(\frac{\sum_{i=1}^{n}\mathrm{CRPS}(\hat{h}_i)_{HN}}{\sum_{i=1}^{n}\mathrm{CRPS}(\hat{h}_i)_{IH}}\right),$$

where $\mathrm{CRPS}(\hat{h}_i)_{HN}$ is the CRPS of the posterior distribution for haplotype effect of haplotype h_i with the HN model. We will refer to this score as the RCRPS.

We also calculated the root mean square error (RMSE) between the mean posterior haplotype effect and true haplotype effects, but the results for the relative RMSE and RCRPS were qualitatively the same. We decided to present the RCRPS results because they take into account the whole posterior distribution, whereas the RMSE only takes into account how closely the posterior mean is to the true effect.

In addition to comparing the haplotype estimates, we compared the estimated mutation effects from the HN model and the mutation model, using the RCRPS (HN model versus mutation model). Although the HN model estimates the haplotype effects $\boldsymbol{h}$, we can obtain mutation effects via $\boldsymbol{v} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{h}$. We could also obtain mutation effects through linear combinations of haplotype effects.

2.4 Simulation study

To test the proposed HN model, we used simulated data. Here, we present data simulated from two different models − the HN model with varying degree of autocorrelation, and a more realistic mutation model where only some mutations have causal effect. We also present the models that were fitted to the simulated data, and how the model fit was evaluated. In the supplemental data (Supplemental 1), we provide an R script and the data file to simulate from and fit the haplotype network model.

2.4.1 Simulation from the haplotype network model

We used the coalescent simulator `msprime` (Kelleher et al., 2016) to simulate the phylogeny shown in Figure 2 with $n = 107$ unique haplotypes. We then simulated phenotypes $\boldsymbol{y}$ for $p = 400$ individuals from the model

$$\boldsymbol{y}_{p\times 1} = \boldsymbol{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{13}$$

where $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)\right)$ with $\boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)$ built from the DAG describing the phylogeny (Figure 2 , (5)), and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2_e)$.

We tested 15 parameter sets, from weak to strong haplotype effect dependency, and from low to high residual variance relative to the conditional haplotype variance

$$\rho = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\sigma^2_e/\sigma^2_{hc} = \{0.5, 1, 2\}$$

We simulated a haploid system for simplicity, so the incidence matrix $\boldsymbol{Z}$ was a zero matrix with a single 1 on each row indicating which individuals had which haplotype. We were particularly interested in estimating the haplotype effect with few or no direct phenotype observations. This is the extreme scenario where the haplotype network model could be beneficial. To achieve this, we designed the incidence matrix to create two different scenarios. In the first scenario, all haplotypes had associated phenotype observation, but some haplotypes only had one observation. We assigned a random sample of 15% of the haplotypes only to one individual each and the rest of the haplotypes randomly to the remaining individuals. In the second scenario, some haplotypes did not have phenotype observations. We selected a random sample of 15% of the haplotypes that did not have phenotype observations and assigned phenotype observations to the rest of the haplotypes randomly.

2.4.2 Simulation from the mutation model

We also simulated haplotype effects from a mutation model using the same phylogeny as in the previous section, shown in Figure 2, and using $p = 400$ individuals. For the 107 unique haplotypes we had 106 mutations in the haplotypes. We used the variants at these mutations to simulate haplotype effects and phenotypes according to the model

$$\boldsymbol{y}_{p\times 1} = \boldsymbol{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{14}$$

where $\boldsymbol{h} = \boldsymbol{U}_{n\times 106}\boldsymbol{v}_{106\times 1}$, $\boldsymbol{v}$ was the mutation effect, $\boldsymbol{U}$ a matrix containing ancestral (reference) alleles coded as zero and alternative alleles coded as 1, and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma^2_e)$. We sampled the mutation effect v from

$$v = \begin{cases} \mathcal{N}(0, \sigma^2_v), \text{ with probability } \lambda \\ 0, \text{ with probability } (1-\lambda) \end{cases}$$

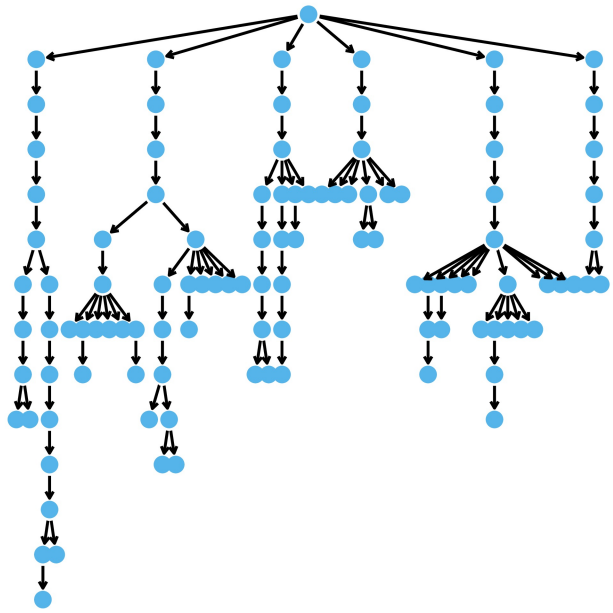


Figure 2: The DAG describing the phylogeny of simulated haplotypes

where we chose σ_v^2 so that the empirical variance of $\boldsymbol{h}$, $\mathrm{Var}(\boldsymbol{h})$, was 1.

Again, we tested 15 parameter sets, from few to many causal variants, and from low to high residual variance relative to empirical haplotype variance

$$\lambda = \{0.1, 0.3, 0.5, 0.7, 0.9\},$$
$$\sigma_e^2/\mathrm{Var}(\boldsymbol{h}) = \{0.5, 1, 2\}$$

We simulated haploid individuals, so the incidence matrix $\boldsymbol{Z}$ was a zero matrix with a single 1 on each row indicating which individuals had which haplotype. The incidence matrix was designed to create the same scenarios as for the data simulated from the HN model in Section 2.4.1 “Simulation from the haplotype network model”.

2.4.3 Models fitted to the simulated data

We fitted the HN model, IH model and the mutation model to the simulated data

$$\boldsymbol{y}_{p\times 1} = \boldsymbol{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1}, \tag{15}$$

where $\boldsymbol{h}$ was assumed to be distributed according to:

$\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_h\left(\rho, \sigma^2_{h_c}\right)\right)$ for the HN model,

$\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma^2_I\right)$ for the IH model and,

$\boldsymbol{h} = \boldsymbol{U}\boldsymbol{v}$, $\boldsymbol{v} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma^2_v\right)$ for the mutation model.

The residual effect was $\boldsymbol{e} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}\sigma^2_e\right)$. We assigned PC priors to the ρ parameter with $u_\rho = 0.7$ and $\alpha_\rho = 0.8$, and to all variance parameters with $u = 0.1$ and $\alpha = 0.8$.

2.4.4 Evaluation

For each combination of phenotype observation distribution across haplotypes, proportion of residual variance relative to the haplotype variance, and ρ or λ parameters, we performed the same experiment 50 times. In 4% of the experiments when the data were simulated from the HN model, the inference method was not able to fit the HN model, and we report results only for cases where all models were successfully fitted. There was no trend for any parameter set in particular causing the inference method to break down.

Since we created different scenarios for how phenotype observations were distributed among the haplotypes, we stratified the results for haplotype effects based on how many times a haplotype was observed in a phenotyped individual. For the first scenario, where some haplotypes were phenotyped either once or multiple times, we computed the RCRPS for these two groups separately. For the second scenario, where some haplotypes were not phenotyped, we present the RCRPS only for haplotypes that were not phenotyped. In both cases, RCRPS less than zero indicates that the HN/mutation model was better than the IH model on average. We present the RCRPS for estimated mutation effects only for the data simulated from the mutation model, because the true mutation effects were not generated when simulating from the haplotype network model.

2.5 Case study: Mitochondrial haplotypes in cattle

In this section we present a case study using the haplotype network model to estimate the effect of mitochondrial haplotypes on milk yield in cattle. We first describe the data and then the fitted model.

2.5.1 Data

We demonstrate the use of the haplotype network model with a case study estimating the effect of mitochondrial haplotypes on milk yield in cattle from Brajković (2019). We chose this case study because mitochondrial haplotypes are passed between generations without recombination and are as such a good case for the haplotype network model. The phenotyped data comprised of information about the first lactation milk yield, age at calving, county, and herd-year-season of calving for 381 cows. Additionally, the data comprised of pedigree information with 6336 individuals (including the 381 cows) and information about mitochondrial haplotypes (whole mitogenome) variation between maternal lines in the pedigree. We inferred the mitochondrial haplotypes by first sequencing mitogenome, aligning it to the reference sqeuence and calling haplotype mutations as described in detail in Brajković (2019). We used PopART (Leigh and Bryant, 2015) to build a phylogentic network of mitochondrial haplotypes. For simplicity we used the median-joining method to show that the haplotype network model can be fit to the output of a standard phylogenetic method. In this process we assumed that the ancestral alleles were the most frequent alleles. The phylogeny contained 63 unique mitochondrial haplotypes each separated by one mutation. Of the 63 haplotypes only 16 haplotypes were observed in the 381 phenotyped cows. There were five haplotypes that did not have a parent haplotype, meaning we treated them as "starting" haplotypes in the haplotype network model.

2.5.2 Model

Let $\boldsymbol{h}_{n\times 1}$ be the effect of the $n = 63$ mitochondrial haplotypes, and let $\boldsymbol{y}_{p\times 1}$ be the phenotypes of the $p = 381$ cows. We fitted the following model to centered and scaled phenotypes being first lactation milk yield

$$\boldsymbol{y}_{p\times 1} = \boldsymbol{X}_{p\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{c}_{p\times 1} + \boldsymbol{a}_{p\times 1} + \boldsymbol{Z}_{p\times n}\boldsymbol{h}_{n\times 1} + \boldsymbol{e}_{p\times 1},$$

where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}1000)$ contained effects of age at calving as a continuous covariate effect and county as a categorical covariate effect with corresponding design matrix $\boldsymbol{X}$, $\boldsymbol{c} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\sigma_c^2\right)$ was the random effect of herd-year-season of calving (contemporary group), $\boldsymbol{a} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}\sigma_a^2\right)$ was additive genetic effect for the whole nuclear genome with the covariance coefficient matrix $\boldsymbol{A}$ derived from the pedigree (Henderson, 1976; Quaas, 1988), and lastly the mitochondrial haplotype effects were fitted with the haplotype network model $\boldsymbol{h} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{V}_h\left(\rho, \sigma_{h_c}^2\right)\right)$ with the covariance matrix $\boldsymbol{V}_h\left(\rho, \sigma_{h_c}^2\right)$ derived from the phylogeny and using the expanded model that accommodates multiple parental haplotypes from Section 2.1.3 "Multiple parental haplotypes". We assumed that residuals were distributed as $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}\sigma_e^2)$. We assigned PC priors to the ρ parameter with $u_\rho = 0.7$ and $\alpha_\rho = 0.8$, to the $\sigma_{h_m}^2$ parameter with $u_{\sigma_{h_m}} = 0.1$ and $\alpha_{\sigma_{h_m}} = 0.3$, and to all remaining variance parameters with $u_{\sigma_*} = 0.1$ and $\alpha_{\sigma_*} = 0.8$.

3 Results

In this section we present results from the simulation study testing the behavior of the haplotype network model, and the case study estimating the effect of mitochondrial haplotypes on milk yield in cattle. In the results from the simulation study, we present the RCRPS between the haploptype network (HN) model and the independent haplotype (IH) model, and between the mutation model and the IH model for the different parameter sets. In the results from the case study, we present the mean and standard deviation of the posterior mitochondrial haplotype effects mapped onto the phylogenetic network, and posterior estimates for the hyper-parameters.

3.1 Simulation study results

3.1.1 Simulation from the haplotype network model

We start by considering the results with the data simulated from the HN model from Section 2.4.1 "Simulation from the haplotype network model" that were fitted with the models from Section 2.4.3 "Models fitted to the simulated data".

The RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) is presented in Figure 3. This figure has three panels denoting haplotypes that were observed in (A)

several phenotyped individuals, (B) only one phenotyped individual and (C) were not observed in a phenotyped individual. The full lines show the RCRPS between the HN model and the IH model, while the dashed lines show the RCRPS between the mutation model and the IH model. Along the x-axis the autocorrelation parameter ρ for the simulated haplotype effects increases from weak to strong phylogenetic dependency, and the three colored lines indicate the amount of phenotypic variation due to residual effects relative to the variation from haplotype effects.

In summary, Figure 3 shows that (i) the HN model outperforms the IH model across a range of model parameter values, (ii) the HN model is more important for haplotypes with fewer phenotypic observations, (iii) the HN model is more important for noisy phenotypic data than for phenotypic data with less noise, and (iv) when haplotypes are more phylogenetically dependent, the HN model and the mutation model have similar performance. We go through each of these findings in detail.

The HN model outperforms the IH model for almost all 15 parameter sets. In all panels of Figure 3, almost all points with the full line are below zero, meaning that the HN model gave better estimates of haplotype effects than the IH model. When the haplotype dependency due to phylogeny was low, the RCRPS was around zero, meaning that the two models were equal in estimating the haplotype effects, which was expected. As the phylogenetic dependency became stronger, the HN model improved relative to the IH model, as seen from the decreasing RCRPS as ρ approaches 0.9.

The improvement in RCRPS with the HN model relative to the IH model increased when haplotypes were observed in a smaller number of phenotyped individuals. This is indicated by the decreasing RCRPS when we compare panels (A), (B) and (C) in Figure 3. The panels correspond to haplotypes observed in several (A), one (B) and no (C) phenotyped individuals. The decrease in RCRPS was the largest in panel (C) followed by panel (B) and panel (A). This means that modeling phylogenetic dependency between haplotypes is most useful when there are some haplotypes with few phenotypic observations, or if we want to predict the effect of new haplotypes. Especially for haplotypes that do not have a direct link to observed phenotypes, the IH model is not useful, because it assigns the average effect of haplotypes with direct link to observed phenotypes to haplotypes without such links, whereas the HN model can assign the haplotype effect based on a phylogenetic network. When the

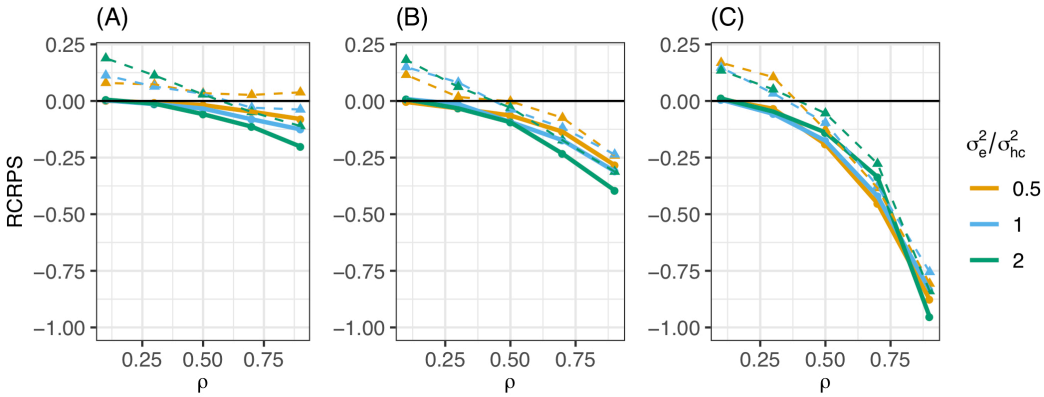


Figure 3: RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) between the HN model and the IH model (solid line) and between the mutation model and the IH model (dashed line). The data were simulated from the HN model with varying ρ parameter and ratio between the residual σ_e^2 and conditional haplotype variance $\sigma_{h_c}^2$. The three panels show RCRPS for the haplotypes that were observed in (A) several phenotyped individuals, (B) only one phenotyped individual and (C) were not observed in a phenotyped individual

haplotype effects had low phylogenetic dependency (ρ was low), there was not much difference in RCRPS between the three panels, as there were no similarities between the haplotype effects from which the HN model could learn from.

The improvement with the HN model relative to the IH model increased when the phenotypic data were noisier. In panels (A) and (B) in Figure 3, the RCRPS was lower with larger residual variance. This indicates that the HN model did a better separation of the environmental and genetic sources of variation than the IH model. Interestingly, we did not observe the same in panel (C), that is for haplotypes that did not have direct link to observed phenotypes. This was because the IH model performed equally poorly in predicting new haplotypes regardless of the amount of residual variance. The HN model on the other hand, performed slightly better as there was less variation due to residual effects for some values of ρ and

similar for other values of ρ compared to the IH model.

As haplotypes became phylogentically more dependent with the increasing ρ, the HN model and the mutation model performed similarly. In all panels, the dashed lines indicate a worse fit for the mutation model than for the IH model, and the HN model when ρ was low. When ρ increased, the mutation model improved relative to the IH model, and had a RCRPS close to the RCRPS for the HN model, but not better than the HN model.

3.1.2 Simulated data from the mutation model

In the previous section we saw that the HN model outperformed the IH model when the simulated haplotype effects were generated from the HN model itself. Now, we consider the results with the haplotype effects simulated from a more realistic mutation model from Section 2.4.2 "Simulation from the mutation model", and fitted with the models from Section 2.4.3 "Models fitted to the simulated data". Here we varied the probability of mutations having a causal effect λ, and we present results using $\lambda = 0.1$, since the results were qualitatively similar for all tested λ values.

The RCRPS is presented in Figure 4 for the three different levels of phenotype observations per haplotype and three different values of residual variance relative to the empirical haplotype variance which was always 1. The full lines show the RCRPS between the HN model and the IH model, while the dashed lines show the RCRPS between the mutation model and the IH model.

In general, the results align with the results from the previous section except for the mutation model: (i) the HN model outperforms the IH model, (ii) the HN model is more important for haplotypes with few phenotypic observations, (iii) the HN model is more important for noisy phenotypic data and (iv) the mutation model was marginally better than the HN model in estimating haplotype effects. We go through each of the findings in detail.

The HN model outperformed the IH model for all tested parameter sets. In Figure 4, all RCRPS values are well below zero. For haplotypes observed in several or one phenotyped individual, the RCRPS was lower than what was seen in panels (A) and (B) in Figure 3. For haplotypes with no direct links to phenotype observations, the RCRPS was not improving as much as seen in panel (C) in in Figure 3.

The improvement with the HN model relative to the IH model increased

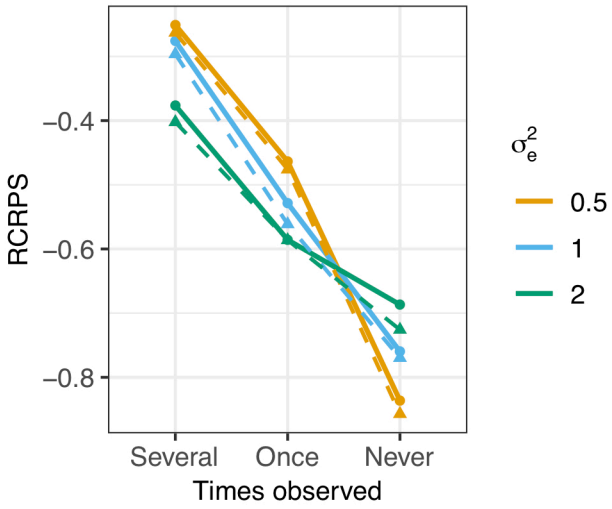


Figure 4: RCRPS (smaller values indicate that the HN or mutation models, respectively, are better than the reference IH model) between the HN model and the IH model (solid line) and between the mutation model and the IH model (dashed line). The data were simulated from the mutation model with varying residual variance σ_e^2 and empirical haplotype variance 1 ($\text{Var}(\boldsymbol{h}) = 1$). The three scenarios show RCRPS for the haplotypes that were observed in several phenotyped individuals (Several), only one phenotyped individual (Once), and were not observed in a phenotyped individual (Never)

with fewer phenotype observations per haplotype. The RCRPS in Figure 4 is lowest for haplotypes with no direct links to phenotype observations, second lowest for haplotypes with one direct link to a phenotype observation, and highest for haplotypes that were observed in several phenotyped individuals.

The improvement with the HN model relative to the IH model increased with increasing residual variation. In Figure 4 the RCRPS for haplotypes observed in several or one phenotyped individual decreases with increasing residual variance. This was again not the case for haplotypes with no direct links to phenotype observations. As mentioned in the previous section, the IH model is predicting new haplotypes equally poorly irrespective of

the residual variance. The HN model on the other hand, improves the prediction of new haplotypes when the phenotypic data is less noisy. This explains why the improvement in RCRPS between the HN model and IH model for haplotypes with no direct links to phenotype observations is largest with less residual variation, seen by the RCRPS in panel (C) being lowest for residual variance 0.5.

The mutation model was marginally better than the HN model in estimating haplotype effects. The dashed lines in Figure 4 indicate the RCRPS between the mutation model and the IH model, and the full lines indicate the RCRPS between the HN model and the IH model. The dashed lines and full lines follow each other closely, and the dashed lines are slightly lower than the full lines, indicating that the mutation model was slightly better than the HN model, although not by much.

In Table 2 we present the average RCRPS between the HN model and the mutation model for the posterior mutation effects. This table has the RCRPS for the two scenarios, where either all haplotypes had associated phenotype observation, or most haplotypes had associated phenotype observation and the rest did not, with different proportions of mutations with causal effect, and for different residual variance. RCRPS above zero indicates that the mutation model had the best estimates, and RCRPS below zero indicates that the HN model had the best estimates. Overall, the difference between the two models is small. The mutation model had the best performance when there were few causal mutations, and the HN model had the best performance when there were many causal mutations.

3.2 Case study results: Mitochondrial haplotypes in cattle

In this section we present results for the case study of estimating the effect of mitochondrial haplotypes on milk yield in cattle presented in Section 2.5 “Case study: Mitochondrial haplotypes in cattle”. We present the posterior mean and standard deviation for the effect of mitochondrial haplotypes mapped onto the phylogeny, the posterior distribution for the autocorrelation parameter ρ, and the mean and 95% confidence interval of the posterior variances in the model.

In summary the results show (i) that there was sharing of information between the mitochondrial haplotypes, (ii) that haplotypes without a direct link to observed phenotypes were estimated with larger uncertainty, (iii) indications of strong phylogenetic dependency between the haplotypes and,

Table 2: RCRPS between the HN model and the mutation model for mutation effects by different values of residual variance σ_e^2, proportion of causal mutations, and for the two scenarios where either all or most haplotypes had direct links to observed phenotypes

	All observed	Most observed
$\boldsymbol{\sigma}_e^2 = 0.5$		
0.1	0.060	0.071
0.3	0.019	0.025
0.5	-0.002	-0.004
0.7	-0.019	-0.021
0.9	-0.027	-0.029
$\boldsymbol{\sigma}_e^2 = 1$		
0.1	0.123	0.111
0.3	0.043	0.037
0.5	0.004	0.000
0.7	-0.024	-0.022
0.9	-0.041	-0.034
$\boldsymbol{\sigma}_e^2 = 2$		
0.1	0.168	0.214
0.3	0.067	0.101
0.5	0.006	0.018
0.7	-0.025	-0.026
0.9	-0.042	-0.048

(iv) a significant proportion of the total phenotypic variation explained by mitochondrial haplotypes. We now go through each of these findings in detail.

The HN model enabled sharing of information from the haplotypes that had a direct link with observed phenotypes, to the other haplotypes. In Figure 5 we present the posterior mean for the effect of mitochondrial haplotypes with node color. Haplotype effect estimates are similar for phylogenetically similar haplotypes, meaning that there was sharing of information between the haplotypes. The figure also shows that the few haplotypes that had direct links with phenotype observations (nodes labeled with 1) were separated from each other with a substantial number

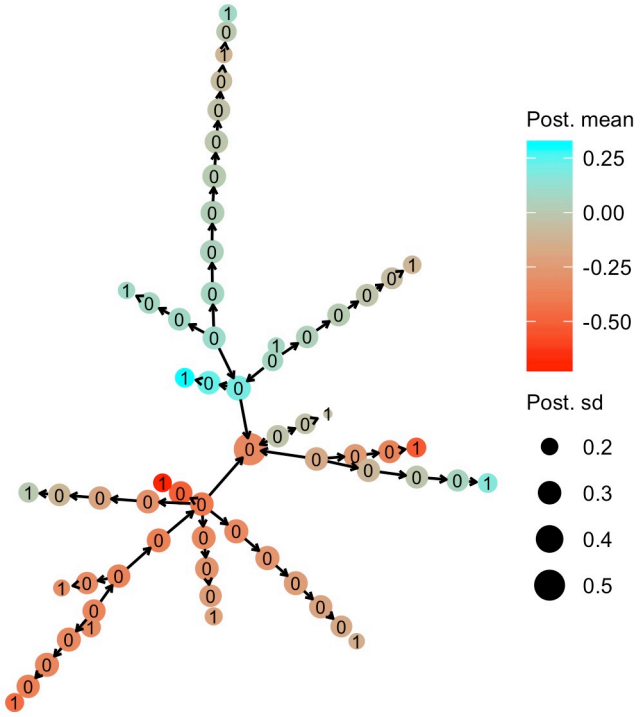


Figure 5: Posterior mean and standard deviation for mitochondrial haplotype effects on milk yield in cattle. Posterior means are denoted with node color, while posterior standard deviations are denoted by the node size. The number on each haplotype node indicates if the haplotype had a direct link to a observed phenotype (1) or not (0)

of mutations, and still information was shared through the phylogeny to all haplotypes, which was the aim of the HN model.

Haplotypes without direct links to observed phenotypes were estimated with larger uncertainty. In Figure 5 we present the posterior standard deviation for the effect of mitochondrial haplotypes with node size. We see that the haplotypes with direct links to observed phenotypes (nodes labeled with 1) had smaller posterior standard deviation than the other hapolotypes (nodes labeled with 0). The posterior standard deviation decreased slightly

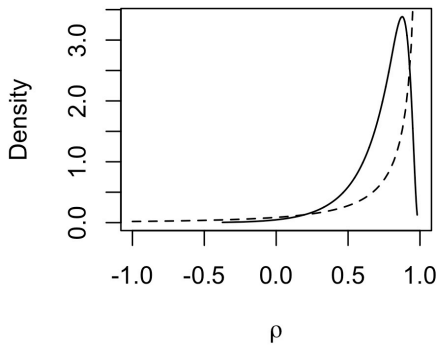


Figure 6: Prior (dashed line) and posterior (solid line) distribution for the autocorrelation parameter ρ for mitochondrial haplotype effects on milk yield in cattle

as the haplotypes without direct links were closer (in number of mutations) to the haplotypes with direct links, which was expected. However, the overall posterior standard deviations for haplotype effects were relatively large, because the data set was small and there were few haplotypes with direct links to observed phenotypes, connected to the other haplotypes with many mutations between them.

The posterior distribution for the autocorrelation parameter ρ indicated strong dependency between haplotype effects. The posterior distribution of ρ is shown in Figure 6 together with the prior distribution. The mode of the posterior distribution lies around 0.85, and the mean lies around 0.73, indicating that neighboring haplotypes had similar effects, which is related to the sharing of information between haplotypes seen in Figure 5. We also note that the posterior distribution shifted to slightly lower values of ρ compared to the prior distribution.

A significant amount of the total phenotypic variation was explained by the mitochondrial haplotypes. In Table 3 we present the posterior mean and 95% confidence interval of each variance component in the model, and how much of the total variation in the model ($\sigma_c^2 + \sigma_a^2 + \sigma_{h_m}^2 + \sigma_e^2$) was explained

Table 3: Posterior mean, 95% confidence interval (CI) for variance parameters, and the proportion of variation explained by each variance component for the case study estimating mitochondrial haplotype effects on milk yield in cattle

Variance parameter	Mean	CI	Prop. variance explained
σ^2_c	0.035	$(0.005, 0.090)$	0.047
σ^2_a	0.329	$(0.194, 0.533)$	0.444
$\sigma^2_{h_m}$	0.113	$(0.033, 0.264)$	0.152
$\sigma^2_{h_c}$	0.048	$(0.007, 0.154)$	0.065
σ^2_e	0.265	$(0.171, 0.416)$	0.357

σ^2_c: variance of contemporary group effects, σ^2_a: variance of nuclear-genome additive effects, $\sigma^2_{h_m}$: marginal variance of mitogenome haplotype effects, $\sigma^2_{h_m}$: conditional variance of mitogenome haplotype effects, σ^2_e: variance of residuals

by each variance component. The posterior distribution of the conditional haplotype variance was obtained by computing $\sigma^2_{h_c} = \sigma^2_{h_m}(1 - \rho^2)$, using 10000 samples from the mean posterior distributions of the marginal haplotype variance and the autocorrelation parameter. We see that the marginal haplotype variance $\sigma^2_{h_m}$ and conditional haplotype variance $\sigma^2_{h_c}$ is smaller compared to the additive genetic variance σ^2_a, and the residual variance σ^2_e. This was expected as the mitogenome ($\sim 1 \times 16kbp$) is much smaller than the nuclear genome ($\sim 2 \times 3Gbp$). In the light of this difference we can say that mitochondrial haplotypes captured a significant amount of phenotypic variation. The variance for the random effect of herd-year-season of calving σ^2_c was also small compared to σ^2_a and σ^2_e.

It should be noted that this is a small data set with few haplotypes with direct links to observed phenotypes. This causes the posterior estimates to be strongly influenced by the prior distributions, especially the posterior for ρ, which we can see in Figure 6. However, we still chose to assign an informative prior to ρ, since it is expected that most mutations have no causal effect, and that phylogenetically similar haplotypes have similar effects.

4 Discussion

The objective of this paper was to propose a hierarchical model that leverages haplotype phylogeny to improve the estimation of haplotype effects. We have presented the haplotype network model, evaluated it using simulated data from two different generative models, and applied it in a case study of estimating the effect of mitochondrial haplotypes on milk yield in cattle. Here, we highlight three points for discussion in relation to the proposed haplotype network model: the importance of the haplotype network model, future development and possible extensions and limitations.

4.1 The importance of the haplotype network model

We see three important advantages of the haplotype network model. These are specifically the ability to share information between related haplotypes, computational advantages when modeling a single region of a genome, and the potential to capture background specific mutation effects.

The haplotype network model utilizes phylogenetic relationships between haplotypes and with this improves estimation of their effects. From the simulation study, we saw the importance of this information sharing when there is limited information per haplotype. For example, we were able to estimate the effect of haplotypes that had few or no direct links to observed phenotypes with much higher accuracy than with a model assuming independent haplotypes. In the haplotype network model the autocorrelation parameter ρ and the conditional variance parameter $\sigma^2_{h_c}$ reflect the effects of phylogenetically similar haplotypes. As the autocorrelation approaches 1, haplotype effects become more dependent. Further, if the conditional variance is small the large dependency and small deviations lead to similar effects for phylogenetically similar haplotypes, suggesting that mutations separating the haplotypes have very small or no effect compared to other shared mutations between haplotypes. If on the other hand conditional variance is large, the large dependency and large deviations lead to haplotype effects that change rapidly along the phylogeny, suggesting that mutations separating the haplotypes have large effects. If the autocorrelation parameter approaches 0, the dependency between phylogenetically similar haplotypes is decreasing, suggesting that haplotypes should be modeled independently.

The three extreme scenarios of hyper-parameter values could denote three real cases. The first case with high autocorrelation and small conditional variance could reflect a situation where the whole haplotype sequence would be used to build a phylogeny and since most mutations do not have a causal effect, but some do, it is expected that similar haplotypes will have similar effects with small differences between the haplotypes. The second case with high autocorrelation and large conditional variance could reflect the situation when the number of causal mutations would be high compared to all mutations (because only such mutations are analyzed), and therefore change of effects along the phylogeny would be larger. The third scenario with no autocorrelation could reflect the situation where phylogeny does not correlate with phenotype change, which can be due to many reasons (analysis of a genome region that is not associated with the phenotype, inadequate genomic platform to capture genotype-phenotype association, etc.).

As mentioned in the introduction, modeling phenotypic variation as a function of haplotype variation has extensive literature (Templeton et al., 1987; Balding, 2006; Thompson, 2013; Morris and Cardon, 2019). The prime motivation for this work was the recent growth in the generation of large scale genomic data sets and methods to build phylogenies (Kelleher et al., 2019). To this end we aimed to develop a general haplotype network model that could exploit phylogenetic relationships between haplotypes in a computationally efficient way. Namely, the model uses phylogeny encoded with a DAG, which enables sharing of information between similar haplotypes in a recursive way that also implies computational benefits. The computational benefits come from the sparse precision matrix $\boldsymbol{V}_h^{-1}$, which is due to the conditional independence structure encoded in the DAG of a network of haplotypes (Rue and Held, 2005). Sparsity is important, because it enables fitting large models due to smaller memory requirement and faster calculations (Rue and Held, 2005). The computational benefits are not critical when the number of haplotypes is small. In that case the matrix $\boldsymbol{V}_h$ is small and easy to invert, but for the autoregressive model we would have to invert it many times during the estimation procedure due to dependency on the autocorrelation parameter. However, it is better to avoid inversions if possible because it can lead to numerical errors and loss of precision (e.g., Misztal, 2016).

While the haplotype network model is different to the pedigree mixed

model (Henderson, 1976; Quaas, 1988) (where we model the inheritance of whole genomes in a pedigree without (fully) observing the genomes) or the phylogenetic mixed model (Lynch, 1991; Pagel, 1999; Housworth et al., 2004; Hadfield and Nakagawa, 2010) (where we model the inheritance of whole genomes in a phylogeny without (fully) observing the genomes), the principles of conditional dependence between genetic effects and the resulting sparsity are the same (Rue and Held, 2005). The key difference of the haplotype network model is that it estimates the effect of observed haplotype sequences as compared to unobserved or partially observed inheritance of whole genomes in a pedigree or phylogeny. To improve the estimation of the haplotype effects we take into account the phylogenetic relationships. A similar model has also been used in spatial disease mapping (Datta et al., 2019), showing potential of this kind of model in several applications.

While the use of phylogenetic relationships might seem redundant if we know (most of) the haplotype sequence, the simulations showed that it improves estimation in most cases, even marginally compared to the mutation model where we directly model mutation effects. The haplotype network model can be seen as a hybrid between the mutation model (that models variation between the columns of a haplotype matrix) and the independent haplotype model (that models variation between the rows of a haplotype matrix). This hybrid view might improve genome-wide association studies (see reviews by Gibson, 2018; Simons et al., 2018; Uricchio, 2019; Morris and Cardon, 2019).

The haplotype network model has the potential to capture background specific mutation effects. Background specific mutation effects are observed when the effect of a mutation depends on other mutations present in an individual (e.g., Chandler et al., 2017; Wojcik et al., 2019; Steyn et al., 2019). Such effects can also manifest when a mutation is marking another unobserved mutation with correlation that varies between backgrounds. The haplotype network model can capture background specific mutation effects through the fact that it is modeling haplotype effects and not mutation effects. If there are background specific mutation effects the haplotype effect differences will capture this, while a mutation model only estimates an average effect of a mutation across multiple backgrounds (haplotypes). However, we must point that the haplotype network model captures only local effects, that is due to interactions between mutations

present on a haplotype (e.g., Clark, 2004; Liu et al., 2019). We have not evaluated how well the model captures background specific mutation effects in this study, and more simulations and applications to a range of data sets are needed to evaluate this aspect.

4.2 Future development and possible extensions

There are several areas for future development with the haplotype network model. We are looking into some areas: making the model more flexible in the number of mutations separating phylogentically similar haplotypes, modeling haplotype differences in a continuous way utilizing branch lengths, and incorporating biological information and phylogenetic aspects of haplotype relationships.

We have developed the haplotype network model by assuming the differences between similar haplotypes are due to one mutation to simplify the model definition. However, in the observed data there might not be haplotypes that are separated by just one mutation. We handle this situation by inserting phantom haplotypes. The order of mutations in such situations is uncertain, and a model could be generalized to account for these larger number of mutations between haplotypes. However, the current "one-mutation" difference model setup has a useful property of inferring the value of unobserved haplotypes, and the sparse model definition does not increase the computational complexity of the model.

The haplotype network model could be generalized to utilize time calibrated distances between haplotypes rather than using the number of mutations. The Ornstein–Uhlenbeck (OU) process is the continuous-time analogue of the autoregressive process of order one used in this study, and plays a major role in the analysis of the evolution of phenotypic traits along phylogenies (Lande, 1976; Hansen and Martins, 1996; Martins and Hansen, 1997; Paradis, 2014; Yang et al., 2018). Relatedly, if the autocorrelation parameter of the autoregressive process of order one is set to 1 we get the non-stationary discrete random walk process, whose continuous-time analogue is the Brownian motion, that is the basic model of phylogenetic comparative analysis (Felsenstein, 1988; Huey et al., 2019). There is a scope to improve computational aspects for these continuous models too by employing recent developments from the statistical analysis of irregular time-series (Lindgren and Rue, 2008).

In the haploptype network model presented in this study, the same

autocorrelation parameter has been assumed for all mutations. However, the autocorrelation parameter could be allowed to vary as Beaulieu et al. (2012) did in the context of adaptive evolution. The stationary autoregressive process of order one for trees with only one ancestral haplotype and no recombination allows for such extensions without having to change the variance parameter. For example, one could use different autocorrelation parameters for different types of mutations to incorporate biological information into the model. This would enable combining the quantitative analysis of mutation and haplotype effects from this study with molecular genetic tools such as Variant Effect Predictor (McLaren et al., 2016). Biological information about haplotypes could also be incorporated using variant specific covariates, for example as implemented by (Susak et al., 2020).

In this study we have assumed that the phylogenetic network is given, and described with a DAG. There is a large body of literature on inferring phylogenies in the form of strict bifurcating trees, more general trees or networks, and recent developments in genomics are rapidly advancing the field (e.g., Anisimova, 2012; Puigbò et al., 2013; Schliep et al., 2017; Uyeda et al., 2018). We have named the model the haplotype network model, because it can work both with phylogenetic bifurcating and multifurcating trees, and phylogenetic networks. The only condition is that we describe the haplotype relationships with a DAG, which gives the structure to the hierarchical haplotype model. Many tools provide such output (e.g., Leigh and Bryant, 2015; Suchard et al., 2018; Kelleher et al., 2019). To accommodate general DAGs, where a haplotype node could potentially have multiple parental haplotype nodes, we have generalized the model construction to allow for network structures. This generalization also enables the model to describe haplotype relationships without paying attention to the directionality, as long as there are no directed loops in the graph.

Knowing the order of mutations, and therefore which haplotypes are parental to other haplotypes, is beneficial because it leads to a tree structure and a sparser model (Rue and Held, 2005). An example of non-optimal sparsity can be seen in our case study. In Figure 5 the "central" haplotype with the largest uncertainty is modeled as a progeny haplotype of four surrounding haplotypes, which means that there is a dense 5×5 block in the precision matrix $\boldsymbol{V}_h^{-1}$. The block is dense because the "central"

haplotype is modeled as a function of the other four "parental" haplotypes, which invokes conditional dependence between the "parental" haplotypes. If however the "central" haplotype would have been used as the parental haplotype the 5×5 block would be sparse since all other haplotypes would be conditionally independent given the "central/parental" haplotype. The same applies also for the other parts of the haplotype network in Figure 5. These examples of non-optimal sparsity are a consequence of the haploptype network we used, but we did this for simplicity and to emphasize the flexibility of the haploptype network model.

The haploptype network model could also work with probabilistic networks where edges have associated uncertainty (weights). If we can encode such a network with a DAG, then the edge weights can be used in model construction, for example in the same way uncertain parentage is handled in pedigree models (Henderson, 1976). An alternative would be to construct a model for each possible realization of a network, run separate models and combine haplotype estimates in the spirit of Bayesian model averaging. This latter approach is obviously computationally more demanding.

4.3 Limitations

The haplotype network model also has limitations that merit further development. We highlight three areas: if the haplotype network model necessary given that we can model mutation effects, the Gaussian assumption and causal mutations, and modeling recombining haplotypes.

For the haplotype network model to achieve its full potential, the data needs to have a certain structure. We saw from fitting the hapolotype network model to a real data set, that having few haplotypes with direct links to observed phenotypes and many haplotypes without, meant that we had large uncertainty in estimated haplotype effects. We also saw from the simulated data, that the mutation model was slightly better at estimating the mutation effects than the haplotype network model, when the data were simulated from a mutation model. However, the magnitude of difference was minimal. In the future, different data structures with balanced and unbalanced structure spanning multiple populations with varying levels of connectedness, small or large number of mutations, and causal or non-causal mutations should be tested to find optimal scenarios for the haplotype network model to achieve its full potential.

The haplotype network model assumes that the haplotype effects follow a Gaussian distribution. If all, or very many, of the haplotypes have the same effect, the true haplotype effect distribution may be quite different from Gaussian, which breaks the model assumptions and perhaps other models should be proposed. Blomberg et al. (2019) describe the underlying theory behind the common Gaussian processes, such as Brownian motion and Ornstein-Uhlenbeck process, and present general methods for deriving new stochastic models, including non-Gaussian models of quantitative trait macroevolution. See also Schraiber and Landis (2015), Landis et al. (2012) and Duchen et al. (2017). Application of these models will depend on the magnitude of deviations from Gaussian assumptions, which might be large on the scale of macroevolution, but might also be large when looking at a specific genome region.

Scaling the haplotype network model to multiple recombining haplotype regions is challenging for two reasons. First, while phasing methods have improved substantially in the last years (Marchini, 2019), determining a recombination breakpoint is challenging due to a limited resolution to resolve exact locus where recombination occured. Second, the sparsity of the haplotype network model comes from the sparsity of the precision matrix $\boldsymbol{V}_h^{-1}$, which is part of the prior distribution for haplotype effects. When developing the extension for recombining haplotypes we observed that the sparsity in the prior is maintained also for multiple consequitive haplotype regions along a chromosome as shown in (11). However, we observed that the design matrices that link phenotype observations with multiple haplotype regions start to create dense cross-products in the system of equations as we increase the number of regions, and the sparsity advantage from one haplotype region is lost. To this end we are exploring alternative ways of formulating the haplotype network model following data structures in Kelleher et al. (2019). Further research is needed to be able to scale the haplotype network model to many haplotype regions or even whole chromosomes and genomes.

Bibliography

Anisimova, M. (2012). *Evolutionary Genomics Statistical and Computational Methods, Volume*, volume 855. Springer.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature reviews genetics*, 7(10):781.

Basseville, M., Benveniste, A., Chou, K. C., Golden, S. A., Nikoukhah, R., and Willsky, A. S. (1992). Modeling and estimation of multiresolution stochastic processes. *IEEE Transactions on Information Theory*, 38(2):766–784.

Beaulieu, J. M., Jhwueng, D.-C., Boettiger, C., and O'Meara, B. C. (2012). Modeling stabilizing selection: Expanding the Ornstein–Uhlenbeck model of adaptive evolution. *Evolution: International Journal of Organic Evolution*, 66(8):2369–2383.

Begum, R. (2019). A decade of genome medicine: Toward precision medicine. *Genome Med*, 11(13).

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blomberg, S. P., Rathnayake, S. I., and Moreau, C. M. (2019). Beyond Brownian motion and the Ornstein-Uhlenbeck process: Stochastic diffusion models for the evolution of quantitative characters. *The American Naturalist*, 195(2):000–000.

Brajković, V. (2019). *Utjecaj mitogenoma na svojstva mliječnosti goveda (Eng: Impact of mitogenome on milk traits in cattle)*. PhD thesis, University of Zagreb. Faculty of Agriculture.

Chandler, C. H., Chari, S., Kowalski, A., Choi, L., Tack, D., DeNieu, M., Pitchers, W., Sonnenschein, A., Marvin, L., Hummel, K., et al. (2017). How well do you know your mutation? Complex effects of genetic background on expressivity, complementation, and ordering of allelic effects. *PLoS genetics*, 13(11):e1007075.

Clark, A. G. (2004). The role of haplotypes in candidate gene studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 27(4):321–333.

Datta, A., Banerjee, S., Hodges, J. S., and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models. *Bayesian Analysis*, 14:1221–1244.

de los Campos, G., Vazquez, A. I., Hsu, S., and Lello, L. (2018). Complex-trait prediction in the era of big data. *Trends in Genetics*, 34(10):746–754.

Duchen, P., Leuenberger, C., Szilágyi, S. M., Harmon, L., Eastman, J., Schweizer, M., and Wegmann, D. (2017). Inference of evolutionary jumps in large phylogenies using Lévy processes. *Systematic biology*, 66(6):950–963.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112.

Ewens, W. J. (2004). *Mathematical population genetics 1.* Springer-Verlag, New York, NY, 2 edition.

Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annual Review of Ecology and Systematics*, 19(1):445–471.

Gardiner, C. (2009). *Stochastic Methods. A Handbook for the Natural and Social Sciences.* Springer Series in Synergetics, 4th ed. Springer.

Gibson, G. (2018). Population genetics and GWAS: A primer. *PLoS biology*, 16(3):e2005485.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hadfield, J. and Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of evolutionary biology*, 23(3):494–508.

Hansen, T. F. and Martins, E. P. (1996). Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. *Evolution*, 50(4):1404–1417.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics*, pages 69–83.

Hickey, J. M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., Jones, C., Canales, C., Grattapaglia, D., Bassi, F., et al. (2017). Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. *Nature genetics*, 49(9):1297.

Housworth, E. A., Martins, E. P., and Lynch, M. (2004). The phylogenetic mixed model. *The American Naturalist*, 163(1):84–96. PMID: 14767838.

Huey, R. B., Garland Jr, T., and Turelli, M. (2019). Revisiting a key innovation in evolutionary biology: Felsenstein's "phylogenies and the comparative method". *The American Naturalist*, 193(6):755–772.

Ibanez-Escriche, N. and Simianer, H. (2016). Animal breeding in the genomics era [Special issue]. *Animal Frontiers*, 6. (Eds.).

Kelleher, J., Etheridge, A. M., and McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput Biol*, 12(5):1–22.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K., and McVean, G. (2019). Inferring whole-genome histories in large population datasets. *Nature genetics*, 51(9):1330–1338.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Lande, R. (1976). Natural selection and random genetic drift in phenotypic evolution. *Evolution*, 30(2):314–334.

Landis, M. J., Schraiber, J. G., and Liang, M. (2012). Phylogenetic analysis using Lévy processes: finding jumps in the evolution of continuous traits. *Systematic biology*, 62(2):193–204.

Leigh, J. W. and Bryant, D. (2015). PopART: Full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, 6(9):1110–1116.

Lello, L., Avery, S. G., Tellier, L., Vazquez, A. I., de los Campos, G., and Hsu, S. D. (2018). Accurate genomic prediction of human height. *Genetics*, 210(2):477–497.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700.

Liu, F., Schmidt, R. H., Reif, J. C., and Jiang, Y. (2019). Selecting closely-linked SNPs based on local epistatic effects for haplotype construction improves power of association mapping. *G3: Genes, Genomes, Genetics*, 9(12):4115–4126.

Lynch, M. (1991). Methods for the analysis of comparative data in evolutionary biology. *Evolution*, 45(5):1065–1080.

Maier, R. M., Zhu, Z., Lee, S. H., Trzaskowski, M., Ruderfer, D. M., Stahl, E. A., Ripke, S., Wray, N. R., Yang, J., Visscher, P. M., et al. (2018). Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nature communications*, 9(1):989.

Marchini, J. (2019). Haplotype estimation and genotype imputation. *Handbook of Statistical Genomics 4e 2V SET*, pages 87–114.

Martins, E. P. and Hansen, T. F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, 149(4):646–667.

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. *Genome biology*, 17(1):122.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Misztal, I. (2016). Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics*, 202(2):401–409.

Morris, A. P. and Cardon, L. R. (2019). *Genome-Wide Association Studies*, chapter 21, pages 597–550. John Wiley & Sons, Ltd.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877.

Paradis, E. (2014). *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice*, chapter Simulation of Phylogenetic Data, pages 335–350. Springer Berlin Heidelberg, Berlin, Heidelberg.

Puigbò, P., Wolf, Y. I., and Koonin, E. V. (2013). Seeing the tree of life behind the phylogenetic forest. *BMC biology*, 11(1):46.

Quaas, R. (1988). Additive genetic model with groups and relationships. *Journal of Dairy Science*, 71(5):1338–1345.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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(2):319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4:395–421.

Rue, H. v. and Held, L. (2005). *Gaussian Markov random fields: Theory and applications*. Chapman and Hall/CRC.

Schliep, K., Potts, A. A., Morrison, D. A., and Grimm, G. W. (2017). Intertwining phylogenetic trees and networks. *Methods in Ecology and Evolution*, (10):1212–1220.

Schraiber, J. G. and Landis, M. J. (2015). Sensitivity of quantitative traits to mutational effects and number of loci. *Theoretical population biology*, 102:85–93.

Simons, Y. B., Bullaughey, K., Hudson, R. R., and Sella, G. (2018). A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS biology*, 16(3):e2002985.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28.

Sørbye, S. H. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38(6):923–935.

Steyn, Y., Lourenco, D. A. L., and Misztal, I. (2019). Genomic predictions in purebreds with a multi-breed genomic relationship matrix. *Journal of Animal Science*.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J., and Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using beast 1.10. *Virus Evolution*, 4(1):vey016.

Susak, H., Serra-Saurina, L., Janssen, R. R., Domènech, L., Bosio, M., Muyas, F., Estivill, X., Escaramis, G., and Ossowsky, S. (2020). Efficient and flexible integration of variant characteristics in rare variant association studies using integrated nested Laplace approximation. *bioRxiv*.

Templeton, A. R., Boerwinkle, E., and Sing, C. F. (1987). A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics*, 117(2):343–351.

Thompson, K. L. (2013). *Using ancestral information to search for quantitative trait loci in genome-wide association studies*. PhD thesis, The Ohio State University.

Uricchio, L. H. (2019). Evolutionary perspectives on polygenic selection, missing heritability, and GWAS. *Human genetics*, pages 1–17.

Uyeda, J. C., Zenil-Ferguson, R., and Pennell, M. W. (2018). Rethinking phylogenetic comparative methods. *Systematic Biology*, 67(6):1091–1109.

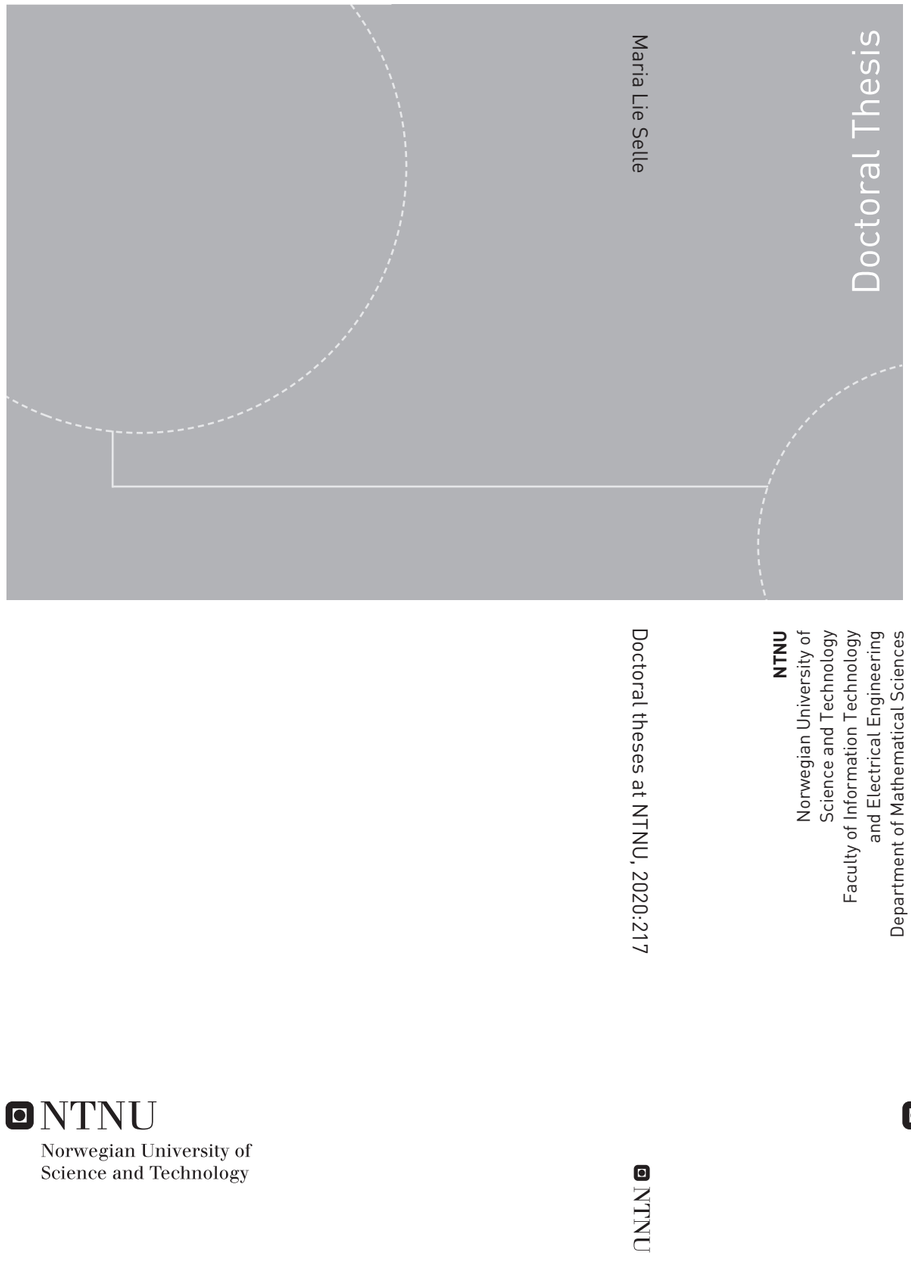
Walsh, B. and Lynch, M. (2018). *Evolution and selection of quantitative traits*. Oxford University Press.

Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., et al. (2019). Genetic analyses of diverse populations improves discovery for complex traits. *Nature*, page 1.

Wu, P., Hou, L., Zhang, Y., and Zhang, L. (2020). Phylogenetic tree inference: A top-down approach to track tumor evolution. *Frontiers in Genetics*, 10:1371.

Yang, Y., Gu, Q., Zhang, Y., Sasaki, T., Crivello, J., O'Neill, R. J., Gilbert, D. M., and Ma, J. (2018). Continuous-trait probabilistic model for comparing multi-species functional genomic data. *Cell systems*, 7(2):208–218.

ISBN 978-82-326-4782-8 (printed version)
ISBN 978-82-326-4783-5 (electronic version)
ISSN 1503-8181



Doctoral theses at NTNU, 2020:217

Maria Lie Selle

Novel statistical variance and dependency models in quantitative genetics

Enabled by recent inference methods

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology
and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Doctoral theses at NTNU, 2020:217

NTNU

NTNU
Norwegian University of
Science and Technology