7-26-2019

Adjusting for Spatial Effects in Genomic Prediction

Xiaojun Mao
Fudan University

Somak Dutta
Iowa State University, somakd@iastate.edu

Raymond K. W. Wong
Texas A & M University - College Station

Dan Nettleton
Iowa State University, dnett@iastate.edu

Adjusting for Spatial Effects in Genomic Prediction

Abstract
This paper investigates the problem of adjusting for spatial effects in genomic prediction. Despite being seldomly considered in genome-wide association studies (GWAS), spatial effects often affect phenotypic measurements of plants. We consider a Gaussian random field (GRF) model with an additive covariance structure that incorporates genotype effects, spatial effects and subpopulation effects. An empirical study shows the existence of spatial effects and heterogeneity across different subpopulation families while simulations illustrate the improvement in selecting genotypically superior plants by adjusting for spatial effects in genomic prediction.

Disciplines
Applied Statistics | Genomics | Plant Sciences | Statistical Models

Comments
This is a pre-print made available through arxiv: https://arxiv.org/abs/1907.11581.

Adjusting for Spatial Effects in Genomic Prediction

Xiaojun Mao* Somak Dutta† Raymond K. W. Wong‡ Dan Nettleton§

Abstract

This paper investigates the problem of adjusting for spatial effects in genomic prediction. Despite being seldomly considered in genome-wide association studies (GWAS), spatial effects often affect phenotypic measurements of plants. We consider a Gaussian random field (GRF) model with an additive covariance structure that incorporates genotype effects, spatial effects and subpopulation effects. An empirical study shows the existence of spatial effects and heterogeneity across different subpopulation families while simulations illustrate the improvement in selecting genotypically superior plants by adjusting for spatial effects in genomic prediction.

Keywords: Gaussian random field; Genomic prediction; Spatial effects; Subpopulation effects.

1 Introduction

In plant breeding, predicting the genetic value of plant genotypes plays an important role in determining which genotype to include in subsequent generations. Recently, several powerful GWAS statistical methods have been

*School of Data Science, Fudan University, Shanghai 200433, China P.R.C. Email: maoxj@fudan.edu.cn

†Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. Email: somakd@iastate.edu

‡Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A. Email: raywong@stat.tamu.edu

§Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A. Email: dnett@iastate.edu

developed that use high-dimensional single-nucleotide polymorphism (SNP) genotypes for genomic prediction. Most of the methods based on mixed linear models (MLM) are quite flexible due to the consideration of fixed and random effects. For instance, population structure (discussed in Pritchard et al. 2000) is often accounted for by modeling the fixed effects of principal components (PCs) derived from the SNPs (Price et al., 2006; Reich et al., 2008; McVean, 2009). For unified MLM approaches (Yu et al., 2006), SNP data are used to determine a kinship matrix that is assumed to be proportional to the variance of a vector of random effects that accounts for dependencies due to relatedness among individuals. For a more computationally efficient Compressed MLM (CMLM) approach (Zhang et al., 2010), data from many individuals are compressed into a smaller number of groups, and the inter-individual kinship matrix is replaced by a lower-dimensional matrix that characterizes correlations among group random effects induced by genetic similarities among groups.

Aside from correlations due to relatedness among individuals or groups, phenotypes measured on plants grown in fields can be spatially correlated. Such correlation can arise because plants growing near each other may share a common micro-environment that differs from the micro-environment experienced by plants in other parts of the field. This micro-environmental variation can induce phenotypic similarity among neighboring plants. When such spatial effects exist but are unaccounted for in the analysis, decisions about which plant genotypes are expected to perform best with regard to one or more phenotypic traits can be adversely affected. With the adjustment of these effects, some high-throughput phenotyping technologies (Cabrera-Bosquet et al., 2012; Masuka et al., 2012; White et al., 2012) can be applied to increase plant yields.

Several works (Crossa et al., 2006; Lado et al., 2013; Bernal-Vasquez et al., 2014) have considered spatial effects in linear mixed-effects models. As suggested by Bernal-Vasquez et al. (2014), fitting a row and column model (RC) (i.e., a model with an effect for each row and for each column in a field experiment layout) can account for a substantial portion of phenotypic heterogeneity that may be due to spatial effects. Lado et al. (2013) compared RC models with approaches that attempt to adjust for spatial effects by using the difference between a plot's response value and the average response of its neighboring plots as a covariate. Such a method, referred to by Lado et al. (2013) as "moving-means as a covariate" (MVNG), was found to best fit the data and lead to the most accurate phenotypic predictions. In this

paper, we propose an alternative modeling strategy that has some conceptual advantages and shows performance improvements relative to the existing approaches for spatial adjustment in genomic prediction.

We study two datasets. One is a maize dataset involving a nested association mapping (NAM) panel consisting of 4660 recombinant inbred lines (RILs) derived from crosses between a reference inbred line B73 and 25 other founder inbreds. More information about the NAM panel is available in Yu et al. (2008) and at `http://www.panzea.org`. The RILs derived by crossing B73 to any one of the 25 other founders from a subpopulation of RILs. Thus, the 4660 RILs we consider can be partitioned into 25 subpopulations. Even after conditioning on SNP genotypes carried by each RIL, phenotypic responses from RILs within a subpopulation are expected to be more strongly correlated than responses from RILs in different subpopulations. This within-subpopulation correlation is expected due to shared genetic material as well as characteristics of the experimental design described in Section 2. The second dataset is a wheat dataset which consists of genotype and phenotype data on 384 advanced lines from two different breeding programs. The data are provided in Lado et al. (2013).

The goal of this paper is to predict the genetic value of each maize RIL or each wheat line from a huge number of SNP marker genotypes, while accounting for the genetic and spatial dependence among phenotypic measurements. We focus on a Gaussian random field (GRF) model with an additive covariance matrix structure that incorporates genotype effects, spatial effects and subpopulation effects. For genotype effects, we adopt a Gaussian kernel (Morota et al., 2013; Ober et al., 2011) to capture general relationships between genotypes and phenotypes. We compare our spatially adjusted genomic predictions with genomic predictions generated by a design-based incomplete block (IB) linear mixed-effects model and existing methods CMLM (Zhang et al., 2010), RC (Bernal-Vasquez et al., 2014) and MVNG (Lado et al., 2013). In a simulation study presented in Section 5, we apply the proposed GRF method to help identify the best plant genotypes.

The rest of the paper is organized as follows. Real data are described in Section 2. The proposed GRF is constructed in Section 3. Within Section 3, we also discuss kernels and corresponding parameter estimation methods. Numerical performances of the proposed method for genomic predictions and for choosing the best plant genotypes are illustrated in an empirical study in Section 4 and a simulation study in Section 5, respectively. The paper concludes with a discussion in Section 6. Some supplementary materials are

provide in Section 8.1.

2 Data

Throughout this paper, we used Data1 to refer to a maize NAM RIL dataset comprised of 4660 RILs genotyped at 687869 SNP markers. The phenotypic value for each RIL is a measurement of the carbon dioxide (CO_2) emitted from plant material incorporated in a soil sample. Scientific interest centers on identifying RILs whose genetic constitution makes them relatively low emitters of CO_2.

The 4660 RILs in Data1 can be partitioned into 25 subpopulations, each produced from a biparental cross of inbred line B73 to one of the 25 NAM founder inbred lines. Due to the large number of RILs and limited field plot availability, the experimental design is unreplicated with a single plot for each RIL distributed across three nearby agricultural fields, with no two plots separated by more than 2.5 miles. RILs from any given subpopulation were randomized to plots within a single subpopulation-specific region (as depicted in Figure 1) to facilitate mapping of quantitative trait loci separately for each subpopulation. In our combined analysis of data from all RILs, we expect correlations among the phenotypic values for RILs within each subpopulation due to both region and subpopulation effects, as well as spatially correlated plot effects, which induce correlations among phenotypic values within any field regardless of subpopulation membership.

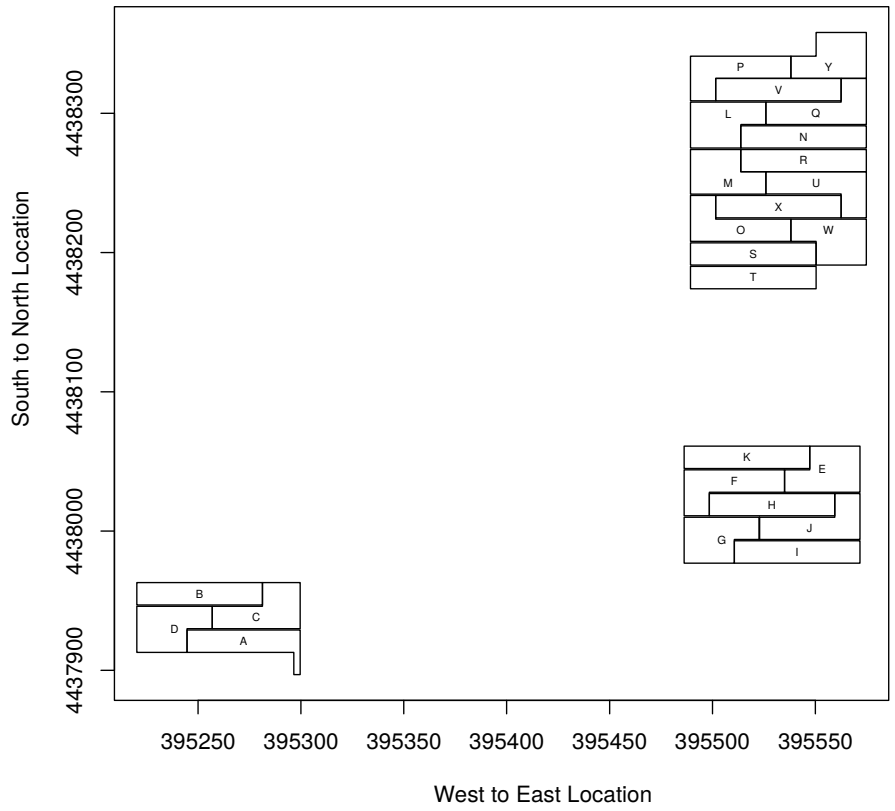

Figure 1: Geographic locations of 25 subpopulations labeled A through Y.

Our second dataset (henceforth labeled Data2) is the 2011 wheat dataset presented in Lado et al. (2013). This dataset contains results for 384 wheat lines genotyped at 102324 bi-allelic markers and phenotyped for grain yield (GY), thousand kernel weight (TKW), the number of kernels per spike (NKS), and days to heading (DH) under two levels of water supply: mild water stress (MWS) and fully irrigated (FI). For both MWS and FI, the 384 wheat lines were planted in an alpha-lattice design with 20 incomplete blocks of size 20 and two complete replications. Within each replications, 382 genotypes were planted on one plot each, while 2 of the 384 genotypes were planted on 9 plots each to cover all 20×20 plots. Lado et al. (2013) also analyzed data collected in 2012 from two separate locations, but the 2012 data contain

measurements of only the grain yield phenotype. We restrict our analysis to the 2011 data only to simplify our presentation.

3 Methods

3.1 Models

We are given a training dataset $\{y_i, \boldsymbol{x}_i, b_i, \boldsymbol{s}_i\}_{i=1}^n$, where $y_i \in \mathbb{R}$ represents a phenotype measurement, $\boldsymbol{x}_i \in \mathcal{X}$ is the corresponding p-dimensional vector of binary marker genotypes, $b_i \in \mathcal{B}$ is the corresponding subpopulation family index of the observation and $\boldsymbol{s}_i \in \mathcal{S}$ is the corresponding spatial location of the observation. Here $\mathcal{X}$, $\mathcal{B}$ and $\mathcal{S}$ represent the sets of possible values of binary marker genotype vectors, subpopulation family indices and spatial locations, respectively.

We propose a Gaussian random field (GRF) approach that carefully models (i) genotype effects, (ii) subpopulation effects, and (iii) spatial effects. More specifically, for $i = 1, \ldots, n$, suppose

$$y_i = \mathbb{Z}\left(\boldsymbol{t}_i\right) + \epsilon_i, \tag{1}$$

where $\boldsymbol{t}_i = (\boldsymbol{x}_i^\intercal, b_i, \boldsymbol{s}_i^\intercal)^\intercal$, $\mathbb{Z}(\boldsymbol{t}_i)$ is an observation at $\boldsymbol{t}_i$ of a GRF $\mathbb{Z}$ defined over index domain $\mathcal{T} = \mathcal{X} \times \mathcal{B} \times \mathcal{S}$, and ϵ_i is a mean zero Gaussian random variable independent of $\mathbb{Z}$. Further, we let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\intercal$ and assume $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma_\epsilon^2 \boldsymbol{I}_{n\times n}$ with $\boldsymbol{I}_{n\times n}$ being the identity matrix of size n. We assume a constant mean function for $\mathbb{Z}$, i.e., $\mathbb{E}(\mathbb{Z}(\boldsymbol{t})) = \mu$ for any $\boldsymbol{t} \in \mathcal{T}$. The power of this model lies in the flexible modeling of the covariance structure of $\mathbb{Z}$.

We consider an additive model for the covariance function that accounts for the three major effects. Specifically, for any $\boldsymbol{t}_i = (\boldsymbol{x}_i^\intercal, b_i, \boldsymbol{s}_i^\intercal)^\intercal, \boldsymbol{t}_k = (\boldsymbol{x}_k^\intercal, b_k, \boldsymbol{s}_k^\intercal)^\intercal \in \mathcal{T} = \mathcal{X} \times \mathcal{B} \times \mathcal{S}$, we assume

$$\mathrm{Cov}[\mathbb{Z}(\boldsymbol{t}_i), \mathbb{Z}(\boldsymbol{t}_k)] = C\left(\boldsymbol{t}_i, \boldsymbol{t}_k\right) = \sigma_g^2 C_g\left(\boldsymbol{x}_i, \boldsymbol{x}_k\right) + \sigma_b^2 C_b\left(b_i, b_k\right) + \sigma_s^2 C_s\left(\boldsymbol{s}_i, \boldsymbol{s}_k\right),$$

where σ_g^2, σ_b^2 and σ_s^2 are variance components and $C_g : \mathcal{X}^2 \to \mathbb{R}$, $C_b : \mathcal{B}^2 \to \mathbb{R}$ and $C_s : \mathcal{S}^2 \to \mathbb{R}$ are unit-diagonal kernel functions that quantify the corresponding dependence structures arising from similarity among observations with respect to genetic markers, subpopulations and spatial locations, respectively. Equivalently, we assume that the GRF $\mathbb{Z}$ can be decomposed into $\mathbb{Z}(\boldsymbol{t}_i) = \mu + \mathbb{Z}_g(\boldsymbol{x}_i) + \mathbb{Z}_b(b_i) + \mathbb{Z}_s(\boldsymbol{s}_i)$, where $\mathbb{Z}_g$, $\mathbb{Z}_b$, $\mathbb{Z}_s$ are mean zero

Gaussian random fields with covariance structures determined by $\sigma_g^2 C_g$, $\sigma_b^2 C_b$ and $\sigma_s^2 C_s$, respectively. We quantify the strength of spatial effects relative to the effects associated with marker genotypes by the variance component ratio $\gamma = \sigma_s^2/\sigma_g^2$.

3.2 Marker Kernel C_g

Following Morota and Gianola (2014) and references therein, we choose the Gaussian kernel

$$C_g\left(\boldsymbol{x}_i, \boldsymbol{x}_k\right) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2}{\tau}\right), \text{ for any } \boldsymbol{x}_i, \boldsymbol{x}_k \in \mathcal{X}$$

where $\|\cdot\|$ represents the Euclidean norm and τ is a parameter greater than zero.

Compared with other common kernels, the Gaussian kernel has been empirically shown to give robust and strong predictive performance. In Ober et al. (2011), the more general Matérn kernel is studied, but the Gaussian kernel performed best among the Matérn family based on their simulation study. Since the marker genotypes take discrete values, there is a temptation to choose a kernel on discrete index space. In Morota et al. (2013), a discretized Gaussian kernel, referred to as a diffusion kernel, was applied to dairy and wheat data for predicting phenotypes using marker information. However, the predictive power of such a kernel was similar to the Gaussian kernel.

Current high-throughput genotyping technology can provide genotype calls for hundreds of thousands of SNPs. Since most SNPs are unassociated with phenotype or conditionally unassociated with phenotype given other SNPs, $C_g(\boldsymbol{x}_i, \boldsymbol{x}_k)$ does not necessarily provide a good representation of correlation between the i-th and k-th lines when all SNPs are included in the vector of marker genotypes. To reduce computation time and improve genomic prediction, we use FarmCPU (Liu et al., 2016) to select important SNPs for inclusion in $\boldsymbol{x}_i$ rather than using the entire ensemble of SNPs. The details of our SNP selection procedure are discussed in Section 8.1 of the supplementary material.

3.3 Subpopulation Kernel C_b

The subpopulation GRF $\mathbb{Z}_b$ is motivated by genetic heterogeneity across different subpopulations and genetic similarity within subpopulations that may not be fully captured by SNP genotypes. We consider $C_b(b_i, b_k) = \mathbb{1}(b_i = b_k)$ for any $b_i, b_k \in \mathcal{B}$, where $\mathbb{1}(\cdot)$ is the indicator function. This covariance structure is equivalent to that induced by a model with independent, constant-variance subpopulation random effects.

3.4 Spatial Kernel C_s

In an agricultural field trial, plots are typically embedded in a regular rectangular array with say m_1 rows and m_2 columns. To adjust for spatial effects that may exist in such trails, the class of spatial autoregressions on regular rectangular lattice has been quite popular following the works of Besag and Green (1993); Besag et al. (1995); Besag and Higdon (1999) and Dutta and Mondal (2015). In this work we focus on the class of stationary autoregressions with modification described as follows. Consider a bigger array with $m_1' = m_1 + 4$ rows and $m_2' = m_2 + 4$ columns obtained by adding two virtual plots (Besag and Higdon, 1999) to each boundary to reduce the boundary effects. Suppose that for any positive integer k, $\boldsymbol{W}_k$ denotes the $k \times k$ matrix with

$$W_k(1,1) = W_k(k,k) = 1,\ W_k(i,i) = 2(1 < i < k),\ W_k(i,i+1) = W_k(i+1,i) = -\mathbb{1}(1 \le i < k),$$

and $W_k(i,j) = 0$ otherwise. Here $W_k(i,j)$ represents the (i,j)-th entry of $\boldsymbol{W}_k$. Next define $\boldsymbol{N}_{01} = \boldsymbol{I}_{m_2'} \otimes \boldsymbol{W}_{m_1'}$, $\boldsymbol{N}_{10} = \boldsymbol{W}_{m_2'} \otimes \boldsymbol{I}_{m_1'}$, and

$$\boldsymbol{W} = \beta_{00}\boldsymbol{I} + \beta_{01}\boldsymbol{N}_{01} + \beta_{10}\boldsymbol{N}_{10},$$

where β_{00}, β_{01} and β_{10} are positive parameters with $\beta_{00} + 2(\beta_{01} + \beta_{10}) = 1$. Let $\boldsymbol{D}$ be the diagonal matrix consisting of the diagonal entries of $\boldsymbol{W}^{-1}$. Next, for any plot s_i suppose $\boldsymbol{h}_i = \boldsymbol{h}(s_i)$ denotes the incidence vector of length $m_1'm_2'$. That is, the j-th entry of $\boldsymbol{h}_i$ is 1 if and only if s_i corresponds to the jth plot in in the $m_1' \times m_2'$ array where the plots in the array are enumerated in a column major format; the rest of the entries of $\boldsymbol{h}_i$ are zeros.

Finally, the spatial covariance of $(\mathbb{Z}_s(s_1), \ldots, \mathbb{Z}_s(s_n))^\intercal$ is given by

$$\sigma^2 C_s(s_i, s_j) = \sigma_s^2 \boldsymbol{h}_i^\intercal \boldsymbol{D}^{-1/2} \boldsymbol{W}^{-1} \boldsymbol{D}^{-1/2} \boldsymbol{h}_j,$$

where σ_s^2 is the spatial variance component introduced in Section 3. The advantage of this spatial kernel is that it makes the marginal variances at observed plots constant, while keeping the pairwise correlations the same as those that would be obtained from the stationary autoregression covariance matrix $\boldsymbol{W}^{-1}$. However, note that the weights on the neighbors are no longer β_{01} and β_{10} but are approximately proportional to them at the interior plots because the variances at the interior plots are approximately constant (Besag and Kooperberg, 1995).

The anisotropy parameters β_{01} and β_{10} play an important role because these parameters are related to the field geometry. In fact, McCullagh and Clifford (2006) found substantial empirical evidence that the non-anthropogenic variability in field trials can be explained by an isotropic spatial process with correlation decaying approximately logarithmically with distance. This would imply, for example, that for square plots the values of β_{01} and β_{10} should be approximately equal. On the other hand, if the plots are rectangular and the spacing between the plots is negligible compared with the plot sizes, the ratio of the β_{01} and β_{10} should be close to the aspect ratio of the plots. Also if the design is single column replication (see the El Batán trial in Besag and Higdon (1999)), then β_{10} is *zero*. In practice the estimates of β_{01} and β_{10} are automatically adjusted to the plot geometry and the inter-plot spacing. For Data1, in our context, the spatial layout in the maize experiment mimics a single-column replicate design because the distance between two east-west neighboring maize plants is much larger than the distance between two north-south neighbors. Thus we apriori expect the estimate of $\beta_{10} \approx 0$. This expectation is corroborated by the ML estimates in Section 4.2, where we see that the MLE of β_{10} occurs at the boundary. On the other hand, the plots in the Data2 are rectangular, and the inter-plot spacings are not very large. Thus we expect the MLE of β_{10} and β_{01} to be somewhere between 0 and 0.5 and this is corroborated by the estimates in Section 4.3.

The parameter β_{00}, on the other hand controls, the strength of the neighboring correlations and the range of the correlation. Interestingly, the boundary value of $\beta_{00} = 0$ gives rise to an intrinsic autoregression process and is the focus of Besag et al. (1995), Besag and Higdon (1999), and Dutta and Mondal (2015) in the context of fertility adjustments in agricultural variety trials. In particular, the foundational work of McCullagh and Clifford (2006) and empirical evidence from Besag et al. (1995), Besag and Higdon (1999), and Dutta and Mondal (2015, 2016) advocate the use of the intrinsic model for spatial adjustments in agricultural trials. Consequently, to

build a proper covariance model and to avoid boundary issues in maximum likelihood estimation, we fix the parameter β_{00} at a small value. To that end, following the suggestion in Besag and Kooperberg (1995), we numerically compute the neighboring correlations for various values of β_{00} (with $\beta_{01} = \beta_{10} = (1 - \beta_{00})/2$). We observe that the theoretical neighboring correlation changes by 11.26% when β_{00} changes from 0.01 to 0.001 and only by 1.69% when the β_{00} changes from 0.001 to 0.0001. A similar conclusion is obtained where β_{01} is held fixed at other values including 0 and 0.5. Because a change of 1.69% in neighboring correlation is practically negligible, we choose to fix the value of β_{00} at 0.001. Our analyses (see supplementary materials Table 4) also shows that the prediction accuracies and ranking do not change appreciably by changing β_{00} from 0.001 to 0.0001.

We end this section with references to other commonly used spatial kernels in such prediction problems. Rather than using the autoregression models to fit the spatial effects, several works (Crossa et al., 2006; Lado et al., 2013; Bernal-Vasquez et al., 2014) considered them as random effects in simple mixed linear models. In the context of agricultural field trials, Gleeson and Cullis (1987), Cullis and Gleeson (1991), Zimmerman and Harville (1991), Gilmour et al. (1995), Gilmour et al. (1997) and Cullis et al. (1998) developed more sophisticated spatial models for fertility adjustments. Although these models draw criticism due to their heavy dependences on the coordinate system by McCullagh and Clifford (2006), they have been quite effective in practice for spatial adjustments. However, their performances in the context of genomic prediction problems remain to be tested.

3.5 Estimation

For the training dataset $\{y_i, \boldsymbol{x}_i, b_i, \boldsymbol{s}_i\}_{i=1}^n$, let $\boldsymbol{C}$ be the $n \times n$ matrix with element i, j equal to $C(\boldsymbol{t}_i, \boldsymbol{t}_j)$ for all $i, j = 1, \ldots, n$. Define $\boldsymbol{C}_g$, $\boldsymbol{C}_b$ and $\boldsymbol{C}_s$ analogously. Then the covariance matrix of the vector of n phenotypic response values $\boldsymbol{y} = (y_1, \ldots, y_n)^\intercal$ can be written as

$$\boldsymbol{\Sigma} = \boldsymbol{C} + \sigma_\epsilon^2 \boldsymbol{I}_{n\times n} = \sigma_g^2 \boldsymbol{C}_g + \sigma_b^2 \boldsymbol{C}_b + \sigma_s^2 \boldsymbol{C}_s + \sigma_\epsilon^2 \boldsymbol{I}_{n\times n}.$$

The variance-covariance matrix $\boldsymbol{\Sigma}$ is a function of the parameters σ_g, τ, σ_b, σ_s and σ_ϵ. We maximize the log-likelihood to estimate these five parameters simultaneously.

It is straightforward to show that, for any given value of $\boldsymbol{\Sigma}$, the likelihood is maximized over μ at $\widehat{\mu} = \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y} / \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}$. Thus, the corresponding

profile log-likelihood function is

$$\ell\left(\sigma_g, \tau, \sigma_b, \sigma_s, \sigma_\epsilon\right) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{\left(\boldsymbol{y} - \widehat{\mu}\mathbf{1}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \widehat{\mu}\mathbf{1}\right)}{2}.$$

Finding maximizers of this profile log-likelihood function yields maximum likelihood estimates (MLEs) $\widehat{\sigma}_g, \widehat{\tau}, \widehat{\sigma}_b, \widehat{\sigma}_s$, and $\widehat{\sigma}_\epsilon$. Let $\widehat{\boldsymbol{C}}_g$ and $\widehat{\boldsymbol{\Sigma}}$ be the estimates of the covariance structures $\boldsymbol{C}_g$ and $\boldsymbol{\Sigma}$ obtained by replacing the unknown parameters with their MLEs.

Considering the joint distribution of $\boldsymbol{y}$, $\boldsymbol{Z}_g = (\mathbb{Z}_g(\boldsymbol{x}_1), \ldots, \mathbb{Z}_g(\boldsymbol{x}_n))^\intercal$, $\boldsymbol{Z}_b = (\mathbb{Z}_b(b_1), \ldots, \mathbb{Z}_b(b_n))^\intercal$ and $\boldsymbol{Z}_s = (\mathbb{Z}_s(\boldsymbol{s}_1), \ldots, \mathbb{Z}_s(\boldsymbol{s}_n))^\intercal$, we have

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{Z}_g \\ \boldsymbol{Z}_b \\ \boldsymbol{Z}_s \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu\mathbf{1} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \sigma_g^2\boldsymbol{C}_g & \sigma_b^2\boldsymbol{C}_b & \sigma_s^2\boldsymbol{C}_s \\ \sigma_g^2\boldsymbol{C}_g & \sigma_g^2\boldsymbol{C}_g & \mathbf{0} & \mathbf{0} \\ \sigma_b^2\boldsymbol{C}_b & \mathbf{0} & \sigma_b^2\boldsymbol{C}_b & \mathbf{0} \\ \sigma_s^2\boldsymbol{C}_s & \mathbf{0} & \mathbf{0} & \sigma_s^2\boldsymbol{C}_s \end{bmatrix} \right).$$

Based on our MLEs, we can estimate the conditional mean and conditional variance of $\boldsymbol{Z}_g$, $\boldsymbol{Z}_b$ and $\boldsymbol{Z}_s$ given $\boldsymbol{y}$, by

$$\widehat{\mathbb{E}}\left(\begin{bmatrix} \boldsymbol{Z}_g \\ \boldsymbol{Z}_b \\ \boldsymbol{Z}_s \end{bmatrix} \middle| \boldsymbol{y} \right) = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \widehat{\sigma}_g^2\widehat{\boldsymbol{C}}_g \\ \widehat{\sigma}_b^2\boldsymbol{C}_b \\ \widehat{\sigma}_s^2\boldsymbol{C}_s \end{bmatrix} \widehat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \widehat{\mu}\mathbf{1}) \tag{2}$$

and

$$\widehat{\mathrm{Var}}\left(\begin{bmatrix} \boldsymbol{Z}_g \\ \boldsymbol{Z}_b \\ \boldsymbol{Z}_s \end{bmatrix} \middle| \boldsymbol{y} \right) = \begin{bmatrix} \widehat{\sigma}_g^2\widehat{\boldsymbol{C}}_g & 0 & 0 \\ 0 & \widehat{\sigma}_b^2\boldsymbol{C}_b & 0 \\ 0 & 0 & \widehat{\sigma}_s^2\boldsymbol{C}_s \end{bmatrix} - \begin{bmatrix} \widehat{\sigma}_g^2\widehat{\boldsymbol{C}}_g \\ \widehat{\sigma}_b^2\boldsymbol{C}_b \\ \widehat{\sigma}_s^2\boldsymbol{C}_s \end{bmatrix} \widehat{\boldsymbol{\Sigma}}^{-1} \begin{bmatrix} \widehat{\sigma}_g^2\widehat{\boldsymbol{C}}_g & \widehat{\sigma}_b^2\boldsymbol{C}_b & \widehat{\sigma}_s^2\boldsymbol{C}_s \end{bmatrix}. \tag{3}$$

4 Empirical Study

4.1 Existing Methods

For the purpose of benchmarking, we compared our method with methods based on the Compressed Mixed Linear Model (CMLM) (Zhang et al., 2010) implemented in the GAPIT R package (Lipka et al., 2012), the Row and Column Model (RC) (Bernal-Vasquez et al., 2014) and the linear regression

with moving means as covariable model (MVNG) (Lado et al., 2013). These competing methods are described in the following.

Compressed Mixed Linear Model: Let $\boldsymbol{M}$ be a matrix whose columns correspond to the first few principal components (usually 3 or 5 by default) computed from the binary genotype matrix to represent population structure. The compressed mixed linear model is

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{Z}\bar{\boldsymbol{u}} + \boldsymbol{e},$$

where $\bar{\boldsymbol{u}}_{r\times 1} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\bar{u}}\bar{\boldsymbol{K}}_{r\times r})$ represents an unknown vector of random additive genetic effects and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \sigma^2_e\boldsymbol{I})$ is the unobserved vector of errors. The random effects in $\bar{\boldsymbol{u}}_{r\times 1}$ are intended to represent the effects of multiple background quantitative trait loci (QTL) on the phenotypic response values. Note that $\bar{\boldsymbol{u}}_{r\times 1}$ is of dimension $r\times 1$ rather than $n\times 1$ as in the MLM because that $\bar{\boldsymbol{u}}_{r\times 1}$ represents different groups $t = 1, \ldots, r$ clustered according to a full kinship matrix $\boldsymbol{K}_{n\times n}$ rather than individuals/lines. Meanwhile, the matrix $\bar{\boldsymbol{K}}_{r\times r}$ is the corresponding kinship matrix that accounts for varying degrees of genetic similarity among groups rather than among individuals/lines. We adopt the formula for the full kinship matrix suggested by (VanRaden, 2008) where:

$$\boldsymbol{K}_{n\times n} = \frac{\widetilde{\boldsymbol{X}}^{(g)}\widetilde{\boldsymbol{X}}^{(g)\intercal}}{\sum_i 2p_i(1-p_i)}, \tag{4}$$

where $\widetilde{\boldsymbol{X}}^{(g)}$ contains allele calls centered so that each row sums to zero and p_i is the frequency of the minor allele at locus i. As for the group kinship matrix $\bar{\boldsymbol{K}}_{r\times r} = (\bar{K}_{st})$ where $s, t = 1$ to r, each of the entry $\bar{K}_{st}$ is defined as the average of a set of $\{K_{hj}\}$ where h belongs to group s and j belongs to group t. For the maize dataset Data1, the Bayesian information criterion (Zhang et al., 2010) selects no principle components in the matrix $\boldsymbol{M}$. For the wheat dataset Data2, we considered one, three, five and ten principle components for $\boldsymbol{M}$. We found no important difference and thus we adopted the default setting with the first three principle components in $\boldsymbol{M}$.

Incomplete Block Model: Motivated by the alpha-lattice experimental design underlying the wheat dataset, we also consider an incomplete block (IB) model defined as follows. Using the same principal component matrix $\boldsymbol{M}$ in CMLM, the IB model assumes

$$\boldsymbol{y} = \mu\mathbf{1} + \boldsymbol{M}\boldsymbol{\beta} + \boldsymbol{Z}_{\boldsymbol{u}_g}\boldsymbol{u}_g + \boldsymbol{Z}_{\boldsymbol{u}_{\text{rep}}}\boldsymbol{u}_{\text{rep}} + \boldsymbol{Z}_{\boldsymbol{u}_{\text{bl(rep)}}}\boldsymbol{u}_{\text{bl(rep)}} + \boldsymbol{e},$$

where $\boldsymbol{u}_g \sim \mathcal{N}(\boldsymbol{0}, \sigma_g^2 \boldsymbol{K})$, $\boldsymbol{u}_{\text{rep}} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\text{rep}}^2 \boldsymbol{I})$, $\boldsymbol{u}_{\text{bl(rep)}} \sim \mathcal{N}(\boldsymbol{0}, \sigma_{\text{bl(rep)}}^2 \boldsymbol{I})$ and $\boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$ represent independent, unknown vectors of additive genetic effects, replication effects, incomplete block effects and errors respectively. Here $\boldsymbol{K}$ is the full kinship matrix defined in (4). We applied this model to the wheat dataset Data2 with the first three principle components in $\boldsymbol{M}$. Because the experimental design that gave rise to the maize dataset involves no replication or blocking, the IB model is not applicable for Data1.

RC and MVNG: For the Row and Column Model (RC) and the linear regression with moving means as covariate model (MVNG), we propose two steps for the prediction as suggested by Lado et al. (2013). The idea is that we first adjust for spatial effects in the observed phenotypic response values, and then we provide genomic predictions by using the rrBLUP R package applied to the spatially adjusted phenotypic response values. Two different kernels, RR and GAUSS (Endelman, 2011), are considered for the genomic predictions.

In the first step, the RC model assumes that

$$y_{ijk} = \mu + \text{row}_i + \text{col}_j + \text{sub}_k + e_{ijk},$$

where row_i (row effect), col_j (column effect) and sub_k (subpopulation effect) are considered as independent random effects with mean-zero normal distributions that have variances specific to the effect type (i.e., one variance for row effects, one for column effects and one for subpopulation effects). For the adjustment, we have $\widehat{y}_{ijk} = y_{ijk} - \widehat{\text{row}}_i - \widehat{\text{col}}_j$ where $\widehat{\text{row}}_i = \widehat{\mathbb{E}}(\text{row}_i|\boldsymbol{y})$ and $\widehat{\text{col}}_j = \widehat{\mathbb{E}}(\text{col}_j|\boldsymbol{y})$ are the corresponding empirical Best Linear Unbiased Predictors (eBLUPs) of row_i and col_j effects.

For MVNG, we adopt the same idea in Lado et al. (2013), namely, we fit the model

$$y_i = \mu + \beta x_i + e_i,$$

where $x_i = y_i - \frac{1}{6}\sum_{k=1}^{6} y_i^{(k)}$ with $y_i^{(k)}$, $k = 1, \ldots, 6$, the phenotypic response values for the spatial neighbors (one up, one down, two left, and two right) of the i-th observation (See Figure 1 in Lado et al. (2013) for details). For Data2, as suggested by Lado et al. (2013), left-right corresponds to spatial neighbors within each row and up-down corresponds to spatial neighbors within each column. For Data1, based on the observation that east-west neighbors are much farther apart than north-south neighbors, we adopt north-south as

left-right and east-west as up-down in this MVNG method. The spatially adjusted values for i-th observation is given by $\widehat{y}_i = y_i - \widehat{\beta} x_i$.

In the second step, the genomic prediction is performed under the model

$$\widehat{\boldsymbol{y}} = \mu \mathbf{1} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

where $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \boldsymbol{K})$ represents an unknown vector of random additive genetic effects and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{I})$ is the unobserved vector of residuals. For kernel RR, $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\intercal$, where $\boldsymbol{X}$ is the original genotype matrix without scaling and centering. For kernel GAUSS, $\boldsymbol{K} = \boldsymbol{C}_g$, the parameter τ is estimated by residual maximum likelihood (REML).

4.2 Data1 Prediction

As described in Section 2, the maize dataset (Data1), can be naturally divided into three fields or into 25 subpopulations (see Figure 1). In this section, we provide evidence of both spatial effects and subpopulation effects in each field and evidence of spatial effects in each subpopulation. To provide such evidence, we fit three reduced versions of the full GRF model defined in Sections 3.2-3.4. For the dataset in each field, we fit both GRF_{-Z_b} and GRF_{-Z_s}, where the corresponding covariances are $\boldsymbol{\Sigma}_{-Z_b} = \sigma^2_{g_{-Z_b}} \boldsymbol{C}_g + \sigma^2_{s_{-Z_b}} \boldsymbol{C}_s + \sigma^2_{\epsilon_{-Z_b}} \boldsymbol{I}$ and $\boldsymbol{\Sigma}_{-Z_s} = \sigma^2_{g_{-Z_s}} \boldsymbol{C}_g + \sigma^2_{b_{-Z_s}} \boldsymbol{C}_b + \sigma^2_{\epsilon_{-Z_s}} \boldsymbol{I}$, respectively; i.e., we ignore subpopulation effects in GRF_{-Z_b} and spatial effects in GRF_{-Z_s}. For any dataset consisting of a single subpopulation, we drop subpopulation effects and fit $\text{GRF}_{-Z_{bs}}$ instead of GRF_{-Z_s}, where $\boldsymbol{\Sigma}_{-Z_{bs}} = \sigma^2_{g_{-Z_{bs}}} \boldsymbol{C}_g + \sigma^2_{\epsilon_{-Z_{bs}}} \boldsymbol{I}$.

In the following, we report the performance of CMLM, RC(RR, GAUSS), MVNG(RR, GAUSS), GRF_{-Z_b}, GRF_{-Z_s}, $\text{GRF}_{-Z_{bs}}$ and the full GRF based on analysis of 1000 independent random partitions of the data in each subpopulation into training (80%) and test (20%) sets. When performing analysis at the field level, we combine the training sets from all subpopulations in a field to form one training set and likewise pool the corresponding subpopulation-specific test sets to form a field-specific test set.

To evaluate the performance of different methods, we consider the accuracy defined as the correlation between predicted response values and observed phenotypic response values in the test set. In Table 1, we report the accuracies for each field, along with estimates of $\widehat{\gamma} = \widehat{\sigma}^2_s / \widehat{\sigma}^2_g$ based on the whole dataset (without splitting). Due to space limitation, the detailed results for each subpopulation are delegated to Table 5 of the supplementary

material. The magnitude of $\widehat{\gamma}$ indicates the estimated strength of spatial effects relative to genotypic variation.

As we can see in Table 1 (and also Table 5), the GAUSS kernel is inferior to the RR kernel in both RC and MVNG results. Thus we present only RC(RR) and MVNG(RR) results in subsequent figures. For each subpopulation, Figure 2 (1-3) shows the comparison of CMLM, RC(RR), MVNG(RR) and the two proposed methods $\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b}.

For most subpopulations, the accuracy of GRF_{-Z_b} is higher than the corresponding accuracies of the existing methods. When the accuracy of GRF_{-Z_b} is close to or lower than accuracies of existing methods, the estimated strength of spatial effects $\widehat{\gamma}$ is close to 0. For the subpopulations with strong spatial effects, it is reasonable that the predictions can be improved relative to CMLM (which ignores spatial effects) by incorporating the spatial kernel $\boldsymbol{C}_s$. Since there is little evidence of horizontal spatial correlation, RC(RR) and MVNG(RR) are based on misspecified spatial models which lead to lower accuracy. For the subpopulations with weak or no spatial effects, accuracy of predictions may be degraded by inclusion of $\boldsymbol{C}_s$ in the model. Comparing CMLM and GRF_{-Z_s} (the methods that ignore spatial effects), we can see that GRF_{-Z_s} has slightly lower average accuracies for many subpopulations. A possible explanation is that CMLM makes greater use of the SNP information. While SNP information enters the marker kernel of GRF_{-Z_s} via simple Euclidean distances, CMLM utilizes this information in both fixed effects and random effects. Specifically, CMLM allows for fixed effects of the PCs of SNPs and adopts the corresponding kinship matrix as the variance-covariance structure for random effects. This may also be the reason that the GAUSS kernel is inferior to the RR kernel in both RC and MVNG methods.

For the field-level analysis, we are able to use the full GRF that includes genotype, subpopulation and spatial effects. Figure 2 (4) and Table 1 show that the full GRF has the highest average accuracy across all methods for every field. These results illustrate that the full GRF can effectively account for heterogeneity across genotype, subpopulation and spatial location effects at the field scale to enhance prediction accuracy.

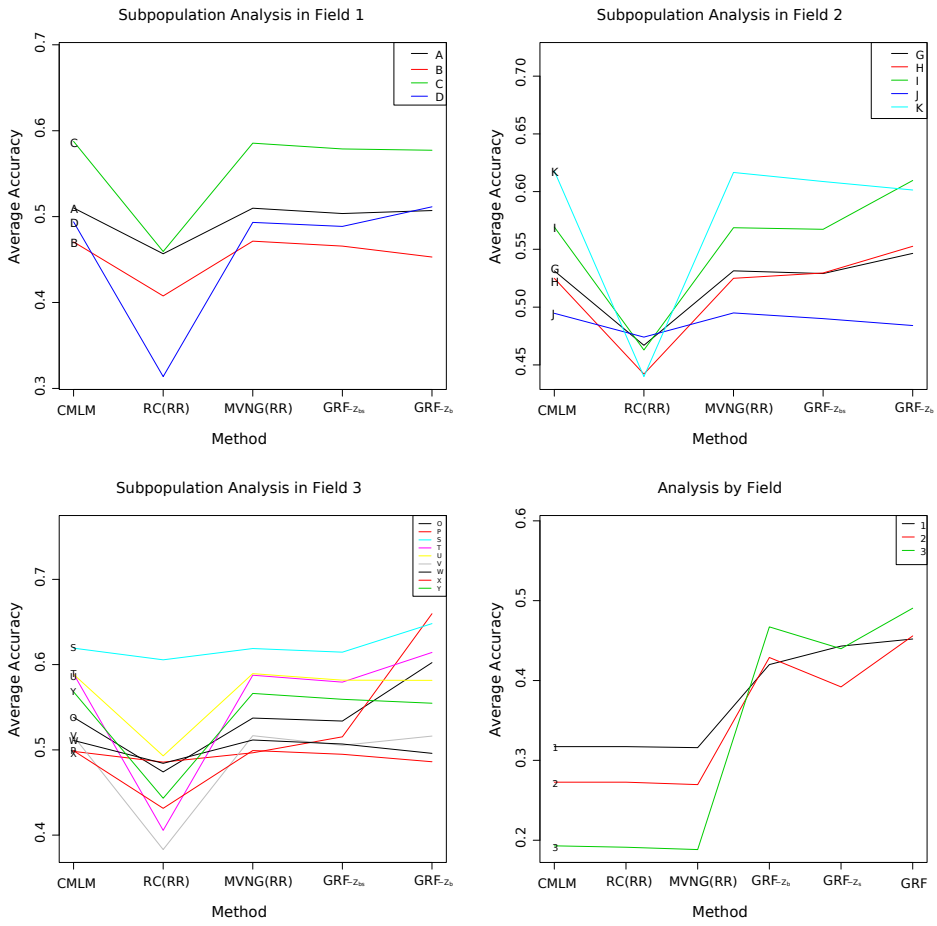

Figure 2: (1-3): Comparison of CMLM, RC(RR), MVNG(RR) and two proposed methods $\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b} for each subpopulation (only 9 of 14 are shown in Field 3 to improve clarity). (4): Comparison of CMLM, RC(RR), MVNG(RR) and three proposed methods GRF_{-Z_b}, GRF_{-Z_s} and GRF for each field.

Table 1: Average accuracies for Data1 by five existing methods and three proposed methods (GRF_{-Z_b}, GRF_{-Z_s} and GRF) for each field based on 1000 independent random partitions of the data into training (80%) and test (20%) sets. The highest average accuracy across methods for each field is shown in bold.

	Method								
Field	CMLM	RC		MVNG		GRF_{-Z_b}	GRF_{-Z_s}	GRF	$\widehat{\gamma} = \widehat{\sigma}_s^2/\widehat{\sigma}_g^2$
		RR	GAUSS	RR	GAUSS				
1	0.3173	0.3173	0.3144	0.3159	0.3131	0.4199	0.4428	**0.4520**	0.0646
2	0.2727	0.2727	0.2729	0.2697	0.2698	0.4289	0.3920	**0.4558**	0.3041
3	0.1930	0.1913	0.1904	0.1883	0.1873	0.4672	0.4395	**0.4904**	1.0087

4.3 Data2 Prediction

For the wheat dataset Data2, there is no subpopulation information. Thus we do not need the component $\boldsymbol{Z}_b$ in the full GRF, and the corresponding subpopulation covariance structure $\boldsymbol{C}_b$ is ignorable. In the following, we report the performance of CMLM, GRF_{-Z_b} and $\text{GRF}_{-Z_{bs}}$ based on 1000 independent training-test partitions for the eight phenotypes in the wheat dataset Data2. Similarly, due to space limitation, the detailed results are delegated to Table 6 of the supplementary material. In addition to the prediction results, the corresponding parameter estimations are reported in Table 7 in the supplementary material as well. We compare the performance of these three methods directly with results for other methods in Table 3 of Lado et al. (2013). For each partition, we split the dataset into training (86%) and test (14%) sets to match the same settings used in Lado et al. (2013). Note that Lado et al. (2013) presented results for an inferior-performing version of our IB approach that involved using genomic prediction techniques on the residuals from the fit of the IB model without genomic information. Results labeled IB in this paper refer to our implementation of the IB model described in Section 4.

Figure 3 shows the comparison of CMLM, IB, GRF_{-Z_b} and $\text{GRF}_{-Z_{bs}}$. It is noted that GRF_{-Z_b} performs best and CMLM performs worst due to the existence of strong spatial effects. For the phenotype grain yield (GY) in Santa Rosa under two levels of water supply, mild water stress (MWS) and fully irrigated (FI), the estimated relative strength of spatial effects $\widehat{\gamma}$ is 2.4231 and 4.5171, respectively. For these phenotypes, the accuracy difference between GRF_{-Z_b} and $\text{GRF}_{-Z_{bs}}$ is much larger than for other phenotypes. Compar-

isons with Table 3 in Lado et al. (2013) show that GRF$_{-Z_b}$ performed the best among all the methods in terms of accuracies. Figure 3 (and also Tables 6 and 7) indicates that none of the results are sensitive to selection of SNPs prior to model fitting and analysis.

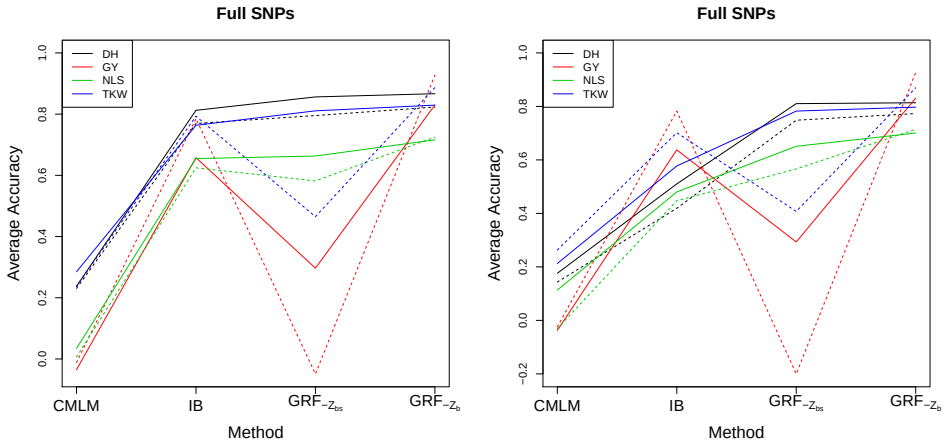

Figure 3: Comparisons of CMLM, IB and two proposed methods GRF$_{-Z_{bs}}$ and GRF$_{-Z_b}$ with full and selected SNPs. The solid lines correspond to different phenotypes under fully irrigated (FI) conditions while dashed lines represent the same phenotypes under mild water stress (MWS).

5 Simulation Study

This section reports results from simulation experiments designed to evaluate numerical performance of genomic predictions after adjusting for spatial effects.

5.1 Data1 Ranking

From the maize dataset Data1, we fit the full GRF model to obtain parameter estimates $\widehat{\mu}, \widehat{\sigma}_g, \widehat{\tau}, \widehat{\sigma}_b, \widehat{\sigma}_s$, and $\widehat{\sigma}_\epsilon$. These estimates provide $\widehat{\boldsymbol{C}}_g$ and $\widehat{\boldsymbol{\Sigma}}$, which determine the estimated mean and variance of the conditional multivariate normal distribution for $\boldsymbol{Z}_g$, $\boldsymbol{Z}_b$ and $\boldsymbol{Z}_s$ according to equations (2) and (3). Given these estimated parameters, let $\widetilde{\boldsymbol{Z}}_g$, $\widetilde{\boldsymbol{Z}}_b$ and $\widetilde{\boldsymbol{Z}}_s$ be generated simultaneously from a multivariate normal distribution where the mean and

variance are specified in (2) and (3). And let $\tilde{\boldsymbol{e}}$ be generated from $\mathcal{N}(\mathbf{0}, \widehat{\sigma}_\epsilon^2 \boldsymbol{I})$. To allow different strengths of spatial effects, we simulate the response vector $\tilde{\boldsymbol{y}} = \widehat{\mu}\mathbf{1} + \widetilde{\boldsymbol{Z}}_g + \widetilde{\boldsymbol{Z}}_b + c\widetilde{\boldsymbol{Z}}_s + \tilde{\boldsymbol{e}}$, where $c \in \{1, 2, 3, 4\}$ controls the strength of spatial effects. Given a simulated dataset, we fit the full GRF, GRF_{-Z_s} and GRF_{-Z_b} to predict $\tilde{\boldsymbol{y}}$. We repeat this simulation and fitting process 1000 times.

In addition to prediction accuracy, we also compare the ability to rank plant genotypes. We compare the true rank-order $\boldsymbol{r}^{(o)}$ of the elements of $\widehat{\mu}\mathbf{1} + \widetilde{\boldsymbol{Z}}_g + \widetilde{\boldsymbol{Z}}_b$, with the rank-orders $\boldsymbol{r}^{(\text{GRF})}$, $\boldsymbol{r}^{(\text{GRF}_{-Z_s})}$ and $\boldsymbol{r}^{(\text{GRF}_{-Z_b})}$ of the predictions by computing Spearman's rank-order correlations $\rho_s(\boldsymbol{r}^{(o)}, \boldsymbol{r}^{(\text{GRF})})$, $\rho_s(\boldsymbol{r}^{(o)}, \boldsymbol{r}^{(\text{GRF}_{-Z_s})})$, and $\rho_s(\boldsymbol{r}^{(o)}, \boldsymbol{r}^{(\text{GRF}_{-Z_b})})$ for each simulation replication.

Table 2 reports both the prediction accuracies and Spearman's rank-order correlations. These two measurements are highly correlated. The full GRF is much better than GRF_{-Z_s} and GRF_{-Z_b} in terms of prediction accuracies and the similarities of rank-orders with the true rank-order $\boldsymbol{r}^{(o)}$. Because spatial effects and subpopulation effects for Data1 in each field are strong enough ($\widehat{\gamma} = 0.0646$, $\widehat{\sigma}_b = 0.3581$; $\widehat{\gamma} = 0.3041$, $\widehat{\sigma}_b = 0.3526$ and $\widehat{\gamma} = 1.0087$, $\widehat{\sigma}_b = 0.3939$ respectively for the three fields.) With spatial strength held constant, prediction performance in Table 2 improves across fields in accordance with the number of observations per field, likely due to the improvement of estimation with more data.

Table 2: Average prediction accuracies and Spearman's rank-order correlations (ρ_s) based on 1000 simulations for Data1 by the full GRF, GRF$_{-Z_s}$ and GRF$_{-Z_b}$ for different spatial strengths. The highest average accuracy and highest average rank-order correlation across methods for each combination of field and spatial strength are shown in bold.

Field	Strength	Accuracies			ρ_s		
		GRF	GRF$_{-Z_s}$	GRF$_{-Z_b}$	GRF	GRF$_{-Z_s}$	GRF$_{-Z_b}$
1	1	**0.8249**	0.8200	0.5041	**0.8076**	0.8036	0.4711
	2	**0.8069**	0.7853	0.4795	**0.7887**	0.7676	0.4459
	3	**0.7860**	0.7343	0.4672	**0.7659**	0.7169	0.4332
	4	**0.7632**	0.6738	0.4571	**0.7421**	0.6595	0.4229
2	1	**0.8395**	0.8317	0.5221	**0.8276**	0.8196	0.5008
	2	**0.8129**	0.7706	0.5070	**0.7995**	0.7563	0.4858
	3	**0.7847**	0.6900	0.4952	**0.7699**	0.6762	0.4740
	4	**0.7554**	0.6067	0.4836	**0.7395**	0.5956	0.4627
3	1	**0.9135**	0.9085	0.2835	**0.8986**	0.8934	0.2760
	2	**0.8818**	0.8505	0.2693	**0.8637**	0.8310	0.2621
	3	**0.8486**	0.7738	0.2601	**0.8286**	0.7512	0.2531
	4	**0.8164**	0.6940	0.2523	**0.7952**	0.6693	0.2453

For each simulated data set, we predict the top l inbred lines are by ranking our predictions of $\mathbb{Z}_g + \mathbb{Z}_b$, for $l \in \{1, \ldots, n\}$. We use T_l as notation for the predicted group of top l lines. Note that, due to estimation and prediction errors, the true rank-orders $\boldsymbol{r}_l^o$ of the lines in T_l may not be $1, \ldots, l$. We evaluate the accuracy by the average median of $\boldsymbol{r}_l^o$ over 1000 simulations. The smaller the average median is, the better the predicted group is. In the following, we study the accuracy of the first ten groups, $T_1, \ldots, T_{10}$, for different methods.

The right panels of Figure 4 shows the average median of $\boldsymbol{r}_l^{(o)}$ for $l = 1, \ldots, 10$ for the full GRF, GRF$_{-Z_s}$ and GRF$_{-Z_b}$ on each field while the left panels zoom in on the results for the full GRF and GRF$_{-Z_s}$. The horizontal axis represents different groups while the vertical axis represents the corresponding average median of $\boldsymbol{r}_I^{(o)}$. We can see in Figure 4 that the full GRF performs consistently better than GRF$_{-Z_s}$ and GRF$_{-Z_b}$ which suggests that accounting for either spatial or subpopulation effects improves selection of the best plant genotypes.

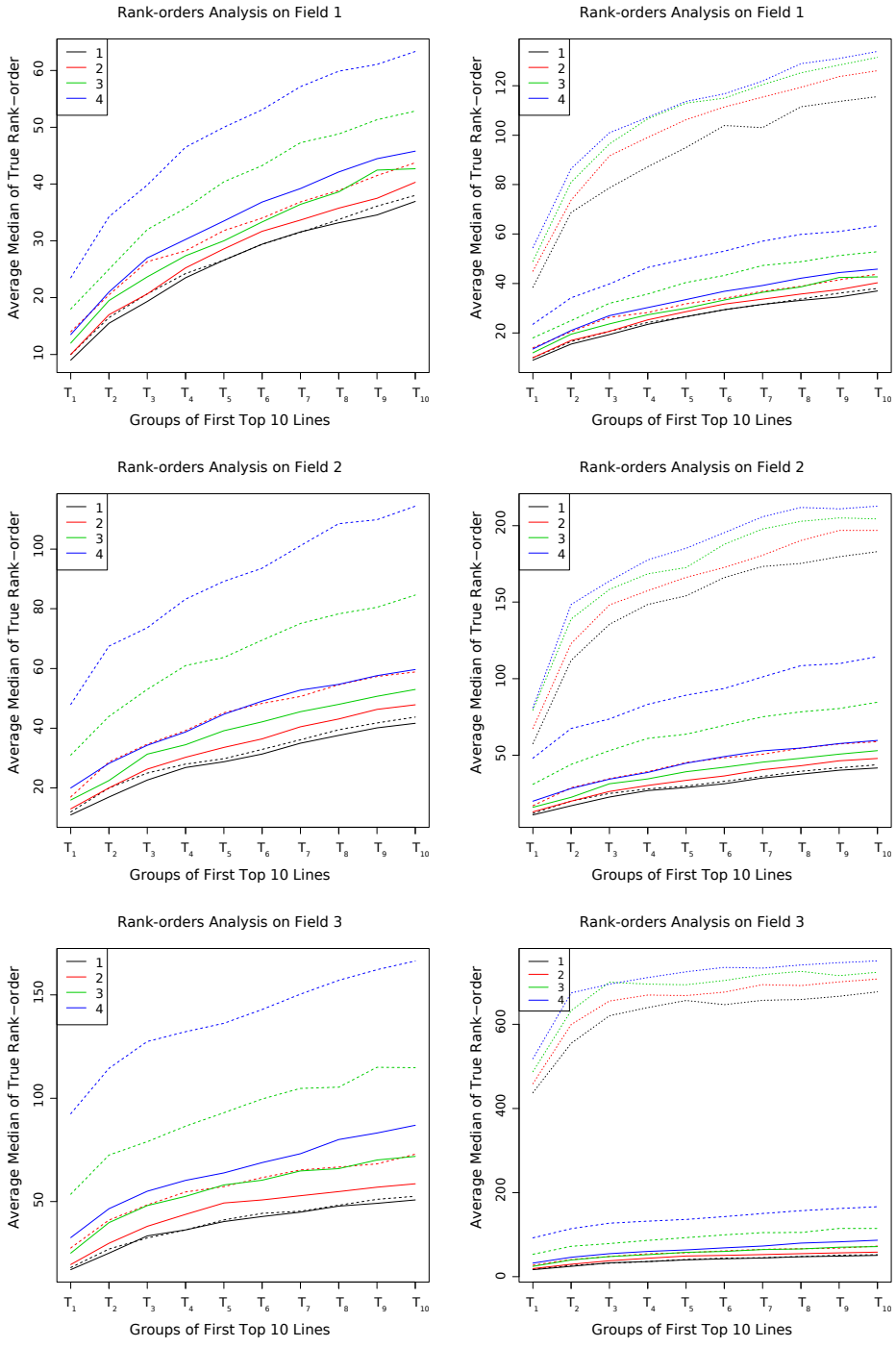

Figure 4: Left: Comparisons of the full GRF and GRF$_{-Z_s}$ for each field. Right: Comparisons of the full GRF, GRF$_{-Z_s}$ and GRF$_{-Z_b}$ for each field. The solid lines are for the full GRF, dashed lines represent for GRF$_{-Z_s}$ and dotted lines are for GRF$_{-Z_b}$.

5.2 Data2 Ranking

For the wheat dataset Data2, we report the performances of GRF_{-Z_b}, $\text{GRF}_{-Z_{bs}}$ and CMLM based on 1000 simulations, for each of the eight phenotypes. We achieve similar conclusions as in Data1. Due to space limitation, the details are delegated to Section 8.5 of the supplementary material.

6 Discussion

This paper investigates the problem of adjusting for spatial effects in genomic prediction. Our analysis of the maize dataset Data1 and the wheat dataset Data2 reveals the existence of spatial effects and heterogeneity across different subpopulation families. The spatial effects and heterogeneity, without proper treatment, can reduce the quality of phenotypic prediction and genotypic ranking. Under the Gaussian random field model, we propose an additive covariance matrix structure that incorporates genotype effects, spatial effects and subpopulation effects. We have also shown that by adjusting for spatial effects, we can improve the selection of top-performing plant genotypes.

7 Acknowledgment

The authors acknowledge the Iowa State University Plant Sciences Institute Scholars Program for financial support and the lab of Patrick S. Schnable and former graduate research assistant Sarah Hill-Skinner for collecting and sharing the maize data. This article is a product of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. IOW03617 is supported by USDA/NIFA and State of Iowa funds. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U.S. Department of Agriculture.

8 Appendix

8.1 Data Pre-Processing

In this section, we provide a detailed description of the pre-processing procedures for the maize dataset Data1 and the wheat dataset Data2.

For the maize dataset Data1, as mentioned in Section 2, we apply LD-kNNi to impute all missing SNP genotypes. Since there are only two alleles at each locus, we code the genotype from each SNP as a binary variable where "1" represents the major more prevalent allele and "0" represents the minor allele. After all missing SNP genotypes are imputed, we subdivide the 4660 observations into subsets corresponding to the three fields and, further, into subsets corresponding to the 25 subpopulations.

In each of these datasets, many SNP markers that are identical to other SNP markers for all observations. We keep only one representative SNP marker in such cases and remove redundant SNPs. However, a huge number of SNP markers are still remaining. It is not only a computational burden for the marker kernel $\boldsymbol{C}_g$, but it also incorporates lots of redundancy information. To get more useful marker kernel, we apply fixed and random effect model to select the important SNP markers. For the fixed effect model step, we test all the genetic markers, one at a time. In each test, we obtain a p-value. For those genetic markers with p-value larger than 0.05, we discard them and keep the remaining. The R package FarmCPU (Liu et al., 2016) is applied for this pre-processing. We repeat the same procedures for each field and each subpopulation to get different selections. Table 3 shows the total number of SNP markers before and after removing the duplicated vectors, the number of selected SNP markers for Data1 on Field 1.

Table 3: The total number of SNP markers before and after removing the duplicated vectors, the number of selected SNP markers for Data1 on Field 1.

Chromosome	1	2	3	4	5	6	7	8	9	10
total	104827	81315	78369	73466	67423	58365	59577	59820	54013	50694
unique	95065	71524	68790	64492	60645	51272	51803	52687	47238	44470
selected	470	2666	789	300	503	290	580	343	124	401

It is shown in Table 3 that, after the selection, the number of SNP markers are much reduced.

For the wheat dataset Data2, the analysis of Lado et al. (2013) is based on the data with full SNPs. For the completeness of our comparison, we provided both results with full and selected SNPs.

8.2 Sensitivity of parameter β_{00}

Table 4: The mean of accuracies for eight phenotypes by method GRF_{-Z_b} ($\lambda_{00} = 0.001$ and $\lambda_{00} = 0.0001$) with full and selected SNPs based on 1000 independent random partitions of the data into training (86%) and test (14%) sets.

		Full		Selected	
Water Supply	Phenotype	GRF_{-Z_b}		GRF_{-Z_b}	
	λ_{00}	0.001	0.0001	0.001	0.0001
FI	DH	0.8667	0.8668	0.8160	0.8185
	GY	0.8299	0.8305	0.8309	0.8314
	NLS	0.7160	0.7133	0.7004	0.6976
	TKW	0.8294	0.8284	0.7972	0.7968
MWS	DH	0.8231	0.8249	0.7732	0.7729
	GY	0.9275	0.9268	0.9269	0.9257
	NLS	0.7248	0.7244	0.7143	0.7140
	TKW	0.8867	0.8863	0.8696	0.8700

8.3 Data1 Prediction Results

Table 5: Average accuracies for Data1 by five existing methods and two proposed methods ($\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b}) for each subpopulation based on 1000 independent random partitions of the data into training (80%) and test (20%) sets. The highest average accuracy across methods for each subpopulation is shown in bold.

		Method							
Field	Subpopulation	CMLM	RC		MVNG		$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}	$\widehat{\gamma} = \widehat{\sigma}_s^2/\widehat{\sigma}_g^2$
			RR	GAUSS	RR	GAUSS			
1	A	**0.5101**	0.4568	0.4196	0.5098	0.5014	0.5036	0.5070	0.0220
	B	0.4706	0.4077	0.3544	**0.4715**	0.4610	0.4657	0.4530	9e-04
	C	**0.5875**	0.4594	0.4388	0.5855	0.5805	0.5788	0.5772	0.0000
	D	0.4939	0.3139	0.0779	0.4933	0.4803	0.4886	**0.5113**	0.0249
2	E	0.5300	0.2966	0.1407	0.5293	0.5295	**0.5377**	0.5308	0.0228
	F	0.4939	0.4102	0.3782	0.4925	0.4847	0.4812	**0.5564**	0.0994
	G	0.5314	0.4670	0.4424	0.5314	0.5278	0.5291	**0.5466**	0.0412
	H	0.5250	0.4421	0.3324	0.5250	0.5281	0.5295	**0.5527**	0.0704
	I	0.5694	0.4629	0.4235	0.5688	0.5689	0.5674	**0.6097**	0.0576
	J	0.4947	0.4741	0.4588	**0.4950**	0.4916	0.4901	0.4841	0.0285
	K	**0.6179**	0.4399	0.3275	0.6166	0.6044	0.6088	0.6014	0.0014
3	L	0.5116	0.4354	0.2586	0.5104	0.5102	0.5102	**0.5455**	0.0498
	M	**0.5036**	0.4427	0.4329	0.5024	0.5014	0.5005	0.5028	0.0276
	N	0.5162	0.1193	0.0646	0.5148	0.5131	0.5084	**0.5166**	0.0375
	O	0.5381	0.4741	0.2939	0.5373	0.5279	0.5338	**0.6024**	0.0670
	P	0.4983	0.4857	0.4807	0.4966	0.5048	0.5153	**0.6598**	0.3922
	Q	0.5562	0.4979	0.4632	0.5529	0.5508	0.5536	**0.5959**	0.1044
	R	0.5563	0.4380	0.1916	0.5567	0.5563	0.5551	**0.5802**	0.0650
	S	0.6193	0.6057	0.5845	0.6188	0.6166	0.6147	**0.6482**	0.0502
	T	0.5892	0.4056	0.3383	0.5875	0.5837	0.5795	**0.6143**	0.0242
	U	0.5886	0.4930	0.4807	**0.5894**	0.5795	0.5818	0.5815	0.0000
	V	0.5160	0.3830	0.3097	**0.5166**	0.5076	0.5054	0.5162	0.0202
	W	0.5110	0.4842	0.4325	**0.5116**	0.5049	0.5068	0.4959	0.0038
	X	0.4990	0.4314	0.4078	**0.4991**	0.4960	0.4947	0.4861	0.0120
	Y	**0.5680**	0.4432	0.3613	0.5662	0.5571	0.5593	0.5547	0.0015

8.4 Data2 Prediction Results

Table 6: Average accuracies for eight phenotypes by methods (IB, CMLM, $\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b}) with full and selected SNPs based on 1000 independent random partitions of the data into training (86%) and test (14%) sets. The highest average accuracy across methods for each of the three methods is printed in bold for each phenotype and each SNP set.

		Full				Selected			
Water Supply	Phenotype	CMLM	IB	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}	CMLM	IB	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}
FI	DH	0.2377	0.8124	0.8562	**0.8667**	0.1760	0.5103	0.8103	**0.8160**
	GY	-0.0351	0.6578	0.2968	**0.8299**	-0.0377	0.6374	0.2932	**0.8309**
	NLS	0.0344	0.6547	0.6631	**0.7160**	0.1131	0.4802	0.6509	**0.7004**
	TKW	0.2849	0.7635	0.8108	**0.8294**	0.2126	0.5772	0.7825	**0.7972**
MWS	DH	0.2307	0.7694	0.7953	**0.8231**	0.1432	0.4191	0.7483	**0.7732**
	GY	-0.0118	0.7852	-0.0490	**0.9275**	-0.0243	0.7827	-0.2004	**0.9269**
	NLS	0.0064	0.6246	0.5817	**0.7248**	-0.0307	0.4470	0.5663	**0.7143**
	TKW	0.2371	0.7930	0.4650	**0.8867**	0.2631	0.7018	0.4072	**0.8696**

Table 7: The spatial parameter estimates for eight phenotypes by method GRF_{-Z_b} with full and selected SNPs.

		Full			Selected		
Water Supply	Phenotype	β_{01}	β_{10}	$\widehat{\gamma}$	β_{01}	β_{10}	$\widehat{\gamma}$
FI	DH	0.0344	0.4656	0.0277	0.0142	0.4858	0.0165
	GY	0.0587	0.4413	2.4231	0.0612	0.4388	2.6715
	NLS	0.0077	0.4923	0.1670	0.0116	0.4884	0.1629
	TKW	0.0264	0.4736	0.0877	0.0270	0.4730	0.0724
MWS	DH	0.0394	0.4606	0.0451	0.0333	0.4667	0.0597
	GY	0.0644	0.4356	4.5171	0.0688	0.4312	4.7537
	NLS	0.0596	0.4404	0.2502	0.0569	0.4431	0.2510
	TKW	0.0861	0.4139	0.4125	0.1109	0.3891	0.4787

8.5 Data2 Ranking

Table 8 reports both the prediction accuracies and Spearman's rank-order correlations. Same as before, these two measurements are highly correlated. It shows that with strong spatial effects, i.e., phenotype GY under full irrigated (FI) conditions and mild water stress, GRF_{-Z_b} is much better than $\text{GRF}_{-Z_{bs}}$ in terms of prediction accuracies and the similarities of rank-orders with the true rank-order $\boldsymbol{r}^{(o)}$. We can see in Figure 5 that GRF_{-Z_b} is consistently better than $\text{GRF}_{-Z_{bs}}$. This provides the evidence that accounting for spatial effects improves selection of the best plant genotypes.

Table 8: Average prediction accuracies and Spearman's rank-order correlations (ρ_s) based on 1000 simulations for different phenotypes by $\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b} with full and selected SNPs under full irrigated (FI) conditions and mild water stress.

		Full				Selected			
		Accuracies		ρ_s		Accuracies		ρ_s	
WS	Phenotype	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}	$\text{GRF}_{-Z_{bs}}$	GRF_{-Z_b}
FI	DH	0.9870	0.9895	0.9851	0.9879	0.9604	0.9606	0.9558	0.9562
	GY	0.7241	0.8401	0.7056	0.8261	0.7123	0.8334	0.6940	0.8196
	NLS	0.9148	0.9302	0.9056	0.9222	0.9067	0.9184	0.8972	0.9098
	TKW	0.9622	0.9697	0.9574	0.9658	0.9505	0.9558	0.9446	0.9504
MWS	DH	0.9749	0.9783	0.9714	0.9751	0.9467	0.9503	0.9405	0.9443
	GY	0.6257	0.8201	0.6060	0.8052	0.6251	0.8195	0.6077	0.8051
	NLS	0.9036	0.9184	0.8938	0.9099	0.8927	0.9064	0.8818	0.8966
	TKW	0.9336	0.9702	0.9258	0.9660	0.9210	0.9533	0.9122	0.9478

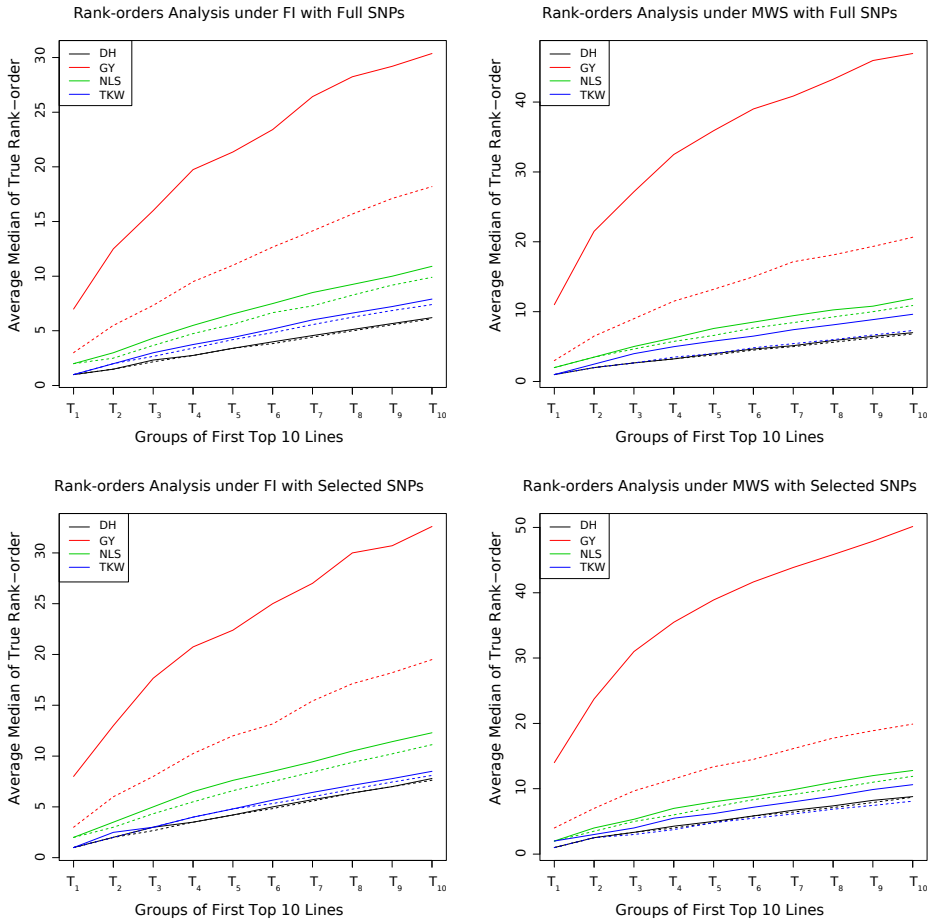

Figure 5: Comparisons of $\text{GRF}_{-Z_{bs}}$ and GRF_{-Z_b} with full and selected SNPs under full irrigated (FI) conditions and mild water stress. The solid lines are for $\text{GRF}_{-Z_{bs}}$, while dashed lines are for GRF_{-Z_b}.

References

Bernal-Vasquez, A.-M., Möhring, J., Schmidt, M., Schönleben, M., Schön, C.-C., and Piepho, H.-P. (2014), "The importance of phenotypic data analysis for genomic prediction-a case study comparing different spatial models in rye," *BMC Genomics*, 15(1), 646.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian com-

putation and stochastic systems," *Statistical science*, pp. 3–41.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(1), 25–37.
URL: *http://www.jstor.org/stable/2346064*

Besag, J., and Higdon, D. (1999), "Bayesian analysis of agricultural field experiments," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4), 691–746.

Besag, J., and Kooperberg, C. (1995), "On conditional and intrinsic autoregressions," *Biometrika*, 82(4), 733–746.

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., and Luis Araus, J. (2012), "High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding ConvergeF," *Journal of Integrative Plant Biology*, 54(5), 312–320.

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., and Krishnamachari, A. (2006), "Modeling genotype× environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes," *Crop Science*, 46(4), 1722–1733.

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

Cullis, B., Gogel, B., Verbyla, A., and Thompson, R. (1998), "Spatial analysis of multi-environment early generation variety trials," *Biometrics*, pp. 1–18.

Dutta, S., and Mondal, D. (2015), "An h-likelihood method for spatial mixed linear models based on intrinsic auto-regressions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(3), 699–726.

Dutta, S., and Mondal, D. (2016), "REML estimation with intrinsic Matérn dependence in the spatial linear mixed model," *Electronic Journal of Statistics*, 10(2), 2856–2893.

Endelman, J. B. (2011), "Ridge regression and other kernels for genomic selection with R package rrBLUP," *The Plant Genome*, 4(3), 250–255.

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*, pp. 269–293.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995), "Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models," *Biometrics*, pp. 1440–1450.

Gleeson, A. C., and Cullis, B. R. (1987), "Residual maximum likelihood (REML) estimation of a neighbour model for field experiments," *Biometrics*, pp. 277–287.

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.

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., and Zhang, Z. (2012), "GAPIT: genome association and prediction integrated tool," *Bioinformatics*, 28(18), 2397–2399.

Liu, X., Huang, M., Fan, B., Buckler, E. S., and Zhang, Z. (2016), "Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies," *PLoS Genet*, 12(2), e1005767.

Masuka, B., Araus, J. L., Das, B., Sonder, K., and Cairns, J. E. (2012), "Phenotyping for abiotic stress tolerance in maizef," *Journal of Integrative Plant Biology*, 54(4), 238–249.

McCullagh, P., and Clifford, D. (2006), Evidence for conformal invariance of crop yields,, in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, Vol. 462, The Royal Society, pp. 2119–2143.

McVean, G. (2009), "A genealogical interpretation of principal components analysis," *PLoS Genetics*, 5(10), e1000686.

Morota, G., and Gianola, D. (2014), "Kernel-based whole-genome prediction of complex traits: a review," *Frontiers in Genetics*, 5.

Morota, G., Koyama, M., Rosa, G. J., Weigel, K. A., and Gianola, D. (2013), "Predicting complex traits using a diffusion kernel on genetic markers with an application to dairy cattle and wheat data," *Genet Sel Evol*, 45, 17.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., and Simianer, H. (2011), "Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data," *Genetics*, 188(3), 695–708.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006), "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, 38(8), 904.

Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000), "Association mapping in structured populations," *The American Journal of Human Genetics*, 67(1), 170–181.

Reich, D., Price, A. L., and Patterson, N. (2008), "Principal component analysis of genetic data," *Nature Genetics*, 40(5), 491–492.

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

White, J. W., Andrade-Sanchez, P., Gore, M. A., Bronson, K. F., Coffelt, T. A., Conley, M. M., Feldmann, K. A., French, A. N., Heun, J. T., Hunsaker, D. J. et al. (2012), "Field-based phenomics for plant genetics research," *Field Crops Research*, 133, 101–112.

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008), "Genetic design and statistical power of nested association mapping in maize," *Genetics*, 178(1), 539–551.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B. et al. (2006), "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nature Genetics*, 38(2), 203–208.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. et al. (2010), "Mixed linear model approach adapted for genome-wide association studies," *Nature Genetics*, 42(4), 355–360.

Zimmerman, D. L., and Harville, D. A. (1991), "A random field approach to the analysis of field-plot experiments and other spatial experiments," *Biometrics*, pp. 223–239.