WHOLE FARM EXPERIMENTATION: MAKING IT PROFITABLE

By

DAVOOD POURSINA

Bachelor of Science in Statistics
Isfahan University of Technology
Isfahan, Iran
2005

Master of Science in Applied Statistics
University of Isfahan
Isfahan, Iran
2007

Doctor of Philosophy in Statistics
University of Isfahan
Isfahan, Iran
2014

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2022

WHOLE FARM EXPERIMENTATION: MAKING IT PROFITABLE

Dissertation Approved:

Dr. B. Wade Brorsen

Dissertation Adviser

Dr. Dayton M. Lambert

Dr. Rodney D. Jones

Dr. Daryl Brian Arnall

Acknowledgements

I thank God for giving me life and fantastic opportunities all along. I would like to thank my advisor Dr. B. Wade Brorsen, which without his help, it was impossible to finish this research. I express my full gratitude to him for "saving me from myself" most of the time. Dr. Brorsen introduced me to spatial Bayesian analysis, which I have completely come to love and enjoy doing. Thank you, Dr. Brorsen. I also acknowledge the contributions of my other dissertation committee members: Dr. Dayton Lambert, Dr. Rodney Jones, and Dr. Brian Arnall. Without the committee's criticisms and approvals, I would not be able to complete this dissertation. I also recognize the contributions of my classmates and friends, Ryan Loy, Dr. Whoi Cho, and all of my officemates. You guys are the best!

Additionally, I would like to express my everlasting thanks to my beloved and amazing wife, Niloofar, for her unfailing encouragement and love and for being my rock at all times. Furthermore, I dedicate this work to my family, including my wife and my lovely elder siblings. Lastly, I also affectionately dedicate this dissertation to the best mom in the world, Mahin, and the most hardworking dad of all time, Mohammadali!

Name: DAVOOD POURSINA

Date of Degree: DECEMBER, 2022

Title of Study: WHOLE FARM EXPERIMENTATION: MAKING IT PROFITABLE

Major Field: Agricultural Economics

Abstract: The first essay considers Bayesian Kriging (BK), which provides a way to estimate spatially varying coefficient regression models where the parameters are smoothed across space. The problem is that previous methods are too computationally intensive when estimating a nonlinear production function. The first essay sought to increase the computational speed by imposing restrictions on the spatial covariance matrix. Two correlation matrices that are sparse in the precision matrix: conditional autocorrelation (CAR) and simultaneous autocorrelation (SAR), were considered. In addition, a new analytical solution is provided for finding the optimal nitrogen value with a stochastic linear plateau model. A comparison among models in the accuracy and computational burden shows that the restrictions reduced the computational burden by 90% (CAR) or 89% (SAR) and led to models that better predicted the missing values.

The second essay starts to deal with the experimentation problem for on-farm experimentation when we know that spatial heterogeneity exists. Nearly Ds-Optimal allocation designs are obtained for an experiment that provides data from estimating the parameters of a linear SVC model in the second essay. This nearly optimal design is far more informative than standard designs such as Latin square (36%), simple random allocation (32%), and randomized strip-plot designs (69%).

The third essay aims to determine the optimal location of treatments when the yield response function is an SVC linear plateau model. The optimal locations are found when the researcher decides to experiment on a portion of the field in addition to when using the whole field. A pseudo-Bayesian approach is taken here because the field's site-specific optimal nitrogen value is unknown and local optimality is impossible. The resulting designs are more efficient than classic Latin square (29%), strip plot (63%), or completely randomized designs (59%) when the underlying yield response directly models field heterogeneity.

In the second and third essays, treatment levels and their corresponding replications are considered predetermined. In the fourth essay, we consider the farmers' net present value over eight years of experimentation and find the optimal levels of treatments, their corresponding replications, the number of experimenting plots, and the quit year for experimenting. Optimal on-farm experimentation is addressed using fully Bayesian decision theory. Of the designs considered, experimenting on 15 plots of a field with treatment levels of 35, 130, 165, and 230 with 2, 3, 5, and 5 replications maximized the farmers' profit over several years. The third year was the best time to quit experimenting.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

Site-Specific Nitrogen Recommendation: Fast, Accurate, and Feasible Bayesian Kriging

Abstract

Bayesian Kriging (BK) provides a way to estimate regression models where the parameters are smoothed across space. Such estimates could help guide site-specific fertilizer recommendations. One advantage of BK is that it can readily fill in the missing values that are common in yield monitor data. The problem is that previous methods are too computationally intensive to be commercially feasible when estimating a nonlinear production function. This paper sought to increase the computational speed by imposing restrictions on the spatial covariance matrix. Previous research used an exponential function for the spatial covariance matrix. The two alternatives considered are the conditional autoregressive (CAR) and simultaneous autoregressive (SAR) models. In addition, a new analytical solution is provided for finding the optimal value of nitrogen with a stochastic linear plateau model. A comparison among models in the accuracy and computational burden shows that the restrictions significantly reduced the computational burden and led to models that better predicted the missing values.

Key Words: Bayesian Kriging, fertilizer, Gaussian spatial process, linear plateau, optimal nitrogen, spatially varying coefficients

1. Introduction

Precision fertilizer application essentially requires finding a production function for each field piece. There are several interesting papers in this field of research (Anselin, et al., 2004; Evans, et al., 2020; Griffin, et al., 2008; Hurley, et al., 2005). Several methods exist to handle the data's spatial behavior. There is a tradeoff between the accuracy and complexity of a model and the ability to estimate it. The objective of the research reported here was to find a precise optimal input value for each part of a field. The specific objective was to find an efficient computational way to make BK feasible for data with a large number of locations. The primary hypothesis of interest was that respecifying the covariance matrix can lead to a model that can be solved more quickly and leads to more accurate forecasts. In addition, an analytical solution for obtaining the optimal nitrogen value for the random-parameter linear plateau model is provided.

There is a rich literature from Heady and Pesek (1954) and Spillman (1933) to the most recent articles (Archontoulis, et al., 2020; Mencaroni, et al., 2021) on finding the optimal value of fertilizer for a given response variable. Early spatial approaches sought to estimate a production functions with the parameters constant for clusters in the data set. A dummy variable was then added for each cluster (Lambert, et al., 2004; Liu, et al., 2006). The dummy variable approach has drawbacks. It requires prior knowledge about how to form the clusters or predefined clustering system. It assumes parameters vary discretely rather than smoothly across a field. Finally, this method could suffer from a lack of degrees of freedom or multicollinearity if the number of variables that should be considered dummies increases which could affect the inference significantly.

The second approach, which is widely used, is geographically weighted regression (GWR). GWR usually uses a neighboring system (with a different number of neighbors) to find

the model's spatial weights (Evans, et al., 2020; Lambert and Cho, 2022; Trevisan, et al., 2021). Although this model can fit the data well, it suffers from the lack of statistical theory for optimality behavior (Dambon, et al., 2021). Wheeler (2014) argues that "GWR is more appropriately viewed as an exploratory approach and not a formal model to infer parameter nonstationary."

The third approach is Bayesian Kriging (Park, et al., 2016). Several papers have been published to compare the GWR and spatially random coefficient models. Wheeler and Calder (2007) showed that the spatially random coefficient model provides more accurate parameter estimates than GWR through a simulation study. Wheeler and Waller (2009) used a public health data set and showed that spatially random coefficient models provide more robust regression coefficients in the moderate to high multicollinearity situation. Finley (2011) compared these two models with several criteria. He concluded that although the GWR was faster and useful in fitting the data, the spatially random coefficient model has a significantly smaller prediction mean square error. Besides, in the GWR, the weight is fixed (a grid search across weights can be done), while in Bayesian Kriging, the optimal weight is estimated simultaneously. The methods from first to last become more complicated and time-consuming. The Bayesian Kriging method has mostly been used with a dense continuous correlation matrix. The integrated nested Laplace approximation (INLA) can only handle models that are linear in the parameters (Rue, et al., 2017) and so it cannot be used with a linear plateau model. Park et al. (2018) used Bayesian Kriging and an exponential covariance matrix to find optimal nitrogen recommendations based on a linear plateau model. However, they only estimate the plateau spatially and restrict the number of locations to 160 to reduce computational time. The model is very time-consuming when the number of random coefficients or sites increases. Hence, finding a method that is not

only feasible for any data but also accurate is essential if Bayesian Kriging is to be competitive with GWR. In this situation, sparsity in the precision matrices (covariance inverse) could make the code faster. Firstly, the code does not need to compute the inverse of a large covariance matrix. Both the CAR and SAR models have sparse precision matrices.

We use Bayesian Kriging to estimate a linear plateau model with spatially varying coefficients and use it to estimate each location's optimal nitrogen level. Both the intercept and plateau parameters vary across the field and each has their own spatial correlation matrix. Calculations involving the spatial correlation matrix are a major reason for the slow computation. Models were estimated with different spatial correlation functions and then compared using computer time and the accuracy of the models to predict the missing values in the data.

Linear plateau models were estimated using the corn (*Zea mays L.*) nitrogen response data from Bongiovanni and Lowenberg-DeBoer (2000). While there is only one observation for each location, the estimated intercept and plateau differ for each location. The Hamiltonian Monte Carlo (HMC) algorithm, provided by Rstan was used to estimate the posterior density function. The optimal N value at each site was obtained by maximizing the expected profile using the posterior density.

1. Bayesian Linear Plateau model

The end goal is to find the optimum amount of nitrogen at each location. A common and effective data generating process for this purpose is a linear plateau model (Llewelyn and Featherstone (1997); Tembo, et al. (2008)). The innovation is assuming that the parameters in these models vary by location. The proposed model is

$$y_i = \min(a_i + bN_i, Plateau_i) + \epsilon_i \quad (1)$$

where y_i and N_i are the yield and the amount of nitrogen input in location i; $a_i$is the intercept and $Plateau_i$ is the plateau parameter. The effect of nitrogen (b) is fixed over space to reduce the computational burden and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. Let $\boldsymbol{a} = (a_1, a_2, \cdots, a_n)^T$and $\boldsymbol{Plateau} = (plateau_1, plateau_2, \cdots, plateau_n)^T$ are the $n \times 1$vector of parameters follow a Gaussian random process with spatial correlation matrices of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_p$. Hence,

$$\begin{aligned} \boldsymbol{a} &\sim MVGP(\alpha \mathbf{1}, \boldsymbol{\Sigma}_0) \\ b &\sim N(\beta, \sigma_b^2) \\ \boldsymbol{Plateau} &\sim MVGP(p\mathbf{1}, \boldsymbol{\Sigma}_p) \end{aligned} \quad (2)$$

where α, β, p are the mean parameters, $\mathbf{1}$ is an $n \times 1$vector with all elements equal to one, $\sigma_b^2$is the variance component for the slope. The parameters a_i and $Plateau_i$ are assumed to vary across locations, and parameters are spatially autocorrelated. Hence, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_p$are the $n \times n$ covariance matrices of the multivariate Gaussian process (MVGP) that depicts this behavior in the parameters. The covariance matrices in the MVGP can have varied structures.

2.1 Optimal nitrogen level recommendation

Assume that all other inputs are fixed, the optimal level of input nitrogen is selected to maximize expected profit:

$$\max_{N_i} E(\pi_i | N_i) = \max_{N_i} \int [Price(\min(a_i + bN_i, Plateau_i)) - rN_i] f(\boldsymbol{\Psi}) \boldsymbol{d\Psi} \quad (3)$$

where the $\boldsymbol{\Psi}$ contains all the parameters which should be estimated, and *f* is the posterior distribution function of parameters. Since the price and cost do not depend on the parameters, the integration is calculated only on the profit equation's production function.

Tembo et al. (2008) (2008) consider a plug-in method to find nitrogen's economic optimal value for a stochastic linear plateau. This method is not applicable in the current situation due to uncertainty in both parts of the linear plateau model.

Ouedraogo and Brorsen (2018) used a grid search and found the expectation using the Monte Carlo sample of the posterior distribution. This method could be used to find the optimal value for each location; however, the grid search method in large data sets would be time-consuming.

The posterior distribution of the parameters converges in limit to the multivariate normal distribution (Van der Vaart (2000)). The analytical solution to find the optimal value is based on the normality assumption of the posterior distribution.

In the problem at hand, one goal is to calculate

$$E(\min(a_i + bN, plateau_i))$$

Nadarajah and Kotz (Nadarajah and Kotz (2008)) provide the distribution and moment generating function of the minimum and maximum of two jointly normal random variables.

Let

$$(X_1, X_2)^T \sim N_2\left((\mu_1, \mu_2)^T, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}\right) \quad (4)$$

and $Y = \min(X_1, X_2)$ then

$$f_Y(y) = f_1(y) + f_2(y)$$

where

$$f_1(y) = \frac{1}{\sigma_1}\phi\left(\frac{y-\mu_1}{\sigma_1}\right)\Phi\left(\frac{\rho(y-\mu_1)}{\sigma_1\sqrt{(1-\rho^2)}} - \frac{(y-\mu_2)}{\sigma_2\sqrt{(1-\rho^2)}}\right)$$

$$f_2(y) = \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)\Phi\left(\frac{\rho(y-\mu_2)}{\sigma_2\sqrt{(1-\rho^2)}} - \frac{(y-\mu_1)}{\sigma_1\sqrt{(1-\rho^2)}}\right)$$

and the mean of Y is

$$E(Y) = \mu_1\Phi\left(\frac{\mu_2-\mu_1}{\theta}\right) + \mu_2\Phi\left(\frac{\mu_1-\mu_2}{\theta}\right) - \theta\mu_1\phi\left(\frac{\mu_2-\mu_1}{\theta}\right) \quad (5)$$

where ϕ, and Φ are the PDF and CDF of normal distribution respectively, and $\theta = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$.

Assuming that the $a_i, b,$ and $plateau_i$ are independent, and the posterior distribution of the parameters is

$$a_i \sim N\left(\bar{a}_i, \sigma_{a_i}^2\right)$$

$$b \sim N\left(\bar{b}, \sigma_b^2\right)$$

$$plateau_i \sim N\left(\bar{p}_i, \sigma_{p_i}^2\right)$$

then

$$a_i + bN \sim N\left(\bar{a}_i + bN, \sigma_{a_i}^2 + N^2\sigma_b^2\right)$$

$$plateau_i \sim N\left(\bar{p}_i, \sigma_{p_i}^2\right)$$

so, the expected value of $Y = \min(a_i + bN, plateau_i)$ is

$$E(Y) = (\bar{a} + \bar{b}N)\Phi\left(\frac{\bar{p} - \bar{a} - \bar{b}N}{\theta}\right) + \bar{p}\Phi(\frac{\bar{a} + \bar{b}N - \bar{p}}{\theta}) - \theta(\bar{a} + \bar{b}N)\phi(\frac{\bar{p} - \bar{a} - \bar{b}N}{\theta}) \quad (6)$$

$$\theta = \sqrt{\sigma_a^2 + N^2\sigma_b^2 + \sigma_p^2}$$

where the index *i* is dropped, for simplicity. The optimization seeks the *N* value to maximize equation (3). The first-order diffrentioation for this profit function in every location is

$$\frac{\partial E\pi}{\partial N} = \bar{b}price\left(1 - \Phi\left(\frac{\bar{a} + \bar{b}N - \bar{p}}{\theta}\right)\right) - r - \phi\left(\frac{\bar{a} + \bar{b}N - \bar{p}}{\theta}\right). \quad (7)$$

The root of equation (7) cannot be obtained analytically. Hence the "optimize" function in R (Team, 2013) was used to find the optimal value.

Some might argue that the intercept, slope, and plateau part cannot be independent in a real situation. By increasing the intercept, the slope will decrease. The plateau part in equation (1) may depend on the linear model, and the independence is questionable. Equation (6) can be adjusted for correlation. Suppose that the parameters are correlated and

$$\boldsymbol{V} = \begin{pmatrix} a \\ b \\ plateau \end{pmatrix} \sim MVN(\begin{pmatrix} \bar{a} \\ \bar{b} \\ \bar{p} \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix})$$

then, the vector $(a + bN, plateau)$ is equal to $A^T V$ where A is

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ N & 0 \\ 0 & 1 \end{pmatrix}$$

and $A^T V$ follows a multivariate normal distribution with mean and variance equal to

$$\boldsymbol{\mu} = (\bar{a} + \bar{b}N, \bar{p})$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} + N^2\sigma_{22} + 2N\sigma_{12} & \sigma_{13} + N\sigma_{23} \\ \sigma_{13} + N\sigma_{23} & \sigma_{33} \end{pmatrix}$$

Rewriting the covariance matrix as (4), gives

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11}^* & \rho\sigma_1^*\sigma_2^* \\ \rho\sigma_1^*\sigma_2^* & \sigma_{22}{}^* \end{pmatrix}$$

where $\rho = \frac{\rho_{13}\sigma_1\sigma_3 + \rho_{23}N\sigma_2\sigma_3}{\sigma_1\sigma_3 + N\sigma_2\sigma_3}$, $\sigma_{11}^* = (\sigma_1 + N\sigma_2)^2$, $\sigma_{22}^* = \sigma_{33}$. Hence the expectation of the linear plateau model is equal to (6) with

$$\theta = \sqrt{(\sigma_1 + N\sigma_2)^2 + \sigma_{33} - 2\frac{\rho_{13}\sigma_1\sigma_3 + \rho_{23}N\sigma_2\sigma_3}{\sigma_1\sigma_3 + N\sigma_2\sigma_3}(\sigma_1 + N\sigma_2)\sigma_3}. \qquad (8)$$

The first-order condition calculation is complicated and unnecessary because equation (6) is maximized with the new θ, given in equation (8), directly.

For the switching regression model of Paris (1992), $y_{it} = \min(\mathrm{a} + bN + \kappa_{it}, \mu_m + \omega_{it})$, the two random variables, κ_{it} and ω_{it} have marginal normal distributions. They do not necessarily have a joint bivariate normal distribution (the copula for the joint distribution is unspecified). So to use this approach in the Paris stochastic linear plateau would require an additional assumption of joint normality that is not imposed in the estimation.

2.2 Spatial correlation matrices and their behavior

The linear plateau model was considered as the data generating procedure. The coefficients vary across space and locations that are closer together have parameters that are more alike. This

behavior can be explained with a spatial covariance function and the normal distribution. Two well-known autoregressive precision matrices (covariance inverse) are the SAR and CAR. The term 'conditional' in the CAR structure shows conditional independence in the distribution of each element dependent on neighbors' values; however, the simultaneous form mostly emphasizes regressing the random part on themselves simultaneously (Hooten, et al. (2014)). Conditional independence between the element i and j can be easily seen in the precision matrix ($q_{ij} = 0$).

The CAR model is usually presented as a conditional distribution

$$\beta_i|\boldsymbol{\beta}_{-i} \sim N(\sum_{j=1}^{n} c_{i,j}\beta_j, m_{i,i})$$

where $\boldsymbol{\beta}_{-i}$ is the vector of all elements of vector $\boldsymbol{\beta}$ except β_i, $c_{i,j}$ are the *i*th and *j*th element of spatial weight matrix $\boldsymbol{C}$ and $\boldsymbol{M}$ is a diagonal matrix with positive diagonal elements of $m_{i,i}$. Following Besag (Besag (1974)), if $(\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}$ is a positive definite matrix then the CAR model can be written as

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{CAR}) \tag{9}$$

where $\boldsymbol{\Sigma}_{CAR}$, the covariance matrix is (Ver Hoef, et al. (2018))

$$\boldsymbol{\Sigma}_{CAR} = (\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{M}.$$

In practice, usually $\boldsymbol{\Sigma}_{CAR} = \frac{1}{\tau^2}(diag(\boldsymbol{W}\mathbf{1}) - \rho_c \boldsymbol{W})^{-1}$ is used where $\boldsymbol{W}$ is contiguity matrix**,** $\mathbf{1}$ is a vector of ones, and ρ_c is the amount of dependency between neighbors (Ver Hoef, et al. (2018)).

In the SAR model, an $n \times n$ weight matrix, B, relates the vector of parameters to themselves. In contrast to the CAR model, the SAR model can directly define the complete distribution of vector. Define

$$\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\vartheta}$$

where the matrix $\boldsymbol{B}$ is a $n \times n$ spatial weight matrix and $\boldsymbol{\vartheta} \sim MVN(\mathbf{0}, \boldsymbol{\Omega})$, so then

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{SAR}). \tag{10}$$

In the SAR model, the $\boldsymbol{\Sigma}_{SAR} = (\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\Omega}(\boldsymbol{I} - \boldsymbol{B}^t)^{-1}$, where $\boldsymbol{B}$ is not necessarily a symmetric matrix since $\boldsymbol{\Sigma}_{SAR}$ is symmetric even if $\boldsymbol{B}$ is not symmetric. In the SAR model, it is enough for $(\boldsymbol{I} - \boldsymbol{B})$ to be a non-singular matrix, $\boldsymbol{\Omega}$ be a diagonal matrix with positive values and $b_{ii} = 0$. In practice, usually consider $\boldsymbol{B}$ as a row standardized non-symmetric contiguity matrix. So the covariance matrix is $\boldsymbol{\Sigma}_{SAR} = (\tau(\boldsymbol{I} - \rho\, \boldsymbol{W}^{*\prime})(\boldsymbol{I} - \rho\, \boldsymbol{W}^{*}))^{-1}$ where $\boldsymbol{W}^{*}$ is the row standardized contiguity matrix.

Another common framework in geostatistics modeling is considering the correlation matrix as an elementwise decreasing function of distance among locations. Suppose that

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}_{ij} = cov(\beta_i, \beta_j) = \sigma^2 e^{-\frac{d_{ij}}{\rho}}$ is a positive definite covariance matrix, d_{ij} is the Euclidean distance between location i and j, ρ is the effective range and σ^2 is the sill. The exponential covariance matrix implies that the observations near each other are highly correlated while the far observations are nearly independent. Although this model uses the correlation matrix directly

and is straightforward to interpret, the researcher needs to specify a point to represent each unit, and for an extensive data set, fitting this model could be time-consuming.

Although the precision matrices in the CAR and SAR models are sparse, which leads to faster computing, the related covariance matrix is dense. So although a structure was assumed on the covariance matrices, no extra independence was considered between the locations. Besides, in both CAR and SAR covariance matrices, the correlation between them decreases by increasing the distance between two places.

2.3 Model fitting and layer specification in the Bayesian framework

The spatial behavior parameters need to be estimated, and there is uncertainty about their actual value. A hierarchical Bayesian perspective implies that some uncertainty may exist in the mean and correlation structures of the prior in the data generating process. The proposed model in the previous section contains three layers: likelihood, process (priors), and hyper prior level. The response variable was assumed to follow a linear plateau model in the likelihood layer, a non-linear model. Also, the parameters in this model were assumed to follow a multivariate Gaussian process. The dependency between the parameters in this model handles the Gaussian process's correlation structures. The third layer contains the hyperparameters priors which assure that the covariance matrix is positive definite. Based on the Bayesian framework, the posterior distribution of the parameters as

$$f(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3 | \boldsymbol{Y}) \propto f(\boldsymbol{Y} | \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \times f(\boldsymbol{\Theta_2} | \boldsymbol{\Theta_3}) \times f(\boldsymbol{\Theta}_3)$$

where $f(\boldsymbol{Y} | \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$, $f(\boldsymbol{\Theta_2} | \boldsymbol{\Theta_3})$, and $f(\boldsymbol{\Theta}_3)$ are the likelihood layer, process layer, and hyper prior layer, respectively. $\boldsymbol{\Theta}_1 = (a_1, a_2, \cdots, a_n, b, plateau_1, plateau_2, \cdots, plateau_n, \sigma_\epsilon)$ is the set

of parameters for the likelihood layer, $\boldsymbol{\Theta}_2 = (\alpha, \beta, p, \rho_0, \rho_p, \tau_0, \tau_p)$, and $\boldsymbol{\Theta}_3$ is the set of all hyperparameters in the distributions of the $\boldsymbol{\Theta}_2$. The likelihood layer is

$$f(\boldsymbol{Y}|\,\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) = \left(\frac{1}{\sqrt{2\pi\sigma_\epsilon^2}}\right)^n \exp \frac{(\boldsymbol{y}-\boldsymbol{\mu})'(\boldsymbol{y}-\boldsymbol{\mu})}{2\sigma_\epsilon^2}$$

where $\boldsymbol{y}$ is the vector of yield data, $\boldsymbol{\mu}$ is the vector with the elements equal to $\mathrm{E}(\min(a_i + bN_i, Plateau_i) + \epsilon_i)$, n is the number of observations, and σ_ϵ^2 is the variance component of ϵ.

The process layer deals with the model's spatial structure and finds a specific estimate for each location. The correlation matrix plays a vital role in the spatial structure of the data. Different parameters have been defined in the three mentioned methods, which should be determined in this layer. In the CAR and SAR model, the parameters are τ, and ρ, and in the exponential model, the parameters are σ, and ρ.

The stochastic spatial process in this model has distribution

$$f(\boldsymbol{\Theta}_2|\boldsymbol{\Theta}_3) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_p|}} \exp\left[-\frac{1}{2}(\boldsymbol{plateau} - \bar{\boldsymbol{P}})'\boldsymbol{\Sigma}_p^{-1}(\boldsymbol{plateau} - \bar{\boldsymbol{P}})\right] \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}_0|}} \exp\left[-\frac{1}{2}(\boldsymbol{a} - \bar{\boldsymbol{a}})'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{a} - \bar{\boldsymbol{a}})\right] \frac{1}{\sqrt{(2\pi)^n\sigma_b^2}} \exp\left[-\frac{1}{2\sigma_b^2}(b - \bar{b})^2\right]$$

where $\bar{\boldsymbol{P}}$, and $\bar{\boldsymbol{a}}$, are $\bar{p}\mathbf{1}$, $\bar{a}\mathbf{1}$; respectively. The covariance matrix in this layer could be any of the covariances defined in the previous section.

The hyper prior layer contains the priors for all the parameters in the process layer and some from the likelihood layer. The priors for p, α, and β are normal, with large variances to be

non-informative. The priors for the variance components are inverse gamma with parameters $\alpha = 2\,, \beta = 1$ for the τ. The covariance matrix in the normal distribution must be positive definite. Since the value of ρ could affect the positive definity of the covariance matrix, a careful selection of prior seems necessary. In practice, the restriction $\frac{1}{\lambda_1} < \rho < \frac{1}{\lambda_n}$ should be imposed with the prior where the λ_i is the eigenvalue of the *W*; however, the restriction turns to $\frac{1}{\lambda_1} < \rho < 1$ when the row standardized form of *W* is used (Haining, 1993). Spatial autocorrelation was restricted to be positive, $0 < \rho < 1$, by using a standard uniform prior. The improper prior proportional to the inverse of standard error for the variance component of σ_ϵ was considered in all three models. The prior for the sill and range parameter in the exponential correlation function consider being an improper distribution of $f(\rho, \sigma) \propto \frac{1}{\sigma}$. Fuglstad, et al. (2015) showed that this improper prior has stable results and can be used widely.

2. Data Analysis

The data used were the corn yield response to nitrogen from Bongiovanni and Lowenberg-DeBoer (2000) and Lambert and Cho (2022). The data were collected from a strip plot design in "Las Rosas" farm in Cordoba's southwestern corner of Argentina. Six different levels of nitrogen, namely 0,19,53,66,106, and 131.5 kg ha^{-1}, were applied to the farm based on a strip plot design. The highest nitrogen rate was higher than the value of nitrogen that was expected to maximize the response. The yield data and the selected nitrogen levels are given in Fig.1 and Fig.2, respectively. The original data contain 1738 locations that were digitalized as polygons. The centroid point was generated and considered as a data point in each area. 486 plots of the fields were selected from the data set such that all six levels of nitrogen were chosen, and the data were unbalanced for each level.

To estimate the linear plateau model in the Bayesian framework, the HMC algorithm was employed through the Rstan package in R. HMC algorithm in Stan uses a dynamic Hamiltonian Markov Chain to reduce the time of calculation and increase the chance of convergence. Different iteration and warmup values were employed for models to meet the convergence criteria for each model. The number of iterations and warmup for each model are given in Table1.1 ($number\ of\ iteration \times number\ of\ chains$). Different convergence criteria such as Gelman-Rubin statistics (Rhat) showed the ratio for the variance of parameters when the chain's data were pooled, and the number of effective samples were considered. The Trace and Trunk plots, which show the Markov property of the data and mixing chain property of chains, respectively, were also monitored to ensure convergence of the estimates. Models with SAR and an exponential correlation matrix converge with less iterations. However, these criteria in the CAR correlation matrix model were not met entirely like the two other models. The amount of time needed to run this model is given in Table 1.1. The site-specific estimates of y_i are given in figure 1.3.

Table 1.1 shows the estimated values for the parameters and their related Gelman-Rubin statistics (Rhat) for all three models. The estimated correlation parameters emphasize the spatial behavior's existence in the model's parameters that should be considered in the data analysis. The posterior likelihood value for the models can be compared since they have the same number of parameters. The time for getting an effective sample indicates that although the CAR model was far faster than the SAR model for creating an iteration, it needs more iterations and hence more time to converge. Also, the Rhat statistics show that the CAR model does not converge as well as the SAR model even with more iterations. The exponential correlation matrix model is not

feasible for a more extensive data set due to the computational burden. The SAR model was faster than the two other models in simulating an effective sample.

The optimal nitrogen value was calculated based on the posterior distribution of the parameters for every specific part of the field. The posterior distributions of each a_i, b and $plateau_i$ were estimated, then the results plugged into equation (6), and the optimal value for nitrogen was obtained. The results are given in figure 1.4. There are some substantial differences in nitrogen recommendations. Such differences are typical of models like this due to the limited number of observations. To provide more robust estimates, future research should explore imposing restrictions such as using informative priors, not allowing the intercept to vary across space, as well as estimating with additional years of data.

An out-of-sample test was used to calculate the accuracy of the models. One hundred randomly selected locations were removed from the data set. Then the posterior distribution was used to predict the missing values based on the nitrogen value and the location of the site. The Diebold and Mariano (2002) test was used to test the null hypothesis of no difference in prediction accuracy. Let y_i be the actual yield and let $\hat{y}_{1i}$ and $\hat{y}_{2i}$ be two forecasts based on two different methods. The Diebold-Mariano test statistic is

$$\sqrt{n}(\bar{d} - \mu) \xrightarrow{d} N(0, 2\pi f_d(0))$$

where $\bar{d}$ is the average of loss (square error) differential between two forecasts, $\mu = E(d)$, $f_d(0) = \frac{1}{2\pi}(\sum_{k=-\infty}^{\infty} \gamma_d(k))$, and $\gamma_d(k)$ is the autocovariance of the loss differential at lag k. Table 1.2 shows the results of the Diebold-Mariano accuracy test for forecasting the missing values in the data set. Package multDM in R was used for Diebold-Mariano test (Drachal, 2018). Results in Table 1.2 show no significant difference between the CAR and SAR models, and between the SAR and exponential covariance matrices in forecasting the missing values.

However the CAR has significantly less error than the exponential covariance matrix. The values for the mean squared errors were 84.64, 91.70, and 95.02 for the CAR, SAR and exponential covariance matrices, respectively.

3. Conclusions

In this paper, intercept and plateau parameters in the plateau model can be determined specifically for each location. Three different correlation matrices were considered.

In this application, all three models perform well in fitting the data set. In the CAR and SAR model, the neighbors' covariance is considered equal without attention to the distance between these two points. In the example at hand, the neighbors have a similar distance. However, in general, the distance between every neighbor could be far different. Table 1.1 shows that the CAR and SAR models were far faster than the exponential covariance model. They can be more easily used for large datasets due to the precision matrix's sparsity. If the number of locations is large and the data have some well-defined equally distant regions, the CAR and SAR models are feasible. Simultaneously, the exponential correlation function cannot be used in large data sets due to the computational burden of calculating the inverse and determinant of an extensive dense matrix. The exponential model was also less accurate in out-of-sample forecasting. Both the CAR and SAR offer speedier computations than the exponential model with no loss in out-of-sample accuracy.

4. References

Anselin, L., R. Bongiovanni, and J. Lowenberg-DeBoer. (2004). "A Spatial Econometric Approach to the Economics of Site-Specific Nitrogen Management in Corn Production." *American Journal of Agricultural Economics* 86:675-687.

Archontoulis, S.V., M.J. Castellano, M.A. Licht, V. Nichols, M. Baum, I. Huber, R. Martinez-Feria, L. Puntel, R.A. Ordóñez, and J. Iqbal. (2020). "Predicting Crop Yields and Soil-Plant Nitrogen Dynamics in the US Corn Belt." *Crop Science* 60:721-738.

Besag, J. (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems." *Journal of the Royal Statistical Society: Series B (Methodological)* 36:192-225.

Bongiovanni, R., and J. Lowenberg-DeBoer (2000) "Nitrogen Management in Corn Using Site-Specific Crop Response Estimates from a Spatial Regression Model." In *Proceedings of the Fifth International Conference on Precision Agriculture.*

Dambon, J.A., F. Sigrist, and R. Furrer. (2021). "Maximum Likelihood Estimation of Spatially Varying Coefficient Models for Large Data with an Application to Real Estate Price Prediction." *Spatial Statistics* 41:100470.

Drachal, K., (2018), "multMDM: Multivariate Version of the Diebold-Mariano Test." https://CRAN.R-project.org/package=multDM.

Evans, F.H., A. Recalde Salas, S. Rakshit, C.A. Scanlan, and S.E. Cook. (2020). "Assessment of the Use of Geographically Weighted Regression for Analysis of Large On-Farm Experiments and Implications for Practical Application." *Agronomy* 10:1720.

Finley, A.O. (2011). "Comparing Spatially-Varying Coefficients Models for Analysis of Ecological Data with Non-Stationary and Anisotropic Residual Dependence." *Methods in Ecology and Evolution* 2:143-154.

Fuglstad, G.-A., D. Simpson, F. Lindgren, and H. Rue. (2015). "Interpretable Priors for Hyperparameters for Gaussian Random Fields."arXiv preprint arXiv:1503.00256.

Griffin, T.W., C.L. Dobbins, T.J. Vyn, R.J. Florax, and J.M. Lowenberg-DeBoer. (2008). "Spatial Analysis of Yield Monitor Data: Case Studies of On-Farm Trials and Farm Management Decision Making." *Precision Agriculture* 9:269-283.

Haining, R. (1993). *Spatial Data Analysis in the Social and Environmental Sciences*: Cambridge University Press.

Heady, E.O., and J. Pesek. (1954). "A Fertilizer Production Surface with Specification of Economic Optima for Corn Grown on Calcareous Ida Silt Loam." *Journal of Farm Economics* 36:466-482.

Hooten, M.B., J.M. Ver Hoef, and E.M. Hanks. (2014). "Simultaneous Autoregressive (SAR) Model." *Wiley StatsRef: Statistics Reference Online*:1-10.

Hurley, T.M., K. Oishi, and G.L. Malzer. (2005). "Estimating the Potential Value of Variable Rate Nitrogen Applications: A Comparison of Spatial Econometric and Geostatistical Models." *Journal of Agricultural and Resource Economics*:231-249.

Lambert, D.M., and W. Cho. (2022). "Geographically Weighted Regression Estimation of the Linear Response and Plateau Function." *Precision Agriculture*: 23(2), 377-399.

Lambert, D.M., J. Lowenberg-Deboer, and R. Bongiovanni. (2004). "A Comparison of Four Spatial Regression Models for Yield Monitor Data: A Case Study from Argentina." *Precision Agriculture* 5:579-600.

Liu, Y., S.M. Swinton, and N.R. Miller. (2006). "Is Site-Specific Yield Response Consistent Over Time? Does it Pay?" *American Journal of Agricultural Economics* 88:471-483.

Llewelyn, R.V., and A.M. Featherstone. (1997). "A Comparison of Crop Production Functions Using Simulated Data for Irrigated Corn in Western Kansas." *Agricultural Systems* 54:521-538.

Mencaroni, M., N. Dal Ferro, J. Furlanetto, M. Longo, B. Lazzaro, L. Sartori, B. Grant, W. Smith, and F. Morari. (2021). "Identifying N Fertilizer Management Strategies to Reduce Ammonia Volatilization: Towards a Site-Specific Approach." *Journal of Environmental Management* 277:111445.

Nadarajah, S., and S. Kotz. (2008). "Exact Distribution of the Max/Min of Two Gaussian Random Variables." *IEEE Transactions on very large scale integration (VLSI) systems* 16:210-212.

Ouedraogo, F., and B.W. Brorsen. (2018). "Hierarchical Bayesian Estimation of a Stochastic Plateau Response Function: Determining Optimal Levels of Nitrogen Fertilization." *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie* 66:87-102.

Park, E., B. Brorsen, and A. Harri (2016) "Using Bayesian Spatial Smoothing and Extreme Value Theory to Develop Area-Yield Crop Insurance Rating. Selected paper." In *2016 Annual Meeting, Boston, MA, USA.*

Park, E., W. Brorsen, and X. Li. (2018). "How to Use Yield Monitor Data to Determine Nitrogen Recommendations: Bayesian Kriging for location specific parameter estimates." *Agricultural and Applied Economics Association*

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

Spillman, W.J. "Use of the Exponential Yield Curve in Fertilizer Experiments."United States Department of Agriculture Technical Bulletin 348.

Team, R.C., (2013), "R: A Language and Environment for Statistical Computing." https://www.r-project.org/.

Tembo, G., B.W. Brorsen, F.M. Epplin, and E. Tostão. (2008). "Crop Input Response Functions with Stochastic Plateaus." *American Journal of Agricultural Economics* 90:424-434.

Trevisan, R., D. Bullock, and N. Martin. (2021). "Spatial Variability of Crop Responses to Agronomic Inputs in On-Farm Precision Experimentation." *Precision Agriculture* 22:342-363.

Van der Vaart, A.W. (2000). *Asymptotic Statistics*: Cambridge University Press.

Ver Hoef, J.M., E.M. Hanks, and M.B. Hooten. (2018). "On the Relationship Between Conditional (CAR) and Simultaneous (SAR) Autoregressive Models." *Spatial Statistics* 25:68-85.

Ver Hoef, J.M., E.E. Peterson, M.B. Hooten, E.M. Hanks, and M.J. Fortin. (2018). "Spatial Autoregressive Models for Statistical Inference from Ecological Data." *Ecological Monographs* 88:36-59.

Wheeler, D.C. (2014) "Geographically Weighted Regression." In M.M. Fischer, and P. Nijkamp eds. *Handbook of regional science.* Springer Heidelberg New York Dordrecht London, Springer, pp. 1435-1459.

Wheeler, D.C., and C.A. Calder. (2007). "An Assessment of Coefficient Accuracy in Linear Regression Models with Spatially Varying Coefficients." *Journal of Geographical Systems* 9:145-166.

Wheeler, D.C., and L.A. Waller. (2009). "Comparing Spatially Varying Coefficient Models: A Case Study Examining Violent Crime Rates and Their Relationships to Alcohol Outlets and Illegal Drug Arrests." *Journal of Geographical Systems* 11:1-22.

Table 1.1. Parameters Estimation and the Rhat Value in Three Models for the Yield

Parameters	CAR (Rhat)	SAR (Rhat)	Exponential (Rhat)
$\bar{a}$	60.86 (2.01)	58.85 (1.01)	58.69(1.01)
ρ_{int}	0.97 (1.00)	0.97 (1.00)	0.04(1.00)
τ_{int}	0.35 (1.01)	0.25(1.00)	10.17(1.00)
$\bar{b}$	0.11(1.10)	0.13(1.00)	0.11(1.01)
$\bar{p}$	69.35(2.25)	68.17(1.01)	70.15 (1.01)
ρ_p	0.89 (1.91)	0.96 (1.00)	1.29 (1.00)
τ_p	0.35 (1.18)	0.28(1.00)	0.60 (1.01)
Max time for effective sample (parameter)	21481 ($\bar{p}$)	348.52 ($\bar{a}$)	108952 (lp)
Time(hours)[a]	30.96	25.58	268.55
Iteration	600000	435000	60000
Warmup	300000	120000	35000

[a] Desktop PC with Intel Core i5-9500 CPU @ 3.00 GHz and 32 GB DDR4

Table 1.2. Diebold-Mariano Test Results

Comparison	Statistics	P-value
CAR-SAR	1.437	0.150
SAR-EXP	-1.214	0.225
CAR-EXP	-1.985	0.048

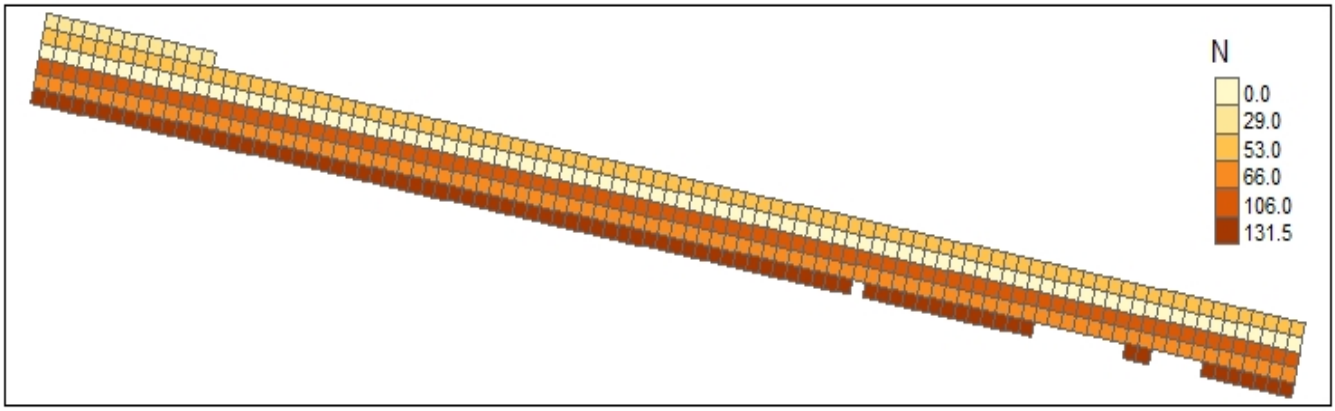

Figure 1.1. The Amount of Applied Nitrogen (kg ha^{-1})

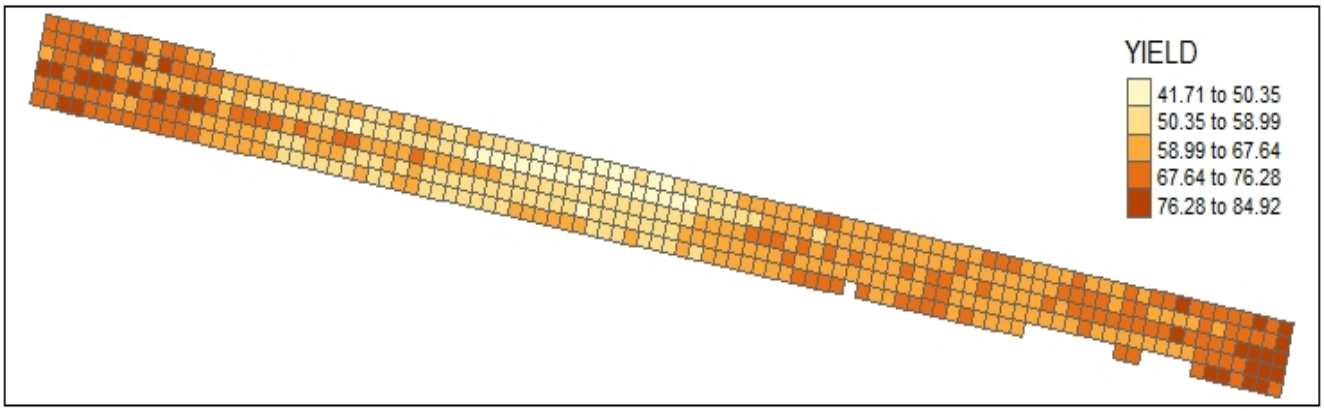

Figure 1.2. Value of Actual Yield (t ha^{-1})

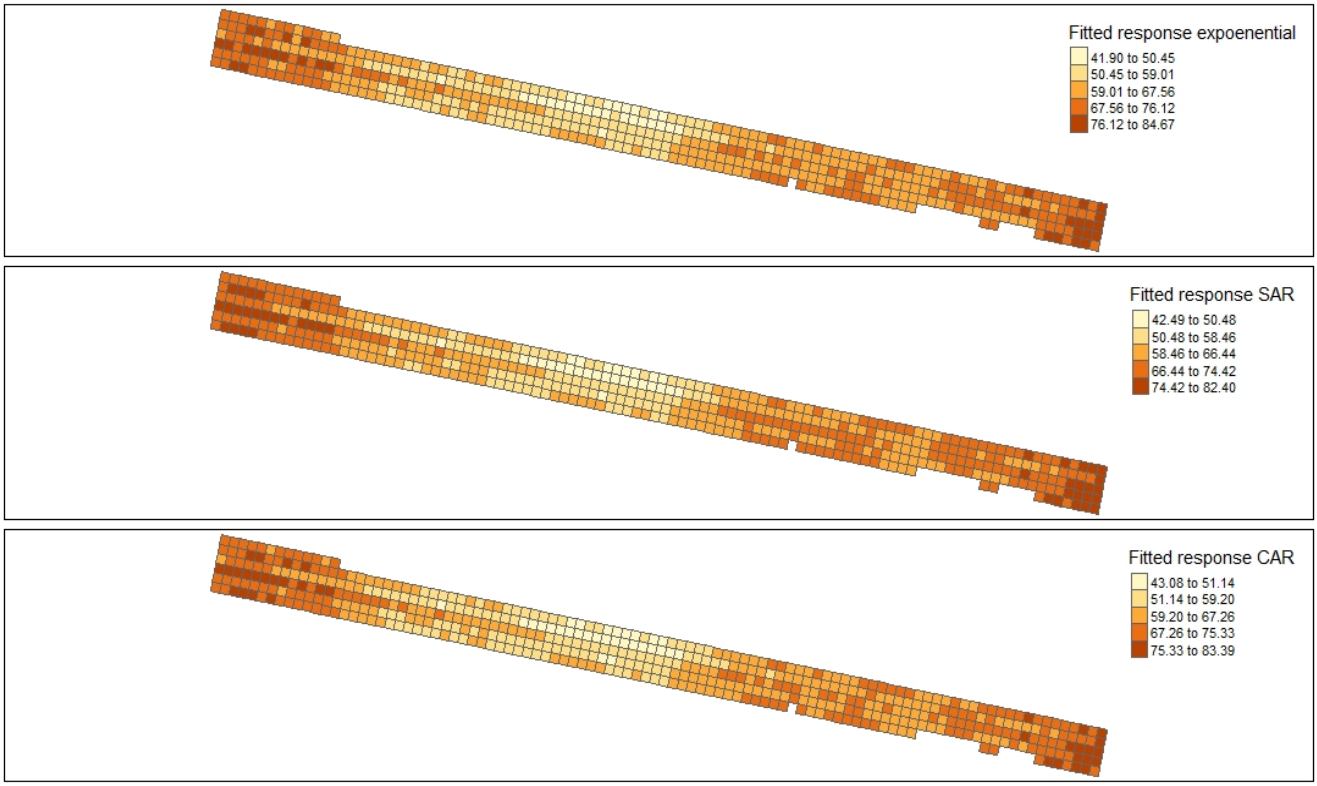

Figure 1.3. The Fitted Value (t ha^{-1})for Exponential, SAR, and CAR Model

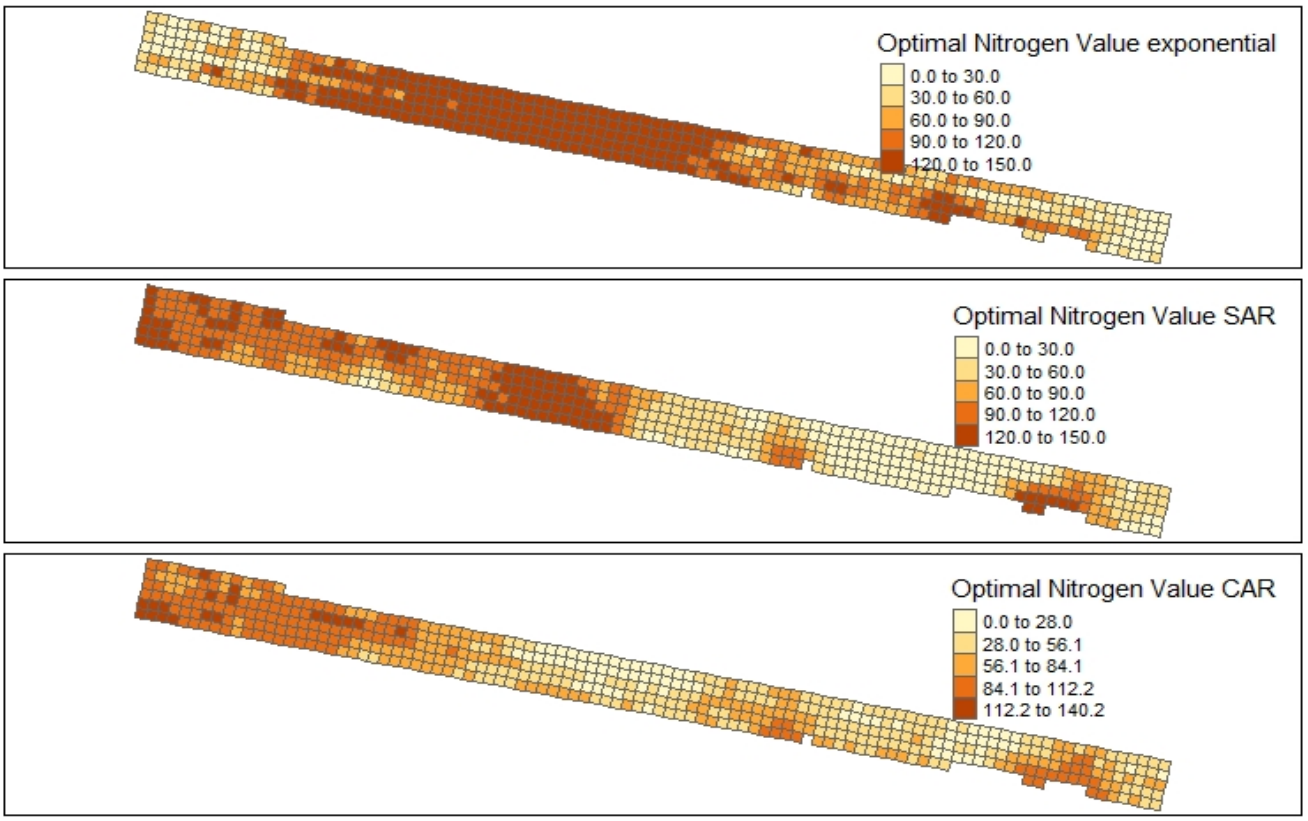

Figure 1.4. The Optimal Nitrogen Value (kg ha^{-1}) for Exponential, SAR, and CAR Model

CHAPTER II

Nearly Ds-Optimal Assigned Location Design for a Linear Model with Spatially Varying Coefficients

Abstract

Spatially varying coefficient (SVC) models are an increasingly popular approach for modeling spatial heterogeneity. One topic that has not been well addressed is how best to design experiments when the data are to be used to estimate a SVC model. The applied problem that motivated this study is that agronomists have begun conducting whole-field experiments on farmers' fields as an alternative to small-scale experiments on experiment stations. The goal is to guide the precision application of nutrients, such as nitrogen fertilizer. The research reported here seeks to optimally allocate treatments to the farm's plots by leveraging information from model designs. Nearly Ds-Optimal allocation designs are obtained for an experiment that provides data from estimating the parameters of a linear SVC model. This nearly optimal design is far more informative than standard designs such as Latin square, simple random allocation, and randomized strip-plot designs; all of which could also be used to generate data for SVC models. Furthermore, the suggested method does not need a regular plot shape for the experimental design, which is necessary for Latin square or strip plot designs.

Keywords: Locally D-optimal Design, On-Farm Experimentation, Spatially Varying Coefficients.

1. Introduction

Due to the increasing computational power of computers and new, faster algorithms, estimating spatially varying coefficients (SVC) models with georeferenced data have become more and more common (Gómez-Rubio, 2020; Rue and Held, 2005; Rue, et al., 2009; Semelhago, et al., 2020). These methods, however, have usually used observational data or nonoptimal experimental designs for collecting and analyzing spatially referenced data from on-farm experiments (Trevisan et al., 2020; Lambert and Cho, 2022). This paper develops a method to determine the optimal placement and location of treatments on plots with pre-trial knowledge on the functional form and treatment levels used to estimate crop response to inputs. The example uses an SVC model, but the approach is generalizable to models with spatially stationary coefficients. The work is motivated by a desire to increase, in terms of cost and precision, the efficiency of whole-field agronomic experiments that are used to guide the precision application of nutrients.

The experimental design literature originated from the implementation of agronomic field trials (Batchelor and Reed, 1918; Maat, 1850). The original purpose of an experimental design was to collect data to estimate treatment differences on crop yield. The introduction of production function estimation pioneered by Heady, et al. (1960) linked data generated from experimental designs to the analysis of economically optimal input use. Heady and Dillon's work was important in that it made a crisp distinction between biologically and economically optimal input management. Experimental designs for agronomic trials continue to play a prominent role in determining economically and biologically optimal input recommendations for a number of crops and a variety of fertilizer nutrient e.g. (Ali, 2020; Hatam, et al., 2020; Tamene, et al., 2017; Walsh, et al., 2018).

For decades, researchers relied on data from small-plot experiments to determine optimal nitrogen recommendations. Through advancements in technology, it is now feasible to conduct on-farm experiments where the experiment is conducted across the entire field. Early work analyzing the effects of spatial correlation on yield response to fertilizer used strip plot experimental designs (Anselin, et al., 2004). Today, the ability to link global positioning systems with fertilizer applicators makes it much easier to change quickly the amount of fertilizer applied, which allows increased flexibility in the design and implementation of agronomic experiments in terms of spatial granularity, the number of replications, and the placement of treatments.

On-farm experimentation has been suggested as a way to guide precision nitrogen applications, but has not yet produced unambiguously positive net returns because using non-optimal values of nitrogen in field experiments can reduce yield (Ng'ombe and Brorsen, 2022). Li, et al. (2021) concluded that randomly assigning treatment levels is not optimal, but judicious allocation of treatment locations could improve model estimation efficiency. Selecting plot locations that maximize the information gained from an experiment is a step toward making on-farm experimentation profitable.

Butler, et al. (2008); Eccleston and Chan (1998) considered linear models with a spatially correlated error term. They found that an A-optimal design was superior to other designs when model residuals are spatially correlated within the rows and columns of the design. These studies assumed that experimental rows and columns were independent, and that population effects were spatially stationary.

Mieno and Bullock (2017) and Bullock and Mieno (2017) suggested choosing treatment locations randomly. Ng'ombe and Brorsen (2022) also considered optimal experimental design

for on-farm trials. These studies ruled out a priori the effects of spatial dependence on the experimental design, so the location of plots did not matter.

Li, et al. (2021) did consider an SVC model in their examination of optimal plot locations. They compared the performance of a few specific designs based on a Monte Carlo simulation. They demonstrate the potential improvement in estimation efficiency from experimental designs, but do not attempt to find general optimal designs.

Alesso, et al. (2021) conducted a simulation study on the design of experiments for on-farm experimentation. Their research used a linear geographically weighted regression model to determine model-based experimental designs for three treatment randomization scenarios. Their measure of accuracy measure was the difference between the estimated and true values of the model parameters. They concluded that treatment randomization and the location of plots play an important role in increasing the accuracy of model estimates.

The optimal design of experiments usually involves selecting a set of design points, $x_1, x_2, \cdots, x_n$, and their corresponding weights, $w_1, w_2, \cdots, w_n$, to fulfill a specific goal, such as the efficient estimation of a model's parameters. For spatially explicit models, the locations of design points need to be incorporated into the design of the experiment. A common criterion for obtaining an optimal design is to maximize the determinant of the information matrix, which equivalently minimizes the volume of the vector of parameter estimates' confidence ellipsoid. In this paper, since the final goal is to estimate a site-specific value for an optimal fertilizer rate, the D-optimal criterion appears to be a completely reasonable design approach. Estimating SVC parameters from a D-optimal experimental design should increase the precision of estimates for the optimal application of nitrogen. Sometimes a researcher is more interested in a subset of the

model parameters; in this situation, the determinant of a submatrix of the information matrix is considered as the optimality criterion. This criterion is the Ds optimality criterion.

The information matrix does not usually depend on unknown parameters when determining the optimal design for a linear model. This means there is a closed-form solution for optimal designs for linear models. However, for linear models with SVC, the information matrix depends on the unknown parameters. This complication results in a chicken-and-egg situation. We want to find a design to estimate SVC parameters efficiently, but the design depends on unknown parameters. To handle this problem, the researcher can make an initial guess for true parameter values, which leads to locally optimal designs (Chernoff and Haitovsky, 1990; Poursina and Talebi, 2014). Locally optimal designs for general linear models are not robust against the misspecification of the true parameter values (Dette, et al., 2008; Dette, et al., 2008; Wiens, 2015). Alternative solutions include pseudo-Bayesian minimax approaches, which declare distributions for parameters rather than informed guesses (Chaloner and Verdinelli, 1995; Dette and Sahm, 1998; King and Wong, 2000; Pukelsheim, 2006).

This paper demonstrates that a 'nearly optimal' allocation of treatment locations significantly increases the information obtainable from a model-based design space. The experimental designs are evaluated in terms of their relative efficiency and precision. Stated differently, the parameter estimates using data generated from comparatively efficient designs have smaller standard errors. The robustness of locally optimal designs to poor guesses about true parameters values is also investigated. We find a locally nearly Ds optimal design for a 4×4, and 10×6 field with equal weights for four nitrogen levels, namely, 20, 50, 100, and 150 units per area. Latin square and strip plot designs are also feasible for 4×4 designs, but when plots are arranged in a 10×6 configuration, these two designs cannot be used. The suggested

method in this paper can handle any shape field. The relative efficiencies of standard randomized strip plot designs, sample random assignment, and Latin square benchmarks against which the proposed procedure is compared. The robustness of the local nearly-Ds-optimal designs is demonstrated since inaccurate prior guesses of parameters do not practically affect treatment locations assigned by the locally optimal design

2. Linear Model with Spatially Varying Coefficients

In some agricultural applications (Park, et al., 2019; Xu and Zhang, 2021), the regression coefficients may vary at locations or subregional levels, and the vector of parameters are correlated with each other. Under the assumption of normality, spatial dependency can be modeled as a Gaussian process that models covariance as a function that decreases as the distance between two locations increases (Cressie, 1993). SVC models are very flexible and capable of modeling the behavior of a response variable in a given region D with spatially correlated random coefficients.

Assume that the true model is

$$\mathbf{y}(\boldsymbol{s}) = \boldsymbol{\mu}(\boldsymbol{s}) + \mathbf{W}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}) \tag{1}$$

where $\boldsymbol{\mu}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta}$; $\boldsymbol{\epsilon}(\mathbf{s})$ is white noise with $\boldsymbol{\epsilon}(\mathbf{s}) \sim N(\mathbf{0}, \tau^2\mathbf{I})$; and $\mathbf{W}(\mathbf{s})$ is a second-order stationary process with mean zero and a valid positive definite variance-covariance matrix. Gelfand, et al. (2003) showed that a hierarchical spatial model aptly represents model (1). If a quadratic SVC yield response model is assumed to be the true yield response function, then (1) becomes:

$$\mathbf{y}(\mathbf{s}) = \beta_0 + \boldsymbol{\beta}_0(\mathbf{s}) + (\beta_1 + \boldsymbol{\beta}_1(\mathbf{s})) \cdot \boldsymbol{N}(\mathbf{s}) + (\beta_2 + \boldsymbol{\beta}_2(\mathbf{s})) \cdot \boldsymbol{N}^2(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \tag{2}$$

where ($\boldsymbol{\beta}_0(\mathbf{s})$, $\boldsymbol{\beta}_1(\mathbf{s})$, $\boldsymbol{\beta}_2(\mathbf{s})$) are random effects and the β_k's are population level effects.

Assume next that the distribution of the vector of spatial parameters $\boldsymbol{\beta}(\mathbf{s})$ follows the distribution:

$$f(\boldsymbol{\beta}_r(\mathbf{s})|\boldsymbol{\Psi}_r) \sim N(\mathbf{0}, \boldsymbol{\Psi}_r) \tag{3}$$

where $\boldsymbol{\Psi}_i$ is a covariance matrix that describes the spatial behavior of the ith parameter.

The parameters in (2) and (3) are not estimated here since it is unnecessary to estimate them to determine the optimal design locations. Maximum likelihood (ML) could be used to estimate the population level effects β_0, β_1, β_2, and $\boldsymbol{\Psi}_i$. Alternatively, Bayesian hierarchical methods could be used to estimate the spatial parameters $\boldsymbol{\beta}_i(\mathbf{s})$, which might lead to a different optimal design.

We can integrate over the $\boldsymbol{\beta}_i$'s and find the marginal likelihood of $\mathbf{y}$ to calculate the information matrix for the data. The marginalized likelihood over the $\boldsymbol{\beta}_i$'s under the independence assumption of spatial processes for $\boldsymbol{\beta}_0(s)$, $\boldsymbol{\beta}_1(s)$, and $\boldsymbol{\beta}_2(s)$ is

$$L(\beta_0, \beta_1, \beta_2, \tau^2, \boldsymbol{\Psi}_0, \boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2; \mathbf{y}) = |\boldsymbol{\Psi}_0 + \mathbf{D}_N \boldsymbol{\Psi}_1 \mathbf{D}_N + \mathbf{D}_{N^2} \boldsymbol{\Psi}_2 \mathbf{D}_{N^2} + \tau^2 \mathbf{I}|^{-\frac{1}{2}}$$
$$e^{-\frac{1}{2}(\boldsymbol{y}-\beta_0\mathbf{1}-\beta_1\boldsymbol{N}-\beta_2\boldsymbol{N}^2)^T(\boldsymbol{\Psi}_0+\mathbf{D}_N\boldsymbol{\Psi}_1\mathbf{D}_N+\mathbf{D}_{N^2}\boldsymbol{\Psi}_2\mathbf{D}_{N^2}+\tau^2\boldsymbol{I})^{-1}(\boldsymbol{y}-\beta_0\mathbf{1}-\beta_1\boldsymbol{N}-\beta_2\boldsymbol{N}^2)} \tag{4}$$

where $\mathbf{D}_N$ and $\mathbf{D}_{N^2}$ are the diagonal matrices with diagonal elements of $\boldsymbol{N}(\boldsymbol{s})$ and $\boldsymbol{N}^2(\boldsymbol{s})$, respectively. The $\mathbf{1}$, $\boldsymbol{N}$, and $\boldsymbol{N}^2$ are a vector of ones, nitrogen rates, and the square of nitrogen rates.

Since the marginal likelihood of the response variable can be written as (3) and the model is linear with a normal distribution, the Fisher information matrix for this SVC model is

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11} & 0 \\ 0 & \mathbf{M}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X} & 0 \\ 0 & \mathbf{M}_V \end{bmatrix} \tag{5}$$

where the $\mathbf{M}_{11}$ is a partitioned matrix related to the parameters in the mean, and $\mathbf{M}_{22}$ is the partitioned matrix related to the variance parameters. The matrix $\mathbf{X}$ is the design matrix, and $\mathbf{\Omega} = \mathbf{\Psi}_0 + \mathbf{D}_N\mathbf{\Psi}_1\mathbf{D}_N + \mathbf{D}_{N^2}\mathbf{\Psi}_2\mathbf{D}_{N^2} + \tau^2\mathbf{I}$ is the covariance matrix of the marginal distribution for the observation vector $\mathbf{y}$. The matrix $\mathbf{M}_V$ is the Fisher information matrix for the $\mathbf{\Psi}$ matrix of parameters and τ. The format of $\mathbf{M}_V$ depends on the spatial covariance function. The $\mathbf{M}_{11}$, and $\mathbf{M}_{22}$ can be investigated separately because the off-diagonal elements of the Fisher information matrix are zero. If we do not consider spatial heterogeneity for a subset of the parameters, the covariance matrix components of those parameters vanish.

3. Optimal Design Theory

An optimal design is a set of points and related weights that ensure the efficient estimation of a model; however, when spatial covariance is introduced into the experiment, treatment locations need to be codetermined with points and weights in the optimal design. Efficiency defined here is based on a utility (or loss) function, which is based on an experimenter's objective. Classical experimental design theory uses the Fisher information matrix as an optimality criterion. In other words, the objective function for a classical optimal design is $\arg\max_{\xi} \phi(\boldsymbol{\xi}, \boldsymbol{\theta}, \mathbf{Y})$, where the $\boldsymbol{\xi}$ are design points and their associated weights (the number of replications in each point) and locations, $\boldsymbol{\theta}$ are model parameters, and $\mathbf{Y}$ is the response vector. The function ϕ is usually determined by the experimenter's goal. For example, suppose the objective is to estimate the model's parameters with the highest degree of accuracy possible. In this case, ϕ is the determinant of the covariance matrix inverse for the estimated parameters. This optimality criterion is called D-optimal, which is a frequently used experimental design criterion.

Sometimes a researcher is interested in a subset of model parameters, with a primary objective of minimizing the volume of the confidence ellipsoid for a subset of the parameters. In

this case, the information matrix partitioned based on the vector of parameters and the variance-covariance of a subset of the parameters is considered. Suppose that the Fisher information matrix for a vector of parameters $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ could be partitioned as

$$\mathbf{M}(\boldsymbol{\xi}, \boldsymbol{\Theta}) = \begin{bmatrix} \mathbf{M}_{11}(\boldsymbol{\xi}, \boldsymbol{\Theta}) & \mathbf{M}_{12}(\boldsymbol{\xi}, \boldsymbol{\Theta}) \\ \mathbf{M}_{21}(\boldsymbol{\xi}, \boldsymbol{\Theta}) & \mathbf{M}_{22}(\boldsymbol{\xi}, \boldsymbol{\Theta}) \end{bmatrix}$$

where $\boldsymbol{\Theta}_2$ is the subset containing the parameters of interest. The resulting covariance matrix is:

$$[\mathbf{M}_{22}(\boldsymbol{\xi}, \boldsymbol{\Theta}) - \mathbf{M}_{21}(\boldsymbol{\xi}, \boldsymbol{\Theta})\mathbf{M}_{11}^{-1}(\boldsymbol{\xi}, \boldsymbol{\Theta})\mathbf{M}_{12}(\boldsymbol{\xi}, \boldsymbol{\Theta})]^{-1}$$

Since the off-diagonal elements of the Fisher information matrix given in (5) are $\mathbf{0}$, it suffices to maximize the determinant of $\mathbf{M}_{11}$ in equation (5) in order to maximize the information pertaining to the $\boldsymbol{\beta}_i$'s.

3.1 Nearly Ds Optimal Design for SVC Models

The Ds optimality criterion based on the information matrix in (5) depends on the covariance parameters that model the heterogeneity of spatial parameters. We obtain a locally optimal design based on an initial best guess for the true parameter values. Extending this design with the Bayesian framework is straightforward (Chaloner and Verdinelli, 1995; Pukelsheim, 2006) and requires maximizing the expectation, given a prior for the information matrix determinant.

Assume that the researcher's budget predetermines the range of treatment levels and the number of replications for each level. In this situation, the only choice variable for a locally optimal design is the location of each treatment level in the field. Three different correlation matrices are considered; namely an exponential decay function, a conditional autoregressive (CAR) process (Besag, 1994), and a Simultaneous Autoregressive (SAR) process (Anselin, 1988). The functional form for the exponential correlation matrix is

$$\text{cov}(\beta_r(s'), \beta_r(s'')) = \sigma_r^2 \exp\left(\frac{d_{s's''}}{\rho}\right), r = 0,1,2 \qquad (6)$$

where d_{ij} is the distance between location i and j, $\sigma_r^2$is the sill, and ρ_r is the effective range for the rth parameter.

The covariance matrix for the SAR process is[1]

$$\boldsymbol{\Sigma}_r^{SAR} = \sigma_r^2((\mathbf{I} - \rho_r\, \mathbf{W}^*)(\mathbf{I} - \rho_r\, \mathbf{W}^{*\prime}))^{-1}, r = 0,1,2 \qquad (7)$$

where σ_r^2 is the common variance for the $\boldsymbol{\beta}_r(\mathbf{s})$ parameters, $\mathbf{W}^*$ is a row standardized contiguity matrix, and ρ_r is the degree spatial dependence. As ρ_r increases and it is positive, then spatial dependence increases.

The covariance matrix for the CAR process is

$$\boldsymbol{\Sigma}_r^{CAR} = \sigma_r^2\big((\mathbf{D} - \rho_r\mathbf{B})\big)^{-1}, r = 0,1,2 \qquad (8)$$

where $\mathbf{D}$ is the summation of the row of contiguity matrix of neighbors and $\mathbf{B}$ is a contiguity matrix. Different assumptions about parameter spatial heterogeneity can be modeled by changing parameter values and contiguity assumptions in the correlation matrix. Therefore, the $\boldsymbol{\Omega}$ matrix for model (1) with SAR, CAR and exponential correlation functions can be obtained by substituting (6), (7) or (8) for $\boldsymbol{\Psi}$ in equation (3).

Suppose we have *K* levels of fertilizer and *N* different plots to which we want to assign a fertilizer amount. Assume that the number of replications for each level is predetermined and equal to $n_1, n_2, \cdots, n_K$ such that $\sum_{i=1}^{K} n_i = N$. Obtaining the best location for a fertilizer rate that maximizes information is a discrete optimization problem with $d = \frac{N!}{n_1!n_2!\cdots n_K!}$ possible permutations since changing the location of similar levels does not change the amount of

[1] It is worth noting that CAR and SAR were originally designed to model spatial dependence, and not spatial heterogeneity. CAR and SAR deal with spatial covariance structures, but originally assumed population effects to be spatially stationary. The SVC approach taken here hybridizes spatial heterogeneity models, such as GWR and previous SVC approaches Anselin, L. (1988). *Spatial Econometrics: Methods and Models*: Springer Science & Business Media, LeSage, J., and R.K. Pace. (2009). *Introduction to Spatial Econometrics*: Chapman and Hall/CRC. with spatial process models designed to explicitly model spatial covariance.

information. The number of all permutations could be a large number (nearly one Septendecillion for 100 locations) and calculating the information matrix for all permutations is impossible, especially for many plots in the field. To address this problem, a nearly optimal design can be determined based on the following algorithm and the distribution functions defining the information matrix's determinant.

Suppose that U is the set of all possible designs includes permutations of the matrix $\boldsymbol{\xi}_i$, which contains the nitrogen levels, the number of plots with the specific nitrogen level, and the location of these plots in a field with N plots, namely $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots$. Also, assume the matrix of random variables, $\mathbf{X}$, is the determinant of the information matrix related to the designs that follow the distribution F. Elements in this matrix of random variables are bounded by a and c. Without loss of generality, suppose that $a < \mathbf{X} < c$. It follows then that design $\boldsymbol{\xi}^*$ maximizes the determinant of the information matrix over all designs, and that this design has a finite determinant of information value $c < \infty$. Assume that distribution of $\mathbf{X}$ has a very thin tail where $P(\mathbf{X}_i < 0.95c) \leq 1 - \alpha$, where α is a very small positive number. In this situation, the sample size for reaching a maximum in the domain of distribution F can be obtained based on the following *lemma*.

> ***Lemma 1****:* Suppose that $X \sim F, a < x < c$ and F has a very thin tail. Let $x_1, x_2, \cdots, x_n$, be a random sample from F and $x_{(n)}$ is the n^{th} order statistics of this sample. If the number of samples is larger than $\frac{\ln(1-p)}{\ln(1-\alpha)}$, then the nearly optimal allocation will occur with probability p.

Proof:

Assume that $x_{(n)}$ provides the largest determinant of the Fisher information matrix among all of samples from U. That is:

$$p\left(X_{(n)} > 0.95c\right) = 1 - p(X_1 < 0.95c, X_2 < 0.95c, \dots, X_n < 0.95c)$$

$$= 1 - (1-\alpha)^n$$

If we assume that $p\left(X_{(n)} > 0.95c\right) \geq p$, then

$$1 - (1-\alpha)^n > p, \text{ and}$$

$$n > \frac{\ln(1-p)}{\ln(1-\alpha)} \tag{9}$$

Equation (9) can be used to find a minimum number of samples required to achieve a desired accuracy in estimating the largest order statistic of a distribution with a boundary. The number of samples for $\alpha = 10^{-6}$ and $p = 0.999$ is 6,907,752 samples, which can be processed easily and quickly.

4. Application and Results

For the first application of the suggested method, suppose we are looking for the best-assigned locations of a 4×4 plot. The number of permutations for these plots with four replications of each level of nitrogen is $\frac{16!}{4!4!4!4!} = 63{,}063{,}000$, which requires 8.1 gigabytes of RAM to store the matrix of all possible permutations. Calculating all possible permutations for these combinations was done with the RcppAlgos package in R (Wood, 2021). We consider the spatial behavior just on the intercept with the CAR and SAR covariance functions, assuming rook contiguity between plot locations. We assume that the true value of $\rho_0 = 0.8$, $\tau_0 = 20$, and $\sigma = 1$.

Figure (2.1) shows the Ds optimal design for this simple situation. The obtained Ds optimal design for an SVC intercept model under SAR and CAR covariance assumptions and

rook contiguity is a Latin square. The number of neighbors for each treatment level was also calculated. Results show that treatment levels for pairs (20, 50; 50,100; 100, 150) are neighbors two times, while the values for treatment with more distance between the levels (e.g., 20,100; 50, 150) are neighbors eight times. In addition, no treatment is a neighbor with itself, which is a property of Latin square designs. When the exponential covariance function is used, the resulting design is not a Latin square (Figure 2.2). This happens because there is no longer only row and column dependency.

For obtaining the local optimal design when all parameters are spatially varying, we consider guesses that are equal to the true values. We calculate the determinant of the Fisher information matrix for each permutation with 0.8, 0.9, and 0.8 as values for ρ_0, ρ_1, and ρ_2 in equation (7) for SAR and equation (8) for CAR, respectively. The value of τ and σ were also set as 20, 10, 30, and '1' for τ_0, τ_1, τ_2, and σ^2. Figure 2.3 shows the best allocation for this experiment, the given CAR parameters and rook contiguity. The number of neighbors for each treatment level are also calculated for the optimal design and the Latin square with maximum information. The number of neighbors shows that when we consider the CAR covariance matrix with rook contiguity, lower-valued treatment are neighbors, while higher-value treatment levels are not. The maximum efficiency for the queen contiguity scenario occurs when all treatment levels are neighbors with themselves exactly two times. This arrangement differs from the Latin square that is maximally efficient.

We repeated all steps mentioned above with the exponential correlation function with values of 4, 3, and 5 for ρ_0, ρ_1, and ρ_2, respectively (equation 6). The sill parameter values are set to 0.1, 0.5, 1, and 1 for σ_0^2, σ_1^2, σ_2^2, and σ^2, respectively. Figure 2.4 depicts the best-obtained allocation with the exponential correlation function and the given parameters. A nearly optimal

allocation design was also calculated for the same scenarios. Figures 2.5 and 2.6 illustrate the nearly optimal design with six million samples out of all possible permutations. The *RcppAlgos* package in R was used to randomly select samples from all possible permutations.

Suppose that the $\boldsymbol{\xi}^*$ is the optimal allocation on all possible permutations, and $\boldsymbol{\xi}$ is any other allocation of the nitrogen levels. The relative efficiency of design $\boldsymbol{\xi}$ with respect to the $\boldsymbol{\xi}^*$can be defined (Poursina and Talebi, 2014) as:

$$E = \frac{|M(\xi, \Theta)|}{|M(\xi^*, \Theta)|} \tag{10}$$

Extending equation (10) for comparing any two designs is straightforward. We calculate the average efficiency for the obtained nearly optimal designs for the CAR, SAR, and exponential correlation matrices and compare with the average and maximum efficiency for commonly used designs. For this purpose, we consider three different designs, namely a design that randomly assigned the locations to nitrogen levels, a strip plot design that assigns nitrogen levels in rectangular strips, and a Latin square design that assigns nitrogen levels such that each level of nitrogen is assigned only once in any row and column. The Magic package in R (Hankin, 2018) is used to select a random Latin square for each simulation. We then calculate the efficiency for each selected design and compare it to the optimal design that was calculated for all possible permutations, $\boldsymbol{\xi}^*$. The results are reported in Table 2.1. We considered 10,000 different random designs, calculated the efficiency of each design, and report the average efficiency score. For the random strip plot and random Latin square designs, the number of all possible combinations is 24 and 576, respectively, and the efficiency is calculated each. Since the nearly optimal design is not unique for the suggested method, we obtained this design 100 times to calculate an average efficiency. Table 2.1 shows the average and maximum efficiency

for each situation. Strip plot designs are one of the most common experimental designs for whole-farm experiments conducted in the past. However, the randomized strip plot design only had a 31 percent efficiency for estimating an SVC model. The Latin square and randomly assigned location designs have 50 to 60 percent efficiency on average for an SVC model. The maximum efficiency for strip plot design can be as low as 51 percent for the exponential. However, some Latin squares could be up to 96 percent efficient, and searching for them is relatively fast and easy.

Rectangular Configuration

The initial guesses for the true parameters value given by equations (6), (7), and (8) are considered fixed at the previous values of the precision and variance parameters, and the plot was extended to a 6×10 grid. Results for nearly optimal allocation with 30 million samples are given in Figures 2.7-2.9 for the CAR, SAR, and exponential correlation functions, respectively. There is no clear pattern in these optimal designs and the designs are not unique.

One important issue that should be addressed in this situation is the robustness of the optimal designs. Here, we assumed that the true value of the variance parameters are known. In practice, this assumption is unrealistic. Table 2.2 reports the efficiency of the optimal designs based on the counterfactual values of the parameters. Two scenarios are considered: small and large misspecification. For the small misspecification scenario, we assume that the guess for the parameters is 0.75, or $\frac{4}{3}$ of the true values. For the large misspecification scenario, we assume that the experimenter's guess is 0.1 and 10 times the true values. For both cases, we change the parameters in the variance matrix to allow more variance and less information. Hence, for the SAR and CAR covariance structures, we consider 0.75 and 0.1 as a multiplier of the true parameters as an initial guess. For the exponential covariance form, we assume $\frac{4}{3}$ and 10 as

multipliers. Table 2.2 reports the results. The optimal design for the SVC model is robust against initial parameter guesses. Designs are more sensitive with respect to the variance parameters in β_0. However, the minimum efficiency is nearly 95 percent for the obtained designs.

The flexible method for finding the nearly optimal allocation design can be used for any irregularly shaped field. For the usual Latin square design, the number of plots must be a multiple of the levels of treatments. For strip plot designs, the number of strips must be a multiple of the treatment levels. However, the method suggested in this paper has no limitations in terms of the number of rows and columns of a design or the shape of the field.

5. Conclusion

For years, small-plot agronomic experiments were the source of information for farm management. Since the data were small and fields were heterogeneous, the results varied form one field to the next. In addition, the parameter estimates exhibited significant variability. Now, a movement toward large-scale on-farm experiments has begun. We assume the researcher decides about the levels of nitrogen and their replications based on a project budget, so the total number of experiments and the nitrogen levels are fixed. The optimal design is determined by maximizing the determinant of the Fisher information matrix for a linear SVC model. The Fisher information matrix for linear SVC models depends on the unknown parameters of the spatial behavior in contrast with the usual linear models. We use the locally Ds-optimal criterion to find the best possible spatial allocation of plots. For a large field with a large number of plots, considering all possible permutations is impossible. Selecting a reasonable number of permutations can guarantee a nearly optimal design with good efficiency. With current technology, changing the locations of the treatment levels does not impose an extra cost on the project; however, it could significantly increase the amount of information compared to

traditional strip or Latin square designs. The suggested method in this paper also has no limitations on the number of plots and the shape of the field, which is a common drawback in traditional experimental methods.

Acknowledgments

The work benefited directly or indirectly from the work of Xiaofei Li, Taro Mieno, and David S. Bullock.

Funding

This work was supported by the A.J. and Susan Jacques Chair, the Sparks Chair in Agricultural Sciences & Natural Resources, the Oklahoma Agricultural Experiment Station, and USDA National Institute of Food and Agriculture [Hatch Project number OKL03170].

5. References

Alesso, C.A., P.A. Cipriotti, G.A. Bollero, and N.F. Martin. (2021). "Design of on-farm precision experiments to estimate site-specific crop responses." *Agronomy Journal* 113:1366-1380.

Ali, A.M. (2020). "Development of an algorithm for optimizing nitrogen fertilization in wheat using GreenSeeker proximal optical sensor. "Experimental Agriculture 56:688-698.

Anselin, L. (1988). Spatial econometrics: methods and models: Springer Science & Business Media.

Anselin, L., R. Bongiovanni, and J. Lowenberg-DeBoer. (2004). "A spatial econometric approach to the economics of site-specific nitrogen management in corn production." American Journal of Agricultural Economics 86:675-687.

Batchelor, L.D., and H.S. Reed. (1918). Relation of the variability of yields of fruit trees to the accuracy of field trials: US Government Printing Office.

Besag, J. (1994). "Discussion: Markov chains for exploring posterior distributions." The Annals of Statistics 22:1734-1741.

Bullock, D., and T. Mieno. (2017). "An assessment of the value of information from on-farm field trials." Unpublished Working Paper, University of Illinois, Champaign, IL.

Butler, D.G., J.A. Eccleston, and B.R. Cullis. (2008). "On an approximate optimality criterion for the design of field experiments under spatial dependence." Australian & New Zealand Journal of Statistics 50:295-307.

Chaloner, K., and I. Verdinelli. (1995). "Bayesian experimental design: A review." Statistical Science:273-304.

Chernoff, H., and Y. Haitovsky. (1990). "Locally optimal design for comparing two probabilities from binomial data subject to misclassification." Biometrika 77:797-805.

Cressie, N.A. (1993). Statistics for spatial data: John Wiley and Sons Inc., New York.

Dette, H., F. Bretz, A. Pepelyshev, and J. Pinheiro. (2008). "Optimal designs for dose-finding studies." Journal of the American Statistical Association 103:1225-1237.

Dette, H., C. Kiss, and W.K. Wong. "Robustness of optimal designs for the Michaelis-Menten model under a variation of criteria." Technical Report.

Dette, H., and M. Sahm. (1998). "Minimax optimal designs in nonlinear regression models." Statistica Sinica 8:1249-1264.

Eccleston, J., and B. Chan (1998) "Design algorithms for correlated data." In Compstat. Springer, pp. 41-52.

Gelfand, A.E., H.-J. Kim, C. Sirmans, and S. Banerjee. (2003). "Spatial modeling with spatially varying coefficient processes." Journal of the American Statistical Association 98:387-396.

Gómez-Rubio, V. (2020). Bayesian inference with INLA: CRC Press.

Hankin, R.K.S., (2018), "Package 'magic'." https://cran.r-project.org/web/packages/magic/index.html.

Hatam, Z., M.S. Sabet, M.J. Malakouti, A. Mokhtassi-Bidgoli, and M. Homaee. (2020). "Zinc and potassium fertilizer recommendation for cotton seedlings under salinity stress based on gas exchange and chlorophyll fluorescence responses." South African Journal of Botany 130:155-164.

Heady, E.O., C. CF, and L.D. John. (1960). "*Agricultural production functions.*" Iowa State University Press, Ames. USA.

King, J., and W.K. Wong. (2000). "Minimax D-optimal designs for the logistic model." Biometrics 56:1263-1267.
LeSage, J., and R.K. Pace. (2009). Introduction to spatial econometrics: Chapman and Hall/CRC.
Li, X., T. Mieno, and S.D. Bullock. (2021). "Economic performances of trial design methods in on-farm precision experimentation: A Monte Carlo evaluation"Unpublished".
Maat, H. (1850). "Statistics and field experiments in agriculture; The emerging discipline of inferential statistics." The statistical mind in modern society. The Netherlands 1940:91-112.
Mieno, T., and D. Bullock. (2017). "Getting to know your yield response better through whole-field randomized experiments." Cornhusker Economics.758.
Ng'ombe, J.N., and B.W. Brorsen. (2022). "Bayesian optimal dynamic sampling procedures for on-farm field experimentation." Precision Agriculture:1-23.
Park, E., B.W. Brorsen, and A. Harri. (2019). "Using Bayesian kriging for spatial smoothing in crop insurance rating." American Journal of Agricultural Economics 101:330-351
Poursina, D., and H. Talebi. (2014). "Modified D-optimal design for logistic model." Journal of Statistical Computation and Simulation 84:428-437.
Pukelsheim, F. (2006). Optimal design of experiments: Society for Industrial and Applied Mathematics.
Rue, H., and L. Held. (2005). Gaussian Markov random fields: theory and applications: CRC press.
Rue, H., S. Martino, and N. Chopin. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71:319-392.
Semelhago, M., B.L. Nelson, E. Song, and A. Wächter. (2020). "Rapid discrete optimization via simulation with Gaussian Markov random fields." INFORMS Journal on Computing.
Tamene, L., T. Amede, J. Kihara, D. Tibebe, and S. Schulz. (2017). "A review of soil fertility management and crop response to fertilizer application in Ethiopia: towards development of site-and context-specific fertilizer recommendation." International Center for Tropical Agriculture (CIAT), Addis Ababa No. 443.
Walsh, O.S., S. Shafian, and R.J. Christiaens. (2018). "Nitrogen fertilizer management in dryland wheat cropping systems." Plants 7:9.
Wiens, D.P. (2015). "Robustness of design." Handbook of design and analysis of experiments:719-753.
Wood, J., (2021), "RcppAlgos: High performance tools for combinatorics and computational mathematics." https://CRAN.R-project.org/package=RcppAlgos.
Xu, H., and C. Zhang. (2021). "Investigating spatially varying relationships between total organic carbon contents and pH values in European agricultural soil using geographically weighted regression." Science of The Total Environment 752:141977.

Tables

Table 2.1. Efficiency of Designs for 16 Locations All Coefficients Spatially Varying

Design of experiment	Number of designs	Average Efficiency for SAR queen	Average Efficiency for Exponential	Average Efficiency for CAR rook	Maximum efficiency for SAR queen	Maximum efficiency For Exponential	Maximum efficiency for CAR rook
Latin Square	576	51.2	61.8	81.9	93.1	91.2	83.2
Randomly Assigned	10000	60.2	65.01	80.5	94.1	96.0	98.9
Strip plot	24	30.2	28.9	68.54	0.56	0.51	82.2
Nearly Optimal	100	98.5	97.7	96.3	-	-	-

Table 2.2. Robustness of Nearly Optimal Designs Against Misspecification

Model	Parameter	Efficiency	Parameter	Efficiency
CAR	$\rho_0 = 0.6$	1	$\rho_0 = 0.08$	0.95
	$\rho_1 = 0.675$	1	$\rho_1 = 0.09$	1
	$\rho_2 = 0.6$	1	$\rho_2 = 0.08$	1
	$\tau_0 = 15$	1	$\tau_0 = 2$	0.97
	$\tau_1 = 7.5$	1	$\tau_1 = 1$	1
	$\tau_2 = 22.5$	1	$\tau_2 = 3$	1
	$\sigma_\epsilon = 1.33$	1	$\sigma_\epsilon = 10$	1
SAR	$\rho_0 = 0.6$	1	$\rho_0 = 0.08$	0.96
	$\rho_1 = 0.675$	1	$\rho_1 = 0.09$	1
	$\rho_2 = 0.6$	1	$\rho_2 = 0.08$	1
	$\tau_0 = 15$	1	$\tau_0 = 2$	0.95
	$\tau_1 = 7.5$	1	$\tau_1 = 1$	1
	$\tau_2 = 22.5$	1	$\tau_2 = 3$	1
	$\sigma_\epsilon = 1.33$	1	$\sigma_\epsilon = 10$	1
	$\sigma_\epsilon = 1.33$			
Exponential	$\rho_0 = 5.33$	1	$\rho_{0=}40$	1
	$\rho_1 = 4$	1	$\rho_1 = 30$	1
	$\rho_2 = 6.66$	1	$\rho_2 = 50$	1
	$\sigma_0 = 0.13$	1	$\sigma_0 = 1$	0.99
	$\sigma_1 = 0.66$	1	$\sigma_1 = 5$	1
	$\sigma_2 = 1.33$	1	$\sigma_2 = 10$	1
	$\sigma_\epsilon = 1.33$	1	$\sigma_\epsilon = 10$	1

Figures

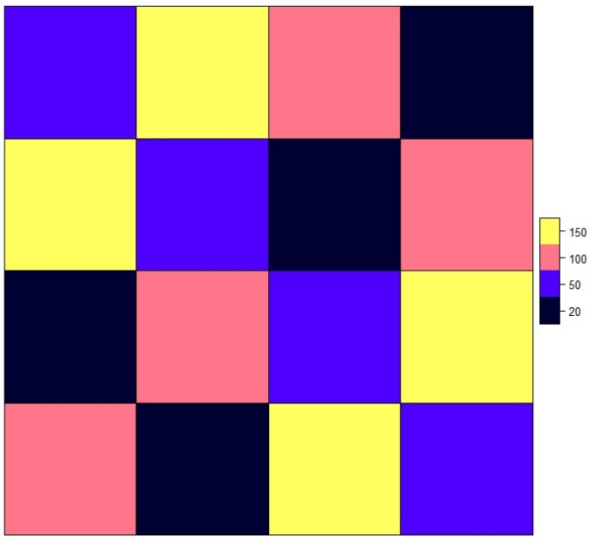

Figure 2.1. Optimal Allocation for SV Intercept with Conditional Autoregrresive and Simultaneous Autoregrresive Rook Behavior

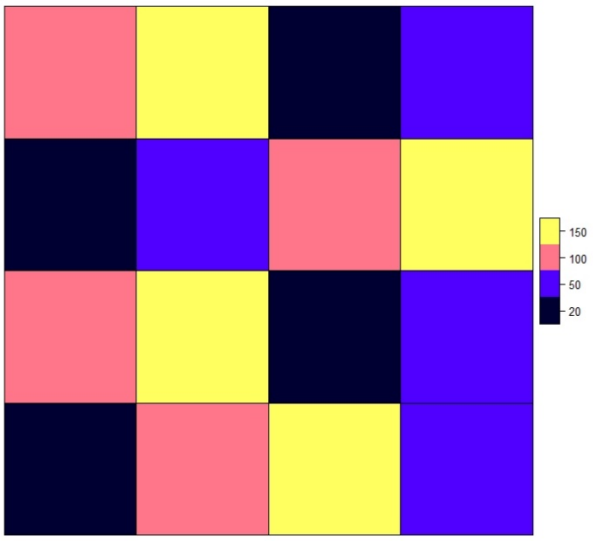

Figure 2.2. Optimal Allocation for SV Intercept with Exponential Covariance

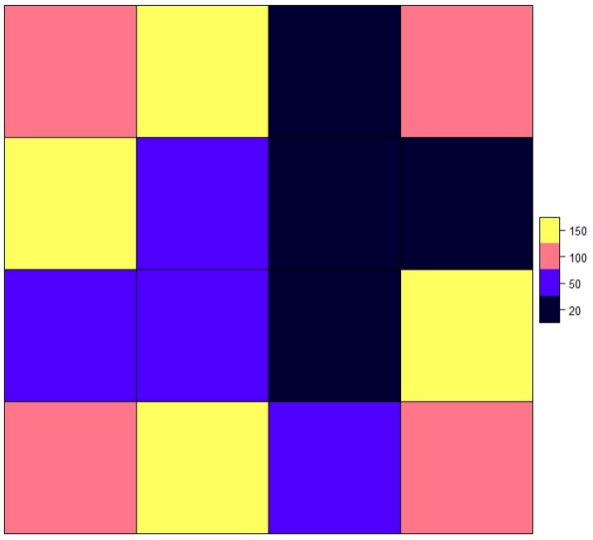

Figure 2.3. Best Allocation for Conditional Autoregrresive Covariance for All Coefficients With Rook Contiguity (All Permutations)

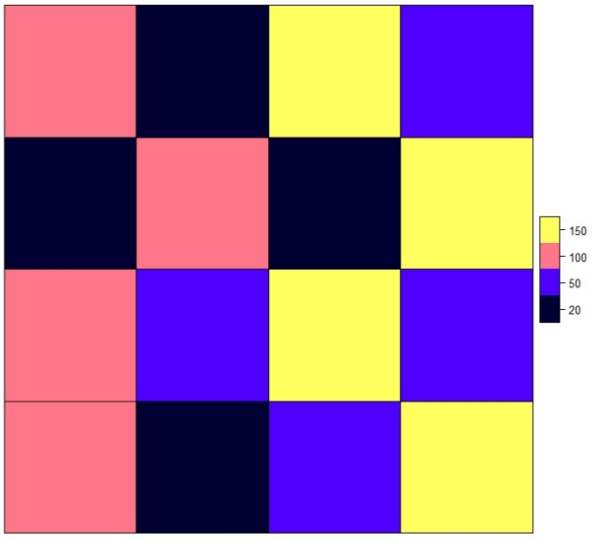

Figure 2.4. Best Allocation for Exponential Correlation Function (All Permutations)

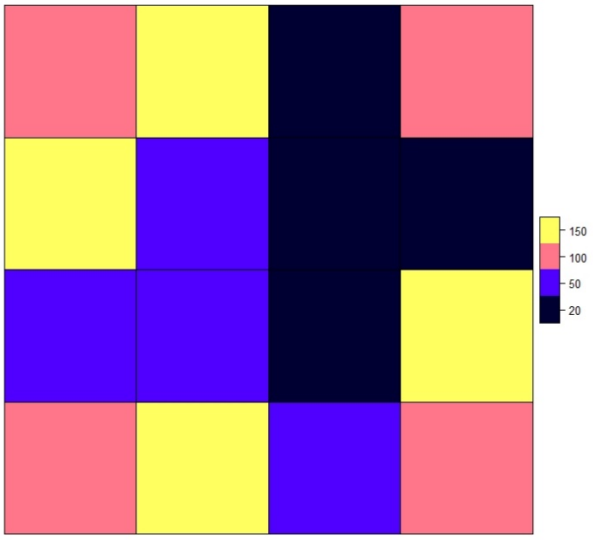

Figure 2.5. Nearly Optimal Design Conditional Autoregrresive Correlation, Rook Contiguity

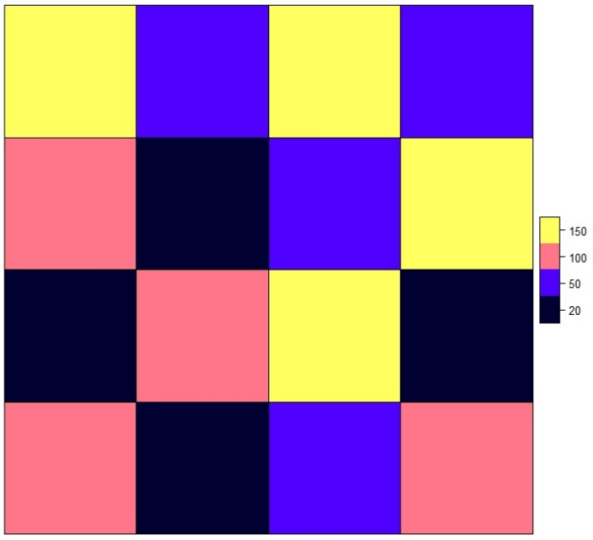

Figure 2.6. Nearly Optimal Design Exponential Correlation

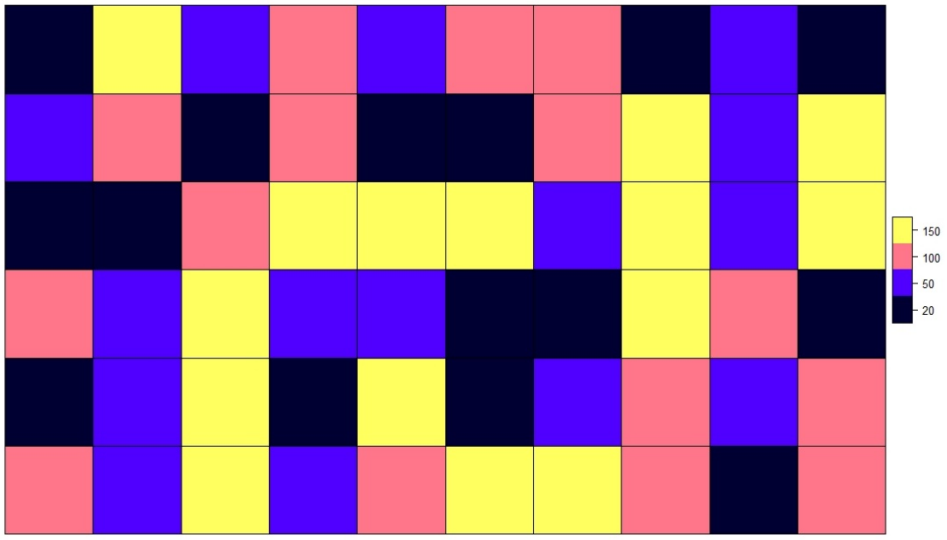

Figure 2.7. Nearly Optimal Design with the Simultaneous Autoregrresive Correlation Function, Queen Contiguity

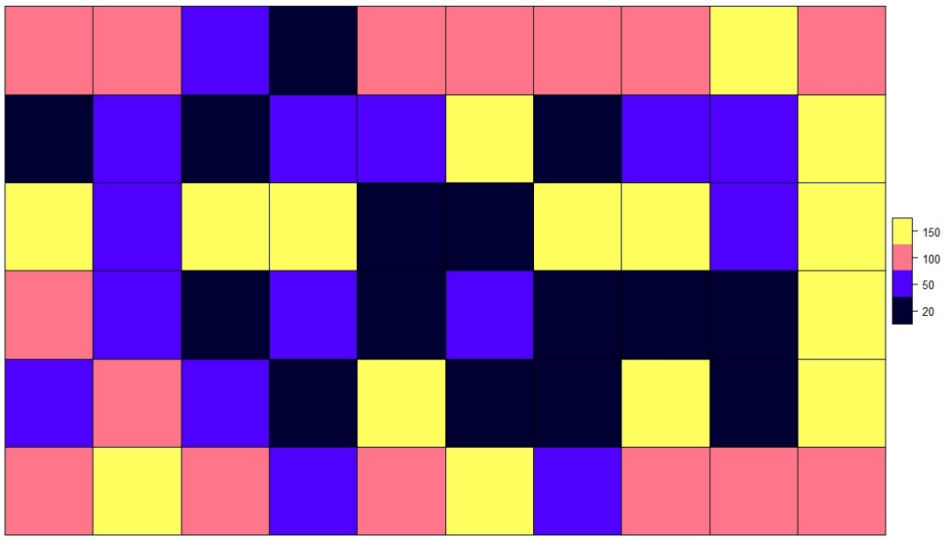

Figure 2.8. Nearly Optimal Design with the Conditional Simultaneous Covariance Function, Rook Contiguity

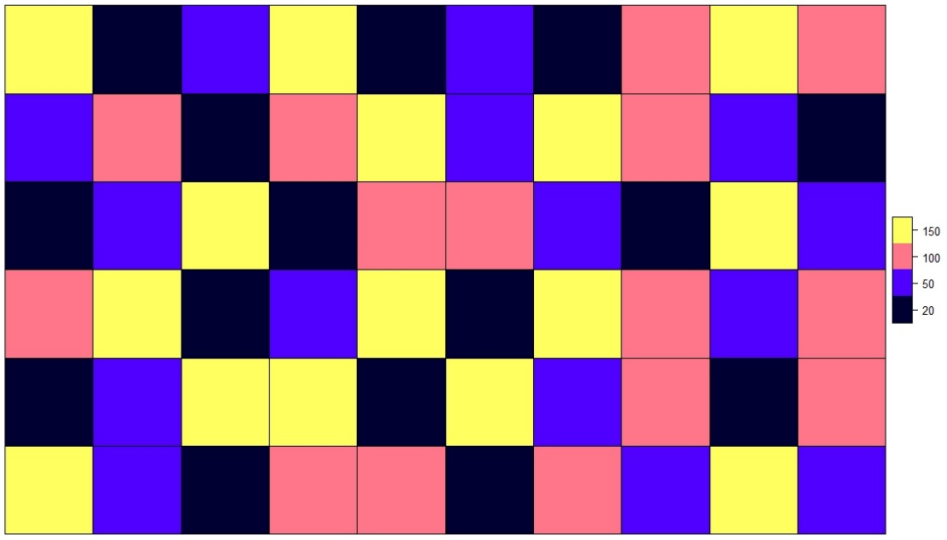

Figure 2.9. Nearly Optimal Design with the Exponential Correlation

CHAPTER III

Optimal Treatment Placement for On-Farm Experimentation

Abstract

The costs of conducting on-farm experiments have decreased with recent technological advances in collecting, storing, and processing geospatial field data. A common Production function for this data is a linear plateau. The ultimate goal is to estimate the linear plateau model with spatially varying coefficients (SVC). A question that has not been well addressed is what spatial experimental design is best to well the goal is to estimating such model. This paper aims to determine the optimal location of treatments when the yield response function is an SVC linear plateau model, and the goal is a D-optimal experimental design. A pseudo-Bayesian approach is taken here because the field's site-specific optimal nitrogen value is unknown. Pseudo-Bayesian D-optimal designs are generated, assuming a fixed number of replications for each treatment level. The resulting designs are more efficient than classic Latin square, strip plot, or completely randomized designs.

Keywords: Linear Plateau Model, On-Farm Experimentation, Pseudo D-optimal Design, Spatially Varying Coefficients.

1. Introduction

Advances in collecting, storing, and processing high-resolution spatial data have lowered the costs of on-farm experimentation. A key innovation is the development of on-the-go applicators, which can precisely deliver treatments to specific plots over a large area. The main goal of this kind of experiment is to obtain accurate, site-specific fertilizer or chemical rates. The data provided from geospatial data layers collected over multiple growing seasons may be voluminous but can provide more information if the treatment locations are selected optimally. An unanswered question for on-farm experimentation is how the treatments should be allocated across space so that the appropriate model can be precisely estimated. The optimal experimental design can vary depending on the assumed data-generating process. Here the yield response to nitrogen is assumed to follow a linear plateau model with spatially varying coefficients (SVC). The objective is to determine where to put the plots to maximize the information gained from the experiment.

Poursina and Brorsen (2022) obtained a Ds-Optimal experimental design assuming that the response model was linear in SVC parameters. They showed that the optimal allocation of treatments based on this criterion was more informative than standard completely randomized designs or randomized strip plots. The Ds-Optimality criterion used by Poursina and Brorsen maximizes the determinant of the Fisher information matrix for a subset of parameters. For crop yield, however, a linear plateau (LP) with spatially varying coefficients is a promising model which is non-linear; hence, the linear response function used by Poursina and Brorsen does not apply. This study examines the performance of an LP with SVC in determining an optimal experimental design for a whole field experiment.

Xiaofei, et al. (2021) conducted a Monte Carlo study to evaluate the performance of several classic experimental designs. They demonstrated that designs based on the random assignment of treatment locations could be improved. Their study used a quadratic plateau yield response model, so they considered non-linearity. They could only select the optimal design from a small set of designs that they chose to simulate. They proposed using blocks such as Latin squares as a practical way to improve over a completely randomized design.

Using the LP model to determine an optimal experimental design creates two problems. First, the LP is a nonlinear, non-differentiable function. Therefore, the information matrix cannot be derived directly from the model's marginal likelihood function. Secondly, the Fisher information matrix for the LP depends on the model's parameters; however, these parameters are unknown. This paper uses a two-step approximation to obtain the Fisher information matrix. The LP model is first approximated with a differentiable model. The differentiable model assumes a known true value for optimal nitrogen levels. Next, the function is linearized to find the Fisher information matrix. We employ a pseudo-Bayesian optimal design approach for the second problem, which considers the parameters as best initial guesses with known distributions. Finally, we assess the robustness of the obtained design by substituting true parameter values for incorrect ones.

Fast algorithms and computational power make SVC models feasible for large data sets (Gelfand, et al., 2003; Mu, et al., 2018; Murakami, et al., 2019). Unfortunately, there is little literature about experimental design when the goal is to estimate SVC models. There is, however, an established literature where parameters are not spatially varying on experimental design (Casler, 2015; Clewer and Scarisbrick, 2013). The main goal of experimental design is to select the levels and number of replications for each treatment such that the production function

is estimable at the highest precision possible (Hanrahan and Lu, 2006). Classical optimal designs assume independence of the observations (e.g., completely randomized designs) or independence within a block (e.g., randomized complete block designs). In practice, these independent assumptions are questionable for most agronomic experiments and are violated when the response follows SVC models.

Optimal experimental designs for LP models have been previously considered, but they did not consider spatial dependence or heterogeneity and its influence on the optimal design. Atkinson and Haines (1996) showed that the optimal design for aspatial LP models has three treatment levels for zero rates (check plots), three for biologically optimal values, and one point for the plateau. Brorsen and Richter (2012) obtained the optimal design of an experiment for a stochastic LP model. Their findings were similar to Atkinson and Haines results, concluding that only three design points were required. Furthermore, both studies were locally optimal designs since they assumed the plateau switch point was known. Ng'ombe and Brorsen (2022) used a Bayesian sampling system to overcome the optimal design problem for the stochastic LP model. They concluded that conducting experiments on a small portion of the field for up to 6 years produced economically optimal outcomes.

The previous research did not consider the potential for model parameters exhibiting spatial structure. The scope and magnitude of the underlying spatial structure affect the amount of information that could be obtained from data. The location of the treatment levels also affects information content and quality, the required number of treatment levels, and the number of replications. These features of SVC models and their extension to optimal experimental design methodology highlight the importance of the role of spatial context in the placement and replication frequency of treatments.

This research considers two scenarios. First, we assume that the experiment is conducted on all potential plots on a farm ('whole farm experimentation'). For this scenario, we obtain nearly optimal designs for 4×4 square and 8×12 rectangular fields. Four equally weighted levels of nitrogen equal to 20, 50, 100, and 150 are considered, each with a uniform prior on the optimal nitrogen value between 90 to 110. We adjusted the method suggested by Poursina and Brorsen (2022) to identify pseudo-Bayesian D optimal designs. Their study showed that the recommended approach is feasible for any field shape. Optimal designs are far more informative than standard experimental designs. They do not impose any extra cost on the application system (assuming that the applicator can switch nitrogen levels between plots). Completely random designs only have 50% efficiency on average. Inefficient designs like strip plots made sense when machinery could not easily apply nutrients and seeds at different levels within a field. Much more is learned from an experiment when more care is given to the placement and replications of treatments.

In a second scenario, we assume that the researcher is interested in conducting experiments on a portion of the farm's fields, for example, 20 percent of the area. Several papers have suggested maximizing the minimum distance between the experimental units to address this issue (Husslage, et al., 2011; Marengo and Todeschini, 1992; Royle and Nychka, 1998). These previous papers did not discuss a coherent approach for allocating treatment levels in an optimal spatial pattern. In this paper, we also find the D-optimal design for allocating treatment levels when locations of plots were selected based on a max-min distance algorithm. We show that, in this situation, the locations of the treatment levels are essential. For both scenarios, we consider two possibilities regarding the spatial structure and spatially varying response patterns. First, we assume that all parameters are spatially varying. For the second scenario, we assume that only

the plateau part of the model exhibits spatial heterogeneity. The intercept and marginal response to fertilizer are assumed to be the same for all locations.

2. Information Matrix for the SVC Linear Plateau Model

When applied to nitrogen applications, von Liebig (1855) 's 'law of the minimum' suggests that nitrogen applications will increase yield up to a point, after which additional nitrogen applications have no effect on yield. The LP model reflects this relationship and has been widely used in agricultural applications (Dhakal, et al., 2019; Harmon, et al., 2016; Hermesch, et al., 1998; Tembo, et al., 2008). Bayesian methods have proven helpful in estimating LP models (Moeltner, et al., 2021; Ouedraogo and Brorsen, 2014). Tembo et al. (2008) argue that the plateau term might vary across fields and years. Poursina and Brorsen (2021) and Lambert and Cho (2022) have previously estimated SVC-LP models.

The SVC-LP model is

$$y_i = \min(\beta_{0i} + \beta_{1i} x_i, P_i) + \epsilon_i \tag{1}$$

where y_i is the yield at location i, β_{0i}, and β_{1i} are location-specific intercepts and slopes, respectively, and P_i is a spatially varying plateau. Assume too that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I})$. The function

$$(\boldsymbol{\theta}_r | \boldsymbol{\Psi}_r) \sim N(\boldsymbol{\mu}_r \mathbf{1}, \boldsymbol{\Psi}_r), r = 1, 2, 3 \tag{2}$$

describes the spatial behavior for each group of parameters in the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{P})$, with $\boldsymbol{\Psi}_r$ as a covariance matrix. We use Bayesian methods to estimate the parameters in (1).

The inverse of the Fisher information matrix can estimate the asymptotic variance of the parameters. The information matrix for the model

$$y = \eta(x, \boldsymbol{\theta}) + \epsilon \tag{3}$$

is

$$\mathbf{I} = \frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \, \boldsymbol{\Omega} \left(\frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^T \tag{4}$$

where $\frac{\partial \eta(x,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ are partial derivatives of the production function with respect to its parameters, and $\boldsymbol{\Omega}$ is the covariance matrix of the response variable y after integrating out the parameters of the model. We can calculate the information matrix with equation (4). Before doing so, an additional step is required because the LP is not differentiable at the biologically optimal value.

Optimal design requires a well-defined positive definite information matrix for a compact set of design spaces. However, Hooshangifar et al. (2022) showed that the design space remains compact when we remove one point from the design area. This allows us to sum over individual information values for each design point to calculate the design's total information. Hence, the following *lemma* can be used to calculate the information matrix for the LP model.

Lemma 1: The approximation of the min of a vector $(X_1, X_2, \cdots, X_n)$ is

$$\min(X_1, X_2, \cdots, X_n) = \lim_{k \to -\infty} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^K\right)^{\frac{1}{k}}$$

For the linear plateau model in each site, we have

$$\min(A, B) = \lim_{k \to -\infty} (0.5 * A^k + 0.5 * B^k)^{\frac{1}{k}}$$

The first derivatives of this function are

$$\frac{\partial}{\partial A} = \lim_{k \to -\infty} 0.5(0.5 * A^k + 0.5 * B^k)^{\frac{1-k}{k}} A^{k-1}$$

$$\frac{\partial}{\partial B} = \lim_{k \to -\infty} 0.5(0.5 * A^k + 0.5 * B^k)^{\frac{1-k}{k}} B^{k-1}$$

Solving one of the limits yields additional, similar information about the other limit. Define $v = \frac{1}{k}$, so $v \to 0^-$. We have

$$\lim_{v \to 0^-} 0.5 \left(0.5\, A^{\frac{1}{v}} + 0.5\, B^{\frac{1}{v}}\right)^{v-1} A^{-\frac{v-1}{v}}$$

$$\lim_{v \to 0^-} 0.5 \left(\left(0.5\, A^{\frac{1}{v}} + 0.5\, B^{\frac{1}{v}}\right) A^{-\frac{1}{v}}\right)^{v-1}$$

$$\lim_{v \to 0^-} 0.5 (0.5 + 0.5 \left(\frac{B}{A}\right)^{\frac{1}{v}})^{v-1}$$

The solution for this limit is

$$\begin{cases} 0 & A > B > 0 \\ \frac{1}{2} & A = B > 0 \\ 1 & 0 < A < B \end{cases}$$

■

Hence, if biologically optimal nitrogen values are known *a priori*, then the derivative of the linear plateau is

$$\frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} 1 & x & 0 \\ 1 & x & 0 \\ \vdots & \vdots & \vdots \\ & \overset{before\ x^*}{\longleftrightarrow} & \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

After linearizing the LP model and integrating out the parameters, the variance of the vector y is approximated by

$$\boldsymbol{\Omega} = Var(\boldsymbol{Y}) = \left(\boldsymbol{D}_1 \boldsymbol{\Psi}_0 \boldsymbol{D}_1 + \boldsymbol{D}_x \boldsymbol{\Psi}_1 \boldsymbol{D}_x + \boldsymbol{D}_p \boldsymbol{\Psi}_2 \boldsymbol{D}_p + \tau^2 \boldsymbol{I}\right) \quad (5)$$

where $\boldsymbol{D}_1, \boldsymbol{D}_x$ and $\boldsymbol{D}_p$ are diagonal matrices with elements equal to the columns of $\frac{\partial \eta(x, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$.

3. Optimal Design

The optimal design of experiments has rich literature from the early 20th century (Smith, 1918). However, the theory of optimal design was developed in a paper by Kiefer (1974). Assume that the experimenter can run N experiments. The theory of optimal design deals with selecting not necessarily distinct N treatment levels. Most classical experimental designs usually assume independence to fit the production function. When production function parameters exhibit spatial heterogeneity, the location of the experiments must also be judiciously selected. Hence, we want to maximize the information an experiment can provide by selecting optimal treatment locations. In the statistics literature, a standard criterion is a D-optimal design, which maximizes the determinant of the Fisher information matrix. Maximizing the Fisher information matrix's determinant is equivalent to minimizing the volume of the confidence ellipsoid of the estimated parameters. Thus, an objective function criterion for the experimental design is required; for example

$$\max_{\xi} \phi(Y, \boldsymbol{\theta}, \xi)$$

where Y is a response variable, $\boldsymbol{\theta}$ is the vector of parameters, ξ is an experimental design, and ϕ is a selection criterion.

The Fisher information matrix does not depend on the unknown model parameters for linear models. So, the experimental designs have closed forms for these models, like factorial designs where treatment levels are selected at the minimum and maximum values of a treatment. However, the Fisher information matrix for nonlinear SVC models does depend on the model parameters. Hence, a chicken and egg situation occurs. The main goal is finding the optimal design to estimate model parameters, but the design depends on these parameters. One solution for this situation is to assume that the parameters are known and then find locally optimal

designs (Chernoff, 1953; Yang and Stufken, 2012). Pseudo-Bayesian methods offer an alternative solution by considering prior distributions for unknown parameters while maximizing the expected value of the Fisher information matrix (Chaloner and Verdinelli, 1995; Dette and Neugebauer, 1997). Pseudo-Bayesian designs are robust to designs when the true values for the parameters are unknown. For example, while the true biologically optimal nitrogen level is unknown, a researcher can specify a prior distribution for model parameters based on results from previous experiments. The following section investigates the robustness of the approach to inaccurate guesses of the initial values for the parameters of the mean response function.

We consider two spatial covariance matrices: a spatial autoregressive (SAR) covariance matrix (Anselin, 1988) and a Spatial Gaussian process (SGP) covariance with negative exponential decay. The SGP exponential covariance function is

$$\mathrm{cov}\big(\beta_r(s_i), \beta_r(s_j)\big) = \sigma_r^2 \exp\left(-\frac{d_{ij}}{\rho_r}\right) \ , r = 0,1,2 \qquad (6)$$

where d_{ij} are distances between location i and j, σ is the sill, and the ρ is the effective range of spatial covariance. The SAR covariance function is

$$\boldsymbol{\Sigma}_r^{SAR} = \sigma_r^2((\boldsymbol{I} - \rho_r \boldsymbol{W}^{*\prime})(\boldsymbol{I} - \rho_r \boldsymbol{W}^{*}))^{-1} \ , r = 0,1,2 \qquad (7)$$

where σ_r^2 is a common variance for the *r*'th parameter, $\boldsymbol{W}^*$ is a row-standardized contiguity matrix, and ρ_r the degree of spatial dependence.

4. Application and results

Assume that the treatment level and the number of replications are fixed, and we want to select the treatment location for 16 experiments in a field. Firstly, we assume that the field is portioned into a 4 × 4 square. Next, we consider a larger field with an 8×12 rectangular shape in the second scenario. The researcher wants to run 16 experiments with the four levels of treatments

given and four replications for each treatment level (a balanced treatment). We first consider a model with only the plateau varying over space and then generalize the model to have the intercept and slope parameters vary according to location. The information matrix given in equation (4) depends on the unknown parameters of spatial covariance as well as the optimal nitrogen value. We apply the pseudo-Bayesian method to overcome this problem. For the pseudo-Bayesian method, we assume that true values of the covariance parameters are known (locally optimal) and a uniform prior distribution between 90 and 110 for the optimal nitrogen values. With this prior distribution, we can be sure that there is at least one point on the plateau and one point in the range of possible optimal nitrogen rates.

Suppose we want to allocate four equally weighted treatment levels in a 4×4 square. There are more than 63 million possible permutations that require 8.1 gigabytes of RAM to store. All possible permutations are calculated using the RcppAlgos package (Wood, 2020). We assume that the true values of the spatial covariance parameters are known and equal to $\rho_1 = 0.8$, $\sigma_1^2 = 20$, $\rho_N = 0.9$, $\sigma_N^2 = 10$, $\rho_p = 0.8$, and $\sigma_p^2 = 30$ for the SAR model. These values for the SGP covariance structure are $4, 0.1, 3, 0.5, 10$, and 1. The value for $\sigma_\epsilon^2 = 1$ for both covariance matrices.

Figure 3.1 shows the optimal allocation when only the plateau is spatially varying. The result is the same for both SAR and SGP covariance functions. In this allocation, higher levels are spread out as much as possible and the lower levels, which are not spatially correlated, clump together in the middle of the field. Figures 3.2 and 3.3 show the optimal allocation for the SAR and SGP covariance structures when all parameters vary across space. This situation has no clear pattern when all parameters exhibit spatial variability. Hence, optimal designs could vary by field depending on the number of plots and the spatial correlation parameters.

To determine the efficiency of the obtained designs, we use relative efficiency as in Poursina and Talebi (2014). Relative efficiency (E) of two designs $\xi_1$and ξ_2 can be calculated as

$$E = \frac{|I(\xi_1, \Theta)|}{|I(\xi_2, \Theta)|}. \quad (8)$$

Since for classic designs like Latin square, strip plot, and randomly assigned designs, more than one allocation is possible, we consider the average efficiency of these designs over all possible Latin square (576) and strip plot (24) designs. Table 3.1 illustrates the efficiency of the classic designs relative to the optimal allocation of the locations.

An important issue that should be addressed here is the robustness of obtained designs against bad guesses about the values of the variance parameters. We consider two different scenarios. First, the assumption is that misspecification is not severe and the $\boldsymbol{\nu}_{real} = \frac{3}{4}\boldsymbol{\nu}_{assumed}$ or $\boldsymbol{\nu} = \frac{4}{3}\boldsymbol{\nu}_{assumed}$ where $\boldsymbol{\nu}$ contains all variance parameters. In another scenario, we consider the severe misspecification where $\boldsymbol{\nu}_{real} = \frac{1}{10}\boldsymbol{\nu}_{assumed}$ or $\boldsymbol{\nu}_{real} = 10\boldsymbol{\nu}_{assumed}$. In both scenarios, we change the parameters to have more variance or less information about the field. Table 3.2 shows the results of the robustness check against the misspecification. In most cases, the optimal design does not change with these assumptions. Even when it does, the lowest efficiency found was still 96%.

Research by Ng'ombe and Brorsen (2021) finds that it is economically optimal to experiment on only a small portion of a field. Several papers consider the selection of a portion of the field when there is spatial behavior in the model (Husslage, et al., 2011; Lin and Tang, 2015; Pronzato and Müller, 2012). We use this past research to decide the plot locations, but we must still allocate the treatment levels to these locations. In the nonspatial case, Ng'ombe and

Brorsen (2022) find that the economic optimum is to experiment on only a portion of the plots, so optimal design on only a portion of the plots is essential to consider.

For this purpose, we consider an 8×12 field where the researcher decides to run 16 experiments over this field with the same treatment levels and the number of replications as before. We use the maximin package in R (Sun and Gramacy, 2021), which provides a space-filling design based on the maximin distance criterion to maximize the information gained from experiments to select the location of the experiments. The spatial covariance matrix is calculated for the whole farm. We selected the portion of the matrix that depicts the covariance matrix for the selected plots based on the maximin criterion. The final covariance matrix is 16×16; however, the initial covariance matrix was 96×96. We use the same values and distribution for the model parameters and biological plateau-level nitrogen.

The optimal designs are given in figures 3.4 and 3.5. In figure 3.5, with the exponential covariance function, the largest nitrogen levels are placed farther apart, much like the case where only the plateau is random. Figure 3.4 shows a more even allocation of treatment levels with the SAR covariance function. If we divided the field into four big plots, we have all levels of nitrogens in each plot. Since the spatial correlation decreases between the plots with more distance, relative efficiency for randomly assigned treatment levels with SAR covariance function increases from 41 percent to 58 percent. In comparison, the relative efficiency surged from 51 to 64 percent for the exponential covariance function. Hence, the locations of the treatment levels do not affect the information as much as before; however, they still play a vital role in gaining information.

5. Conclusion and disscusion

In this paper, we consider a non-differentiable production function (LP). The researchers' budget predetermines the treatment levels and the number of replications of each level. The optimal location for the treatment levels is found based on the D optimal criterion that maximizes the determinant of the Fisher information matrix for the LP model with spatially varying coefficients. Current technology lets us apply different levels of treatment without extra cost. So, finding the optimal location for the treatments helps to increase the amount of information gained from an experiment without imposing an extra cost on the project. The obtained designs are far more informative than the classical designs like Latin square, strip plot, and random designs. These designs also are robust against the misspecification of the parameters, but optimal designs are obtained, assuming that LP is the true functional form and it is assumed that the spatial covariance functional form is also known.

Funding

This work was supported by the A.J. and Susan Jacques Chair, the Oklahoma Agricultural Experiment Station, and USDA National Institute of Food and Agriculture [Hatch Project number OKL03170].

6. References

Anselin, L. (1988). Spatial econometrics: methods and models; Springer Science & Business Media.

Atkinson, A.C., and L.M. Haines. (1996). "14 Designs for Nonlinear and Generalized Linear Models." *Handbook of Statistics* 13:437-475.

Brorsen, B.W., and F.G.-C. Richter. (2012). "Experimental Designs for Estimating Plateau-Type Production Functions and Economically Optimal Input Levels." *Journal of Productivity Analysis* 38:45-52.

Casler, M.D. (2015). "Fundamentals of Experimental Design: Guidelines for Designing Successful Experiments." *Agronomy Journal* 107:692-705.

Chaloner, K., and I. Verdinelli. (1995). "Bayesian Experimental Design: A Review." *Statistical Science* 10:273-304.

Chernoff, H. (1953). "Locally Optimal Designs for Estimating Parameters." *The Annals of Mathematical Statistics* 24:586-602.

Clewer, A.G., and D.H. Scarisbrick. (2013). *Practical Statistics and Experimental Design for Plant and Crop Science*: John Wiley & Sons.

Dette, H., and H.-M. Neugebauer. (1997). "Bayesian D-Optimal Designs for Exponential Regression Models." *Journal of Statistical Planning and Inference* 60:331-349.

Dhakal, C., K. Lange, M.N. Parajulee, and E. Segarra. (2019). "Dynamic Optimization of Nitrogen in Plateau Cotton Yield Functions with Nitrogen Carryover Considerations." *Journal of Agricultural and Applied Economics* 51:385-401.

Gelfand, A.E., H.-J. Kim, C. Sirmans, and S. Banerjee. (2003). "Spatial Modeling with Spatially Varying Coefficient Processes." *Journal of the American Statistical Association* 98:387-396.

Hanrahan, G., and K. Lu. (2006). "Application of Factorial and Response Surface Methodology in Modern Experimental Design and Optimization." *Critical Reviews in Analytical Chemistry* 36:141-151.

Harmon, X., C.N. Boyer, D.M. Lambert, J.A. Larson, and C.O. Gwathmey. (2016). "Comparing the Value of Soil Test Information Using Deterministic and Stochastic Yield Response Plateau Functions." *Journal of Agricultural and Resource Economics* 41:307-323.

Hermesch, S., K. Egbert, and J. Eissen. (1998). "Description of a Growth Model: The Linear-Plateau Model." *AGBU, Wageningen Agricultural University, Animal Breeding and Genetics Group, Wageningen, The Netherlands.*

Hooshangifar, M., H. Talebi, and D. Poursina. (2022). "D-Optimal Design for Logistic Model Based on More Precise Approximation." *Communications in Statistics-Theory and Methods* 51:1975-1992.

Husslage, B.G., G. Rennen, E.R. Van Dam, and D. Den Hertog. (2011). "Space-Filling Latin Hypercube Designs for Computer Experiments." *Optimization and Engineering* 12:611-630.

Kiefer, J. (1974). "General Equivalence Theory for Optimum Designs (Approximate Theory)." *The Annals of Statistics* 2:849-879.

Lambert, D.M., and W. Cho. (2022). "Geographically Weighted Regression Estimation of the Linear Response and Plateau Function." *Precision Agriculture* 23:377-399.

Lin, C.D., and B. Tang. (2015). "Latin Hypercubes and Space-Filling Designs." *Handbook of design and analysis of experiments*:593-625.

Marengo, E., and R. Todeschini. (1992). "A New Algorithm for Optimal, Distance-Based Experimental Design." *Chemometrics and Intelligent Laboratory Systems* 16:37-44.

Moeltner, K., A.F. Ramsey, and C.L. Neill. (2021). "Bayesian Kinked Regression with Unobserved Thresholds: An Application to the Von Liebig Hypothesis." *American Journal of Agricultural Economics* 103:1832-1856.

Mu, J., G. Wang, and L. Wang. (2018). "Estimation and Inference in Spatially Varying Coefficient Models." *Environmetrics* 29:e2485.

Murakami, D., B. Lu, P. Harris, C. Brunsdon, M. Charlton, T. Nakaya, and D.A. Griffith. (2019). "The Importance of Scale in Spatially Varying Coefficient Modeling." *Annals of the American Association of Geographers* 109:50-70.

Ng'ombe, J.N., and B.W. Brorsen. (2022). "Bayesian Optimal Dynamic Sampling Procedures for on-Farm Field Experimentation." *Precision Agriculture*:1-23.

Ouedraogo, F.B., and B.W. Brorsen. (2014). "Bayesian Estimation of Optimal Nitrogen Rates with a Non-Normally Distributed Stochastic Plateau Function." Paper presented at SAEA. Dallas, Texas, USA.

Poursina, D., and B.W. Brorsen. (2022). "Nearly Ds Optimal Assigned Location Design for a Linear Model with Spatially Varying Coefficients." Paper presented at SAEA. New Orleans, Louisiana, USA.

Poursina, D., and W. Brorsen. (2021). "Site-Specific Nitrogen Recommendation: Using Bayesian Kriging with Different Correlation Matrices." Paper presented at AAEA. Austin, Tx, USA.

Poursina, D., and H. Talebi. (2014). "Modified D-Optimal Design for Logistic Model." *Journal of Statistical Computation and Simulation* 84:428-437.

Pronzato, L., and W.G. Müller. (2012). "Design of Computer Experiments: Space Filling and Beyond." *Statistics and Computing* 22:681-701.

Royle, J.A., and D. Nychka. (1998). "An Algorithm for the Construction of Spatial Coverage Designs with Implementation in Splus." *Computers & Geosciences* 24:479-488.

Smith, K. (1918). "On the Standard Deviations of Adjusted and Interpolated Values of an Observed Polynomial Function and Its Constants and the Guidance They Give Towards a Proper Choice of the Distribution of Observations." *Biometrika* 12:1-85.

Sun, F., and R.B. Gramacy, (2021), "Space-Filling Design under Maximin Distance." https://CRAN.R-project.org/package=maximin.

Tembo, G., B.W. Brorsen, F.M. Epplin, and E. Tostão. (2008). "Crop Input Response Functions with Stochastic Plateaus." *American Journal of Agricultural Economics* 90:424-434.

von Liebig, J.F. (1855). *Principles of Agricultural Chemistry: With Special Reference to the Late Researches Made in England*: Walton & Maberly.

Wood, J., (2020), "Rcppalgos: High Performance Tools for Combinatorics and Computational Mathematics." https://cran.r-project.org/web/packages/RcppAlgos/index.html

Xiaofei, L., M. Taro, and B. David. (2021). "Economic Performances of Trial Design Methods in on-Farm Precision Experimentation: A Monte Carlo Evaluation"Unpublished".

Yang, M., and J. Stufken. (2012). "Identifying Locally Optimal Designs for Nonlinear Models: A Simple Extension with Profound Consequences." *The Annals of Statistics* 40:1665-1681.

Tables

Table 3.1. Relative Efficiency of Designs Based on Different Correlation Matrices for Whole-Farm Experimentation

		Average Efficiency		Maximum Efficiency	
Design of Experiment	Number of Designs	SAR	Exponential	SAR	Exponential
Latin Square	576	71.29	61.81	0.94	0.91
Random	1000	40.85	51.01	0.48	0.96
Strip plot	24	37.59	28.89	0.56	0.51

Table 3.2. Robustness of Nearly Optimal Designs Against Misspecification in Variance Parameters

Model	Parameter	Efficiency	Parameter	Efficiency
	$\rho_1 = 0.6$	1	$\rho_1 = 0.08$	0.96
	$\rho_N = 0.675$	1	$\rho_N = 0.09$	1
	$\rho_p = 0.6$	1	$\rho_p = 0.08$	1
SAR	$\tau_1 = 15$	1	$\tau_1 = 2$	0.95
	$\tau_N = 7.5$	1	$\tau_N = 1$	1
	$\tau_p = 22.5$	1	$\tau_p = 3$	1
	$\sigma_\epsilon = 1.33$	1	$\sigma_\epsilon = 10$	1
	$\rho_1 = 5.33$	1	$\rho_{1=}40$	1
	$\rho_N = 4$	1	$\rho_N = 30$	1
	$\rho_p = 6.66$	1	$\rho_p = 50$	1
Exponential	$\sigma_1 = 0.13$	1	$\sigma_1 = 1$	0.99
	$\sigma_N = 0.66$	1	$\sigma_N = 5$	1
	$\sigma_p = 1.33$	1	$\sigma_p = 10$	1
	$\sigma_\epsilon = 1.33$	1	$\sigma_\epsilon = 10$	1

Figures

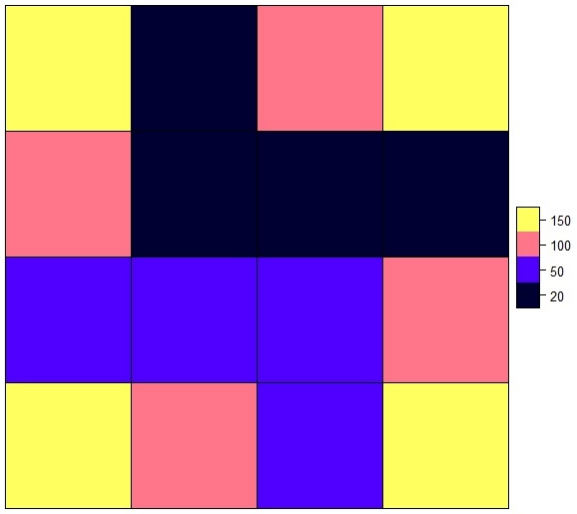

Figure 3.1. Optimal Allocation for Simultaneous Autoregrresive and Exponential when Only the Plateau Is Spatially Varying

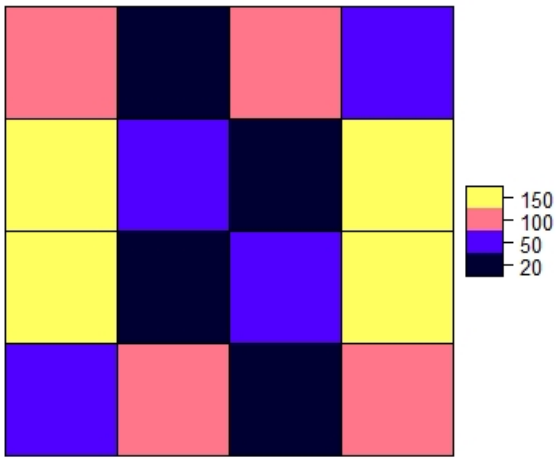

Figure 3.2. Best Allocation for exponential Covariance with Whole-Field

Experimentation

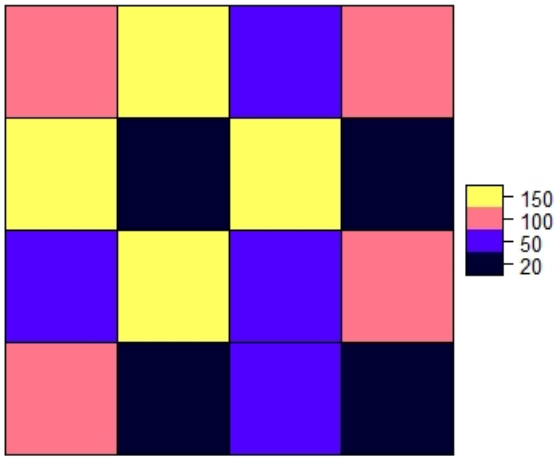

Figure 3.3. Best Allocation for Simultaneous Autoregrresive Covariance and Whole-Field Experimentation

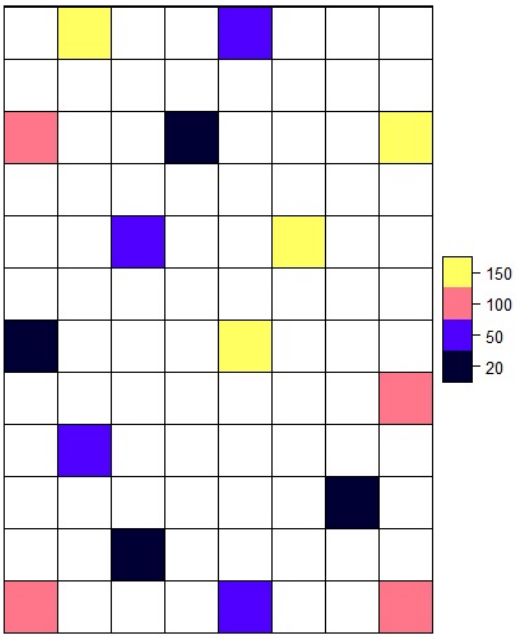

Figure 3.4. Best Allocation for 16 Selected Locations with Simultaneous Autoregrresive Covariance

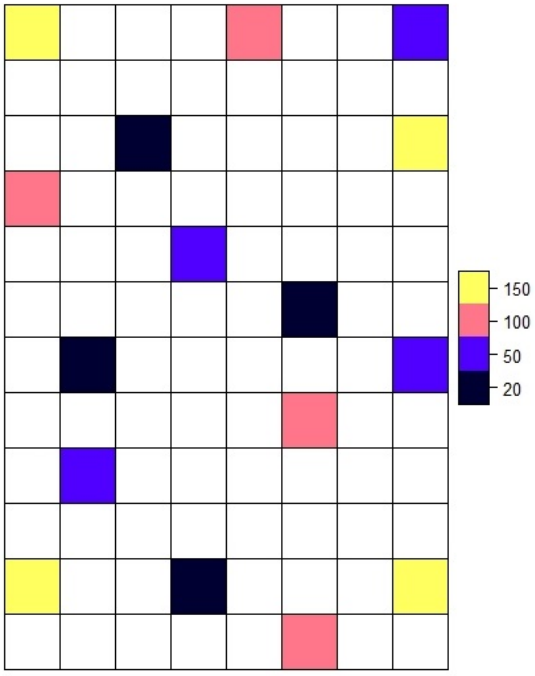

Figure 3.5. Best Allocation for 16 Selected Locations with Exponential Covariance

CHAPTER IV

Fully Bayesian Economically Optimal Design for Spatially Varying Coefficient Linear Stochastic Plateau Model in Several Consecutive Years

Abstract

Experimentation to guide optimal fertilizer selection is moving toward on-farm experimentation due to the uncertainty of small on-station experiments. However, there is no agreement on the optimal way to conduct on-farm experimentation, which motivated this paper. Optimal on-farm experimentation is addressed using fully Bayesian decision theory. Monte Carlo integration was used, assuming a linear stochastic plateau model with spatially correlated plateau parameters. The spatially varying coefficient model can be used to guide the application of site-specific nitrogen. The actual economic optimal nitrogen values vary from 130 to 180 for different plots based on the data-generating process used for simulation. Of the designs considered, the results show that experimenting on 15 plots of a field with treatment levels of 35, 130, 165, and 230 with 2, 3, 5, and 5 replications maximized the farmers' profit over several years. The third year was the best time to quit experimenting.

Keywords: Linear stochastic plateau, Profit function, Simulated based Bayesian design, Spatially varying coefficients, Utility function.

1. Introduction

Next-generation precision agriculture technologies could contribute more than $47 billion to the economy (US Department of Agriculture, 2019). One crucial goal of precision agriculture is applying a site-specific amount of nitrogen fertilizer. Improving nitrogen efficiency can save farmers lots of money and help reduce environmental emissions (Matson, et al., 1998). Computer scientists and engineers have developed the technical ability to apply site-specific nitrogen. However, a significant hurdle to adoption has been acquiring information on which to base site-specific recommendations. On-farm experimentation is a promising source of information (Bullock, et al., 2019; Paccioretti, et al., 2021; Tanaka, et al., 2022). The question addressed here is, what is the expected profit-maximizing way to use on-farm experimentation to guide precision nitrogen recommendations?

There is a vast literature on optimal nitrogen fertilizer value over the last two centuries. Researchers relied on small experiments conducted on agricultural stations for more than a century (Colyer and Kroth, 1968; Hanumantha, 1965; Huelsen, 1932; Singh and Sharma, 1968). Although small experiments provide valuable information, the uncertainty regarding the results and heterogeneity between different fields make these results unreliable for many farmers (Bullock and Mieno, 2017; Rodriguez, 2014). Sellars, et al. (2020) showed that farmers use a higher nitrogen rate than the maximum return to nitrogen (MRTN) which is a rational way of responding to the uncertainty about the results of small experimentations for their field.

Large-scale on-farm experimentation can overcome the uncertainty of small experiments (de Oliveira Ferreira, et al., 2021; Evans, et al., 2020; Lacoste, et al., 2022). However, there are two extreme points of view on conducting on-farm experimentation. Lambert and Cho (2022) run a geographically weighted regression (GWR) on a strip plot design that contains the zero nitrogen rate in a part of the field, which could cause a significant cost to farmers due to yield loss. On the other hand, Trevisan, et al. (2021) reduced the variation in nitrogen levels. They considered seven levels of nitrogen rate from 154 to 235 (kg ha^{-1}), where they believe the optimal value is in this range. There is a trade-off between the information gained from the experiments by varying the nitrogen levels more and the cost of conducting the experiments due to reduced yield. This paper determines what nitrogen levels and how much replication for each level is optimal to maximize the net present value of farmers' profit over the years.

Ng'ombe and Brorsen (2022) obtained an optimal Bayesian simulated design for on-farm experimentation. They considered a linear stochastic plateau model and found the best time to quit the experiment based on the net return value. In addition, they suggested using a portion of the field to reduce the cost and increase the net present value of the experimentation. This research, however, has two significant limitations. First, they consider the level of nitrogen and their replications fixed and predetermined. In addition, they ignored the spatial behavior for on-farm experimentations, so they had to use uniform optimal nitrogen rates. In this paper, we address both limitations and find the optimal levels, replications, and the portion of the field needed for the experiment for a linear stochastic plateau model with a spatially varying plateau parameter that can provide site-

specific optimal nitrogen values. In addition, we also consider the year when it is optimal to quit the experimentation.

This paper will answer an essential question in experimental design and agricultural economics literature. The first question is: "do we need to experiment on all fields when doing on-farm experimentation to find the optimal economic value for the fertilizer?" If the answer to this question is no, then how many plots are needed to maximize the farmer's profit over some consecutive years? We need an optimal proportion of plots to conduct the experimentation. The second necessary question is: "what treatment levels and replications can provide this information?" The last question we will answer in this paper is: “if we use the optimal levels and replications, in which year should we quit the experimentation?” This paper uses a two-step procedure to answer these questions. In the first step, we simulate the data for different designs considering all scenarios. In the second step, we consider a surface on the simulated design and find the design that maximizes the utility function. This optimization results in a fully Bayesian optimal design based on the farmers' profit for T periods of time.

2. Spatially Varying Coefficient Stochastic Linear Plateau

The linear plateau (LP) model is a well-known production function in agricultural applications (Dhakal, et al., 2019; Poursina and Brorsen, 2021; Villacis, et al., 2020). The LP models reflect the yield behavior when the fertilizer response increases to a certain point and gets flat. Bayesian estimation is a reliable and efficient way to estimate LP models (Cho, et al., 2020; Moeltner, et al., 2021; Ng'ombe and Lambert, 2021). Tembo, et al. (2008) argued that the plateau function could vary across different years and within

fields. Lambert and Cho (2022); Park et al. (2018); Sarkar and Lupi (2022) consider the spatially varying plateau function, which shows the popularity and acceptability of spatially varying stochastic plateau.

The LP model for several years with stochastic plateau and spatial behavior can be presented as

$$y_{it} = \min(\beta_0 + \beta_1 N_{it}, P_i + \nu_t) + \phi_t + \epsilon_{it} \quad (1)$$

where y_{it} depicts the yield in plot i at time t, β_0, and β_1 are intercept and slope, P_i is the plateau value for each plot, ν_t is the plateau year random effect, and ϕ_t is the intercept year random effect. Also, assume that $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, $\boldsymbol{\nu} \sim N(0, \sigma_\nu I)$, and $\boldsymbol{P} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is the covariance matrix that describes the spatial behavior of the plateau, and all random parts are independent of each other.

3. Methodology

Muller (1999) first defined Monte Carlo Bayesian simulation-based optimal designs as a decision problem. Bayesian decision methodology now dominates the optimal experimental design for complicated non-standard classical optimality criteria (Ng'ombe and Brorsen, 2022; Overstall and Woods, 2017; Ryan, et al., 2015; Ryan, et al., 2015; Seeger, et al., 2007). Consider the experimental design problem as a decision problem that can be explained by a utility function of $U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})$ where $\boldsymbol{d}$ is the design of the experiment, $\boldsymbol{\theta}$ contains the model parameters, and $\boldsymbol{y}$ is the observable data. Here $\boldsymbol{d}$ is the experimental design, and we are looking for a design that maximizes the expected utility function over all possible scenarios of the parameters and observable response through the probability density function $p_{\boldsymbol{d}}(\boldsymbol{\theta}, \boldsymbol{y})$. We assume that the decision maker is risk neutral, so the

decision maker is a rational person who selects an action (design of experiment) that maximizes the expected profit averaging all relevant unknown values of the model.

Consider the general form of a decision problem based on a given utility function $U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})$ and the probability model $p_{\boldsymbol{d}}(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{\theta})p_{\boldsymbol{d}}(\boldsymbol{y}|\boldsymbol{\theta})$. Müller (2005) formulated an optimal design as

$$\boldsymbol{d}^* = \arg\max_{\boldsymbol{d}} U(\boldsymbol{d}) \tag{2}$$

where $U(\boldsymbol{d}) = \iint U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\boldsymbol{\theta}d\boldsymbol{y}$ is the expected utility. Usually, the target function $U(\boldsymbol{d})$ cannot be calculated in a closed form, especially when the model is non-linear. However, this function can be approximated by a Monte Carlo integration since the prior distribution $p(\boldsymbol{\theta})$, and the sampling model $p_{\boldsymbol{d}}(\boldsymbol{y}|\boldsymbol{\theta})$ are available. Hence, we can generate the Monte Carlo sample for the $(\boldsymbol{\theta}_j, \boldsymbol{y}_j), j = 1, \dots, M$ and obtain the approximated $\widehat{U}(\boldsymbol{d}) = \frac{1}{M} \sum U(\boldsymbol{d}, \boldsymbol{\theta}, \boldsymbol{y})$.

Optimal design is an optimization problem. So, the next step to finding the design of the experiment is to maximize the approximated utility function. One solution to this problem is looking at the whole surface for all simulated designs. Hence the next step is to find a surface all over the obtained utility functions and find the optimal design of the experiment.

To find the optimal design for the spatially varying stochastic linear plateau model, we simulated on-farm field trials over T periods for each farm. So, the objective is to maximize the farmer's net present value (NPV) over T years of the experimentations. Therefore, the optimal design is

$$d^* = \arg\max_d U(d) = \arg\max_d \sum_{t=1}^{T} NPV_t\,(d)$$

where NPV_t is the tth period NPV. So the NPV is the discounted expected profit over the years and can be calculated as

$$NPV_t(d) = \sum_{f=1}^{F}\sum_{i=1}^{I} \frac{\pi_{ift}}{(1+rate)^t} = \sum_{f=1}^{F}\sum_{i=1}^{I} p\frac{E\left(y_{ift}\middle|d,\theta\right)}{(1+rate)^t} - \frac{rN_{ift}}{(1+rate)^t}$$

where i shows the plots in farm f , p is the price of output, N_{ift} is the selected level of fertilizer in location i for farm f at time t, $y_{ift}$is the production function for the spatially varying coefficient stochastic plateau model, and $rate$ is the interest rate.

The maximizing problem involves the treatment levels and their replications in addition to the proportion of a field used for experimenting. In this paper, we restrict the number of treatment levels to four but do not specify the treatment level values and their corresponding replications. In addition, we consider three different percentages of experimentation for the first year and reduce the number of experiments over the years. The rest of the farm is filled by the farmer's optimal value for the first year. Then, the site-specific optimal values are calculated for the following years based on the posterior distribution. The site-specific optimal nitrogen value can be obtained from

$$\max_{N_{it}} E\left(\pi_{it}|N_{ift}\right) = \max_{N_{ift}} \int [p(min(\beta_0 + \beta_1 N_{it}, P_i + \nu_t)) - rN_{it}]f(\boldsymbol{\Psi})\boldsymbol{d\Psi}$$

where the $\boldsymbol{\Psi}$ contains all the parameters which should be estimated, and f is the posterior distribution function of parameters. Since, in this problem, both sides in the

plateau model are random, we use the analytical solution given by Poursina and Brorsen (2021) to find the site-specific optimal nitrogen value for each plot in each farm.

4. Monte Carlo Simulation

We simulate 30 farms with similar spatial behavior with 100 plots over eight years where treatment levels and replications are selected randomly. We assume the plot sizes are large enough for the machines to easily switch the nitrogen rate from one plot to the next. The true simulated production function is given in (1). Since the model is extraordinarily complex, simulation and fitting models are time-consuming. Simulating eight years of data for each design point takes three days for 30 farms. Hence, we simulate forty designs for every proportion (70%, 30%, and 15%). The proportion was reduced by 10 percent of the total number of experiments for each consecutive year. So, the number of plots for experimentation starts from 70 for the first year and is reduced to 46 for the last year in the first simulation. These numbers are 30 and 20 for the second scenario; and 15 and 10 for the last scenario. The experimented plots are selected based on the maximin optimal criterion on the filling space experimental design, which maximizes the minimum distance between two selected plots to maximize the gained information through the maximin package in R (Sun and Gramacy, 2021). Figure 4.1 depicts the simulation process for one farm over eight years. This process ran for 30 farms to consider all possible randomness in the data-generating process. Each design point which contains 30 farms takes three days to simulate and find the optimal nitrogen value on a desktop

computer with a i-5 9500 and 32 gigabytes of RAM. Hence, we simulated 40 designs for each selected proportion.

The true model for Monte Carlo simulation is given in 1. Poursina and Brorsen (2021) argued that the conditional autocorrelation function for spatial behavior in LP models could fit and predict better in the spatial data when the distance between the plots' centers is fixed. In addition, this model is much faster to estimate than decreasing correlation functions such as the exponential. Hence, the simulated model for all of the designs is

$$y_{it} = \min(105 + 0.7N_{it}, P_i + \nu_t) + \phi_t + \epsilon_{it} \tag{3}$$

$$P \sim N(194, \frac{1}{0.003}(diag(\boldsymbol{W}\mathbf{1}) - 0.5\boldsymbol{W})^{-1})$$

$$\nu \sim N(0,400)$$

$$\phi \sim N(0,25)$$

$$\epsilon \sim N(0,9)$$

where $\boldsymbol{W}$ is the contiguity matrix of the field, and $\mathbf{1}$ is a vector of ones. To reduce the variation in the model, the antithetic method is used for simulating the year random effect in both plateau and intercept year random effect . The corn price is \$4.5 $bu^{-1,}$ and the nitrogen price is \$0.45 $kg^{-1,}$ so the expected profit function is

$$E(\pi) = \sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{i=1}^{I} 4.5E(Y_{ift}) - 0.45N_{ift}$$

where the yield function is given in (3).

When the corn yield is simulated on each plot, we use the Hamiltonian Monte Carlo (HMC) algorithm provided by Rstan (Carpenter, et al., 2017) to fit a hierarchical Bayesian model and find the posterior distribution of the parameters. We consider a normal distribution with large variance (non-informative) as priors for the intercept and plateau to make sure that the experimentation can provide information. The priors are

$$\beta_0 \sim N(100, 2500), \beta_1 \sim N(0,4), P \sim N\left(\bar{p}, \frac{1}{\tau}(diag(\boldsymbol{W}\boldsymbol{1}) - \rho \boldsymbol{W})^{-1}\right)$$

$$\bar{p} \sim N(100,2500), \rho \sim U(0,1), \nu \sim N(0, \sigma_\nu), \phi \sim N\left(0, \sigma_\phi\right)$$

$$\epsilon \sim N(0, \sigma_\epsilon), \sigma_\epsilon \sim half - cauchy(3,6), \sigma_\phi \sim improper\ uniform$$

$$\sigma_\nu \sim improper\ uniform, \tau \sim U(0,1).$$

The only informative prior is the $\beta_1$since there is a vast literature on the effect of nitrogen on yield. Still, this prior also considers an extensive range concerning the values gained for this effect on many papers (Alotaibi, et al., 2018; Boyer, et al., 2013; Vetsch and Randall, 2004).

We use four chains with 5000 iterations and 2000 warmups in each Bayesian estimation process. Since the HMC is more efficient (Girolami and Calderhead, 2011; Ng'ombe and Lambert, 2021) than the usual MCMC methods, this method is used to fitting the models on simulated data. The convergence Gelman-Rubin criterion $\hat{R}$ (Gelman and Rubin, 1992) and trace plots are checked to confirm the convergence of the Bayesian method.

5. Optimal Experimental Design

The next step for finding the optimal design is to analyze the simulated data and consider a curve over the possible conducted designs. Figure 4.2 shows three field proportions' (15%, 30%, 70%) empirical cumulative density function (ECDF). Based on Figure 4.2, it is clear that the 15% experimentation dominates the two others in first-order stochastic dominance (FSD). Hence, we do not need to conduct experiments over a significant part of the field to maximize the farmers' profit. In other words, increasing the experimentation could not provide enough information to cover the cost of these designs.

The site-specific optimal nitrogen values and the locations of the experimentation plots change from years 1 to 8 since the number of experiments reduce over time. Figures 4.3 to 4.6 illustrate the NPV versus the nitrogen levels for 15 plots of field experimentations. The labels across the borders indicate the number of replications for each point. Based on Figures 4.3 to 4.6, we can select the optimal nitrogen values and replications for each level. The optimal nitrogen values for making the experimentation are 35 with 2 plots replication, 130 with 3 plots replication, 165 with 5 plots replication, and 230 with 5 plots replication. Two values of the nitrogen selected at the beginning and the end of linear part, one level is in the plateau change point domain, and one nitrogen valu at the plateau.

Figure 4.7 depicts the farmers' eight-year NPV when the best designs are selected for experimentation. There is always a significant surge in the first year of experimenting since the prior knowledge of the site-specific optimal nitrogen is very off (optimal average is 160.46 and the farmers optimal is 300). Figure 4.7 shows farmers' NPV maximizes at

the third year of experimentation. The experimentation can be quit after this year if the design is selected wisely by the researcher in the first place.

Figure 4.8 depicts the actual (left) and the average estimated (right) site-specific optimal nitrogen values for the optimal selected design. This Figure demonstrates that the average optimal design coverage to the actual optimal values. Table 4.1 shows the mean squared error (MSE) value for 8 consecutive years.

Table 4.2 shows the Bayesian estimation of the parameters for model 1 based on the third year of simulation for one of the best-selected experimental designs. For all these parameters, the model converges nicely. The results show that the estimated parameters and the actual values are close when the profit maximizing design is selected.

6. Conclusion

There is no optimal experimenting system for on-farm experimentation, which motivated this paper. This research assumes an LP model with year stochastic behavior and spatial plateau. A fully Bayesian decision approach is used. We consider three different scenarios for the proportion of fields on which to make the experimentation. Forty different designs for each scenario are simulated with a random selection of nitrogen levels and replication for eight years. The number of experimentations is reduced by ten percent for the following years to cover the cost of losing yield for the extensive experiments.

Results show that experimenting on 15 plots of the field FSD dominated the two scenarios that experimented on more of the field. The optimal levels of nitrogen for

experimenting are 35, 130, 165, and 230, with corresponding replications of 2, 3, 5, and 5 plots, respectively. The best year for quitting the experiment is the third year since the marginal revenue is almost zero adjusted for the year's random effect.

This study has several limitations. First, since the optimal percentage happened in the lowest value of the experiment plots (boundary point), the results may vary if a lower proportion was considered. Another limitation is the reduction in the experiment percentage for the following years. We consider this value fixed. Continued experimentation might have been profitable if the number of plots had decreased more quickly.

7. References

Alotaibi, K.D., A.N. Cambouris, M. St. Luce, N. Ziadi, and N. Tremblay. (2018). "Economic Optimum Nitrogen Fertilizer Rate and Residual Soil Nitrate as Influenced by Soil Texture in Corn Production." *Agronomy Journal* 110:2233-2242.

Boyer, C.N., J.A. Larson, R.K. Roberts, A.T. McClure, D.D. Tyler, and V. Zhou. (2013). "Stochastic Corn Yield Response Functions to Nitrogen for Corn after Corn, Corn after Cotton, and Corn after Soybeans." *Journal of Agricultural and Applied Economics* 45:669-681.

Bullock, D., and T. Mieno. (2017). "An Assessment of the Value of Information from on-Farm Field Trials." *Unpublished Working Paper, University of Illinois, Champaign, IL*.

Bullock, D.S., M. Boerngen, H. Tao, B. Maxwell, J.D. Luck, L. Shiratsuchi, L. Puntel, and N.F. Martin. (2019). "The Data-Intensive Farm Management Project: Changing Agronomic Research through on-Farm Precision Experimentation." *Agronomy Journal* 111:2736-2746.

Carpenter, B., A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76(1):1-32.

Cho, W., D.M. Lambert, A. Fornah, and W.R. Raun. (2020). "Bayesian Estimation and Economic Analysis of under-Replicated Field Trials with a Linear Response Plateau Function." *Journal of Agricultural Science* 12:1-15.

Colyer, D., and E.M. Kroth. (1968). "Corn Yield and Economic Optima for Nitrogen Treatments and Plant Population over a Seven-Year Period 1." *Agronomy Journal* 60:524-529.

de Oliveira Ferreira, A., T.J.C. Amado, C.W. Rice, D.R.P. Gonçalves, and D.A. Ruiz Diaz. (2021). "Comparing on-Farm and Long-Term Research Experiments on Soil Carbon Recovery by Conservation Agriculture in Southern Brazil." *Land Degradation & Development* 32:3365-3376.

Dhakal, C., K. Lange, M.N. Parajulee, and E. Segarra. (2019). "Dynamic Optimization of Nitrogen in Plateau Cotton Yield Functions with Nitrogen Carryover Considerations." *Journal of Agricultural and Applied Economics* 51:385-401.

Evans, F.H., A. Recalde Salas, S. Rakshit, C.A. Scanlan, and S.E. Cook. (2020). "Assessment of the Use of Geographically Weighted Regression for Analysis of Large on-Farm Experiments and Implications for Practical Application." *Agronomy* 10:1720.

Gelman, A., and D.B. Rubin. (1992). "Inference from Iterative Simulation Using Multiple Sequences." *Statistical science*:457-472.

Girolami, M., and B. Calderhead. (2011). "Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73:123-214.

Hanumantha, R. (1965). "Agricultural Production Functions, Costs and Returns in India." *Agricultural production functions, costs and returns in India.*

Huelsen, W. (1932). "Efficiency Factors and Their Use in Determin-Ing Optimum Fertilizer Ratios." *Journal of Agricultural Research* 45.

Lacoste, M., S. Cook, M. McNee, D. Gale, J. Ingram, V. Bellon-Maurel, T. MacMillan, R. Sylvester-Bradley, D. Kindred, and R. Bramley. (2022). "On-Farm Experimentation to Transform Global Agriculture." *Nature Food* 3:11-18.

Lambert, D.M., and W. Cho. (2022). "Geographically Weighted Regression Estimation of the Linear Response and Plateau Function." *Precision Agriculture* 23:377-399.

Matson, P.A., R. Naylor, and I. Ortiz-Monasterio. (1998). "Integration of Environmental, Agronomic, and Economic Aspects of Fertilizer Management." *Science* 280:112-115

Moeltner, K., A.F. Ramsey, and C.L. Neill. (2021). "Bayesian Kinked Regression with Unobserved Thresholds: An Application to the Von Liebig Hypothesis." *American Journal of Agricultural Economics* 103:1832-1856.

Muller, P. (1999) "Simulation Based Optimal Design, Bayesian Statistics 6." *In Proceedings of the Sixth Valencia International Meeting 6–10 June 1998* (eds J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith), pp. 459–474. Oxford University Press, Oxford, UK.

Müller, P. (2005). "Simulation Based Optimal Design." *Handbook of Statistics* 25:509-518.

Ng'ombe, J.N., and D.M. Lambert. (2021). "Using Hamiltonian Monte Carlo Via Stan to Estimate Crop Input Response Functions with Stochastic Plateaus." *Journal of Agriculture and Food Research* 6:100226.

Ng'ombe, J.N., and B.W. Brorsen. (2022). "Bayesian Optimal Dynamic Sampling Procedures for on-Farm Field Experimentation." *Precision Agriculture*:1-23.

Overstall, A.M., and D.C. Woods. (2017). "Bayesian Design of Experiments Using Approximate Coordinate Exchange." *Technometrics* 59:458-470.

Paccioretti, P., C. Bruno, F. Gianinni Kurina, M. Córdoba, D. Bullock, and M. Balzarini. (2021). "Statistical Models of Yield in on-Farm Precision Experimentation." *Agronomy Journal* 113:4916-4929.

Park, E., W. Brorsen, and X. Li. (2018). "How to Use Yield Monitor Data to Determine Nitrogen Recommendations: Bayesian Kriging for Location Specific Parameter Estimates."

Poursina, D., and W. Brorsen. (2021). "Site-Specific Nitrogen Recommendation: Using Bayesian Kriging with Different Correlation Matrices."

Rodriguez, D.G.P. (2014). *Testing Two Existing Fertilizer Recommendation Algorithms: Stanford's 1.2 Rule for Corn and Site-Specific Nutrient Management for Irrigated Rice*: University of Illinois at Urbana-Champaign.

Ryan, E.G., C.C. Drovandi, and A.N. Pettitt. (2015). "Fully Bayesian Experimental Design for Pharmacokinetic Studies." *Entropy* 17:1063-1089.

___. (2015). "Simulation-Based Fully Bayesian Experimental Design for Mixed Effects Models." *Computational Statistics & Data Analysis* 92:26-39.

Sarkar, S., and F. Lupi. (2022). "Modelling Mid-Western Corn Yield Response to Phosphorus Fertilizer in Michigan."

Seeger, M., F. Steinke, and K. Tsuda (2007) "Bayesian Inference and Optimal Design in the Sparse Linear Model." In *Artificial Intelligence and Statistics.* PMLR, pp. 444-451.

Sellars, S.C., G.D. Schnitkey, and L.F. Gentry. (2020). "Do Illinois Farmers Follow University-Based Nitrogen Recommendations?" Paper presented at Agricultural and Applied Economics Association meetings.

Singh, I., and K. Sharma. (1968). "Response of Some Mexican Red and Indian Amber Wheats to Nitrogen." *Indian Journal of Agricultural Economics* 23:86-93.

Sun, F., and R.B. Gramacy, (2021), "Space-Filling Design under Maximin Distance." https://CRAN.R-project.org/package=maximin.

Tanaka, T.S., S. Kakimoto, T. Mieno, and D.S. Bullock (2022) "Comparison between Spatial Predictor Variables for Machine Learning in Site-Specific Yield Response Modeling Based on Simulation Study of on-Farm Precision Experimentation." In *Abstracts of Meeting of the CSSJ The 253rd Meeting of CSSJ. CROP SCIENCE SOCIETY OF JAPAN*, pp. 63-63.

Tembo, G., B.W. Brorsen, F.M. Epplin, and E. Tostão. (2008). "Crop Input Response Functions with Stochastic Plateaus." *American Journal of Agricultural Economics* 90:424-434.

Trevisan, R., D. Bullock, and N. Martin. (2021). "Spatial Variability of Crop Responses to Agronomic Inputs in on-Farm Precision Experimentation." *Precision Agriculture* 22:342-363.

US Department of Agriculture. (2019). *A Case for Rural Broadband: Insights on Rural Broadband Infrastructure and Next Generation Precision Agriculture Technologies.* Washington, DC: USDA.

Vetsch, J.A., and G.W. Randall. (2004). "Corn Production as Affected by Nitrogen Application Timing and Tillage." *Agronomy Journal* 96:502-509.

Villacis, A.H., A.F. Ramsey, J.A. Delgado, and J.R. Alwang. (2020). "Estimating Economically Optimal Levels of Nitrogen Fertilizer in No-Tillage Continuous Corn." *Journal of Agricultural and Applied Economics* 52:613-623.

Tables

Table 4.1. Mean Square Error for Optimal Nitrogen Value for Selected Optimal Design

Year	1	2	3	4	5	6	7	8
MSE	950.86	1062.33	688.77	1030.92	804.19	794.75	648.68	750.32

Table 4.2. Parameters and Their Estimates for One of the Best-Selected Designs in the Third Year

Parameters	True value	Estimate (SD)	Gelman- Rubin Statistics
β_0	105	102.71 (0.315)	0.998
β_1	0.7	0.71 (0.020)	0.998
$\bar{p}$	194	202.5 (0.600)	1.012
τ	0.003	0.003 (0.000)	1.000
ρ	0.5	0.46 (0.008)	1.001
σ_v	20	9.8 (3.456)	1.003
σ_ϕ	5	10.3 (3.871)	1.002
σ_ϵ	3	3.47 (0.115)	1.003

Figures

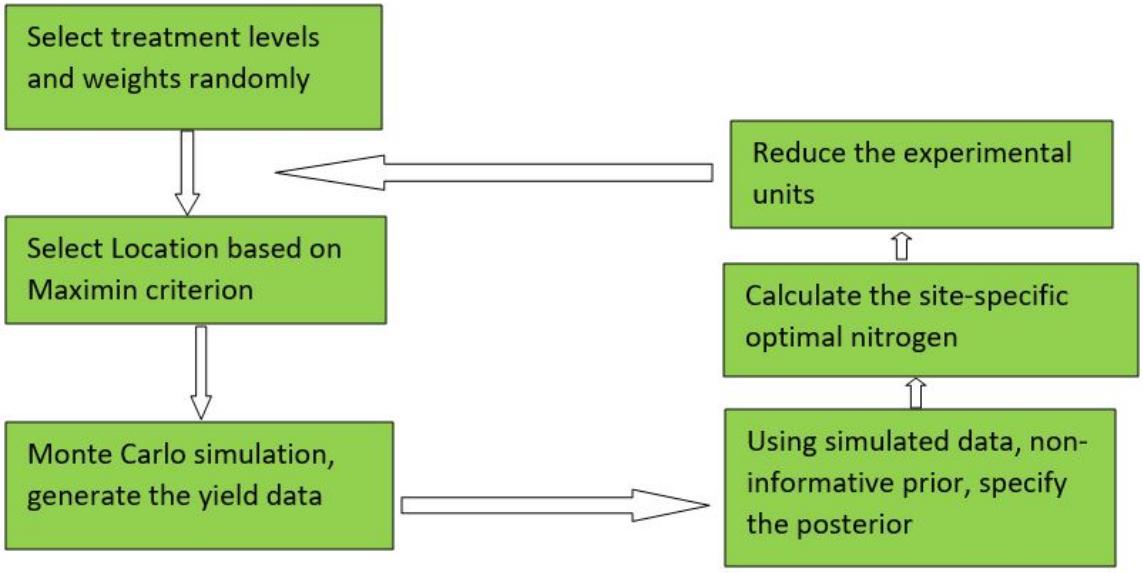

Figure 4.1. Flowchart for Simulation of One Farm

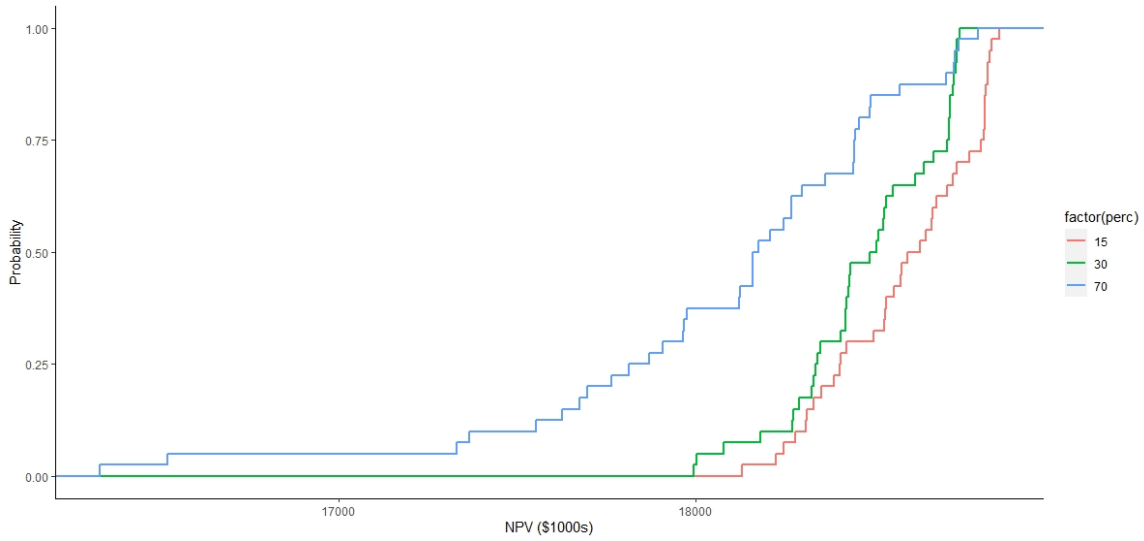

Figure 4.2. Empirical Cumulative Distribution Function for Farmers' Profit ($1000) Over Eight Consecutive Years Based on Field Experimentation Proportion

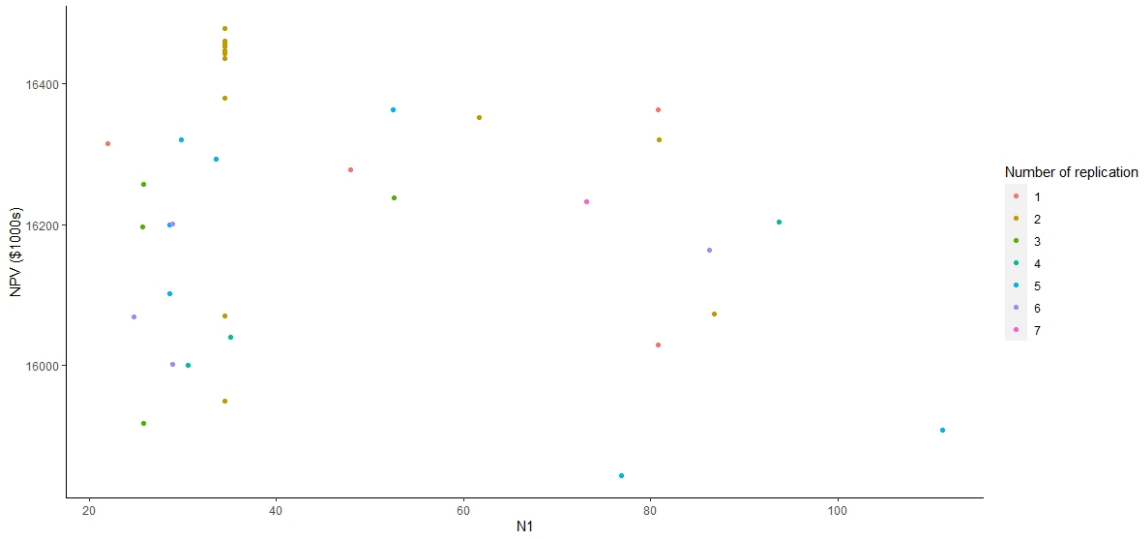

Figure 4.3. Total Farmers' Profit ($1000) Over Eight Consecutive Years vs. the First Level of Nitrogen

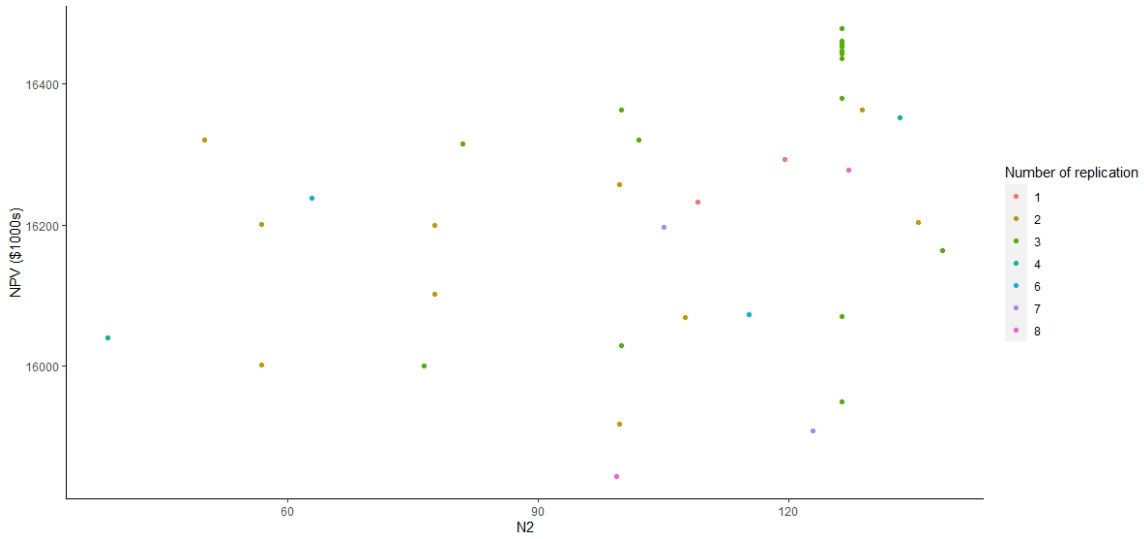

Figure 4.4. Total Farmers' Profit ($1000) Over Eight Consecutive Years vs. the Second Level of Nitrogen

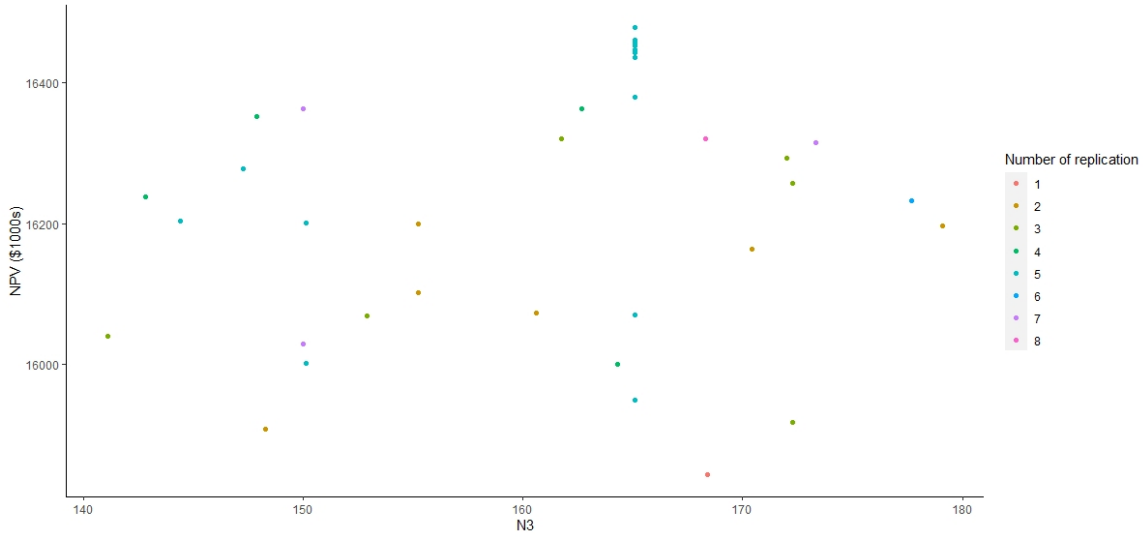

Figure 4.5. Total Farmers' Profit ($1000) Over Eight Consecutive Years vs. the Third Level of Nitrogen

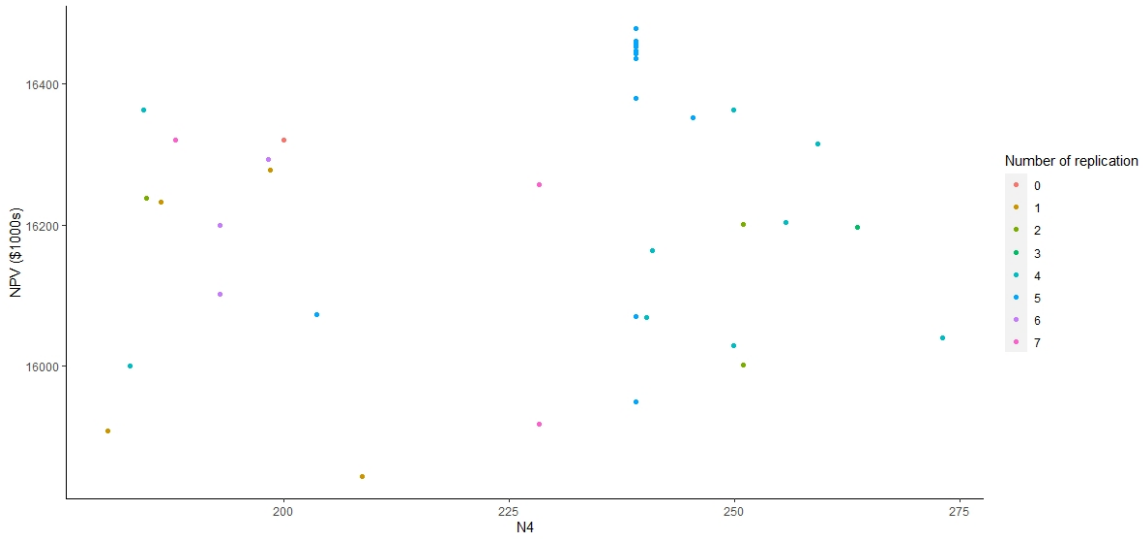

Figure 4.6. Total Farmers' Profit ($1000) Over Eight Consecutive Years vs. the Last Level of Nitrogen

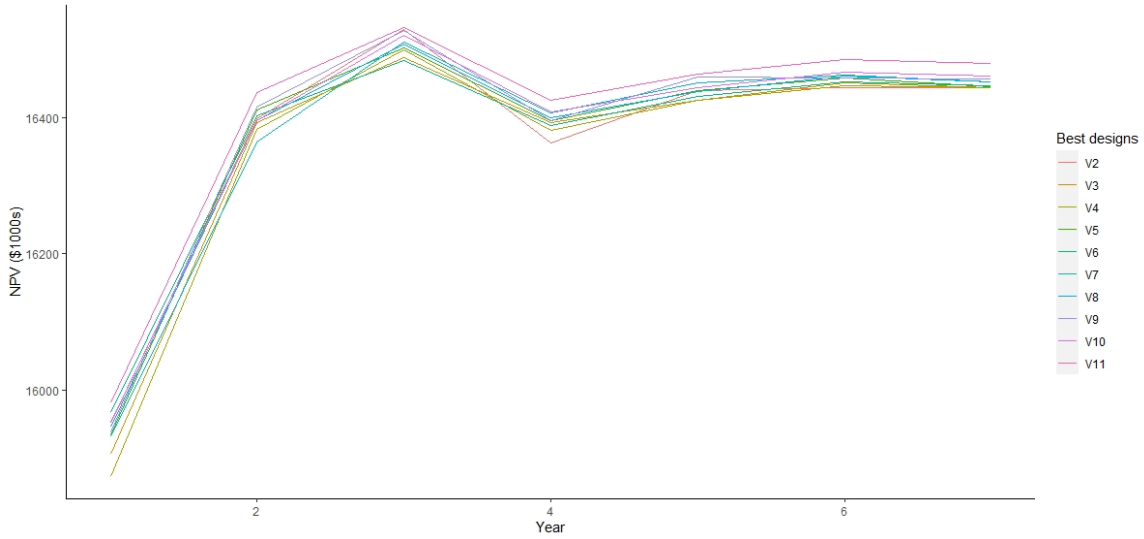

Figure 4.7. Farmers' NPV ($1000) vs. Year for Best-Selected Designs

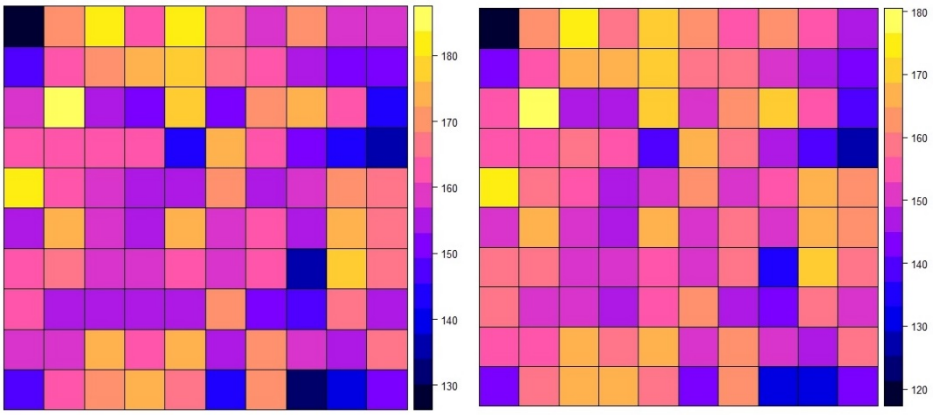

Figure 4.8. Actual (Left) and Estimated (Right) Optimal Nitrogen Values for the Profit Maximizing Design in the Third Year

VITA

Davood Poursina

Candidate for the Degree of

Doctor of Philosophy

Dissertation: WHOLE FARM EXPERIMENTATION: MAKING IT PROFITABLE

Major Field: Agricultural Economics

Biographical:

Education:
Completed the requirements for the Doctor of Philosophy in Agricultural Economics at Oklahoma State University, Stillwater, Oklahoma in December, 2022.
Completed the requirements for the Doctor of Philosophy in Statistics at University of Isfahan, Isfahan, Iran in September, 2014.

Completed the requirements for the Master of Science in Applied Statistics at University of Isfahan, Isfahan, Iran in September, 2007.

Completed the requirements for the Bachlor of Science in Statistics at Isfahan University of Technology, Isfahan, Iran in September, 2005.