Estimating genotype sensitivity according to location and year effects

by

Mriga Kher

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Industrial Engineering

Program of Study Committee:
Sigurdur Olafsson, Major Professor
Guiping Hu
Julie Dickerson

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2022

DEDICATION

This work is entirely dedicated to my family in India as well as in the USA and to my colleagues at Iowa State without whom this thesis paper would not have been possible.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS

I would like to thank my committee chair, Dr. Sigurdur Olafsson, and my committee members, Dr. Guiping Hu and Dr. Julie Dickerson, for their guidance and support throughout the course of this research. In addition, I would also like to thank my colleagues, the department faculty and staff for making my time at Iowa State University a wonderful experience.

ABSTRACT

In crop breeding, breeders aim to select crop varieties that consistently perform well in both stability and mean performance across target environments. A classic model to determine this stability is Finlay-Wilkinson (FW) two-stage regression model. Several modifications to this model have been proposed in the past that select for genotypes that have a minimum level of performance across all contexts while still responding well in favorable circumstances. Our research aims to select favorable crop varieties based on their sensitivity/interaction with a specific location and not based on stability across all the planted environments. Thus, we propose an approach like FW regression but incorporating a geostatistical method (kriging) accounting for the spatial interdependence of data points. The kriged estimate at the unknown location considers the standard deviation of the prediction error. This model seems to perform well inter ms of identifying varieties that are sensitive at certain locations. The new model is validated using a study that assesses the genotype sensitivity of the 28 late-stage soybean varieties planted in a mid-western experiment. Specifically, we assess the sensitivities of the genotype to the environment, using five models. The classic FW regression is used as a benchmark, as well as two straightforward variants that separate the environmental effects into Location-Environment (LE) effects and Location-Year (LY) effects. We compare those with our proposed Deviation-from-Expectation (DE) two-stage regression model, and the Kriging-DE that replaces the first-stage regression with a kriging approach. The study shows that the proposed DE and Kriging-DE models can discover significant interaction effects (sensitivities) between genotypes and the environmental effect.

CHAPTER 1. INTRODUCTION

Background

The world population is expected to rise to around nine million people by 2050 (Lenaerts et al., 2019). This immense increase in the population makes it imperative to address and find solutions to future food demand & agriculture. And to add on to the increasing population, deteriorating land fertility, and climate change have also caused a surge in the need for looking into alternate solutions for food security (Lenaerts et al., 2019).

One of the solutions to addressing this problem is improving quality and yield of the harvested crops via selective crop breeding. In selective crop-breeding, an important aspect to consider is the genotypic variation. It is the genetic variation which is essentially a variation in biochemical markers in cells, that along with the environmental effects that is responsible for determining the phenotypic traits such as the crop yields and its tolerance to changes in the environment (van Dijk et al., 2021). Therefore, it is essential to understand effects of genotypic variation on phenotypic trait such as crop yield in order to gain insights and make crop yield predictions in the field of crop breeding. In the past few years, the agriculture industry, owing to its complex nature, has become part of the big data phenomenon. Decision-makers use data mining techniques on such agriculture datasets to extract valuable insights that are then used to achieve desired crop forecasting, in precision farming, smart agriculture, and in achieving high-quality seeds (Bose & Author, 2020).

A commonly used predictive model in the agricultural industry is regression. Regression is known to output a numerical estimation and is especially of use in predicting yields of crops (Bose & Author, 2020). Several regression models such as Finlay–Wilkinson model, Additive mean effects and multiplicative interaction analysis (AMMI) , Best linear unbiased prediction

(BLUP), Genotype plus GE interaction biplot analysis (GGE), Joint regression analysis (JRA), Yield–stability statistic (YS) have been used for estimating crop yields. The best model is then selected based on statistical parameters such as model coefficients, R^2, RMSE, SSE.

Motivation

In the field of crop breeding, genotype stability or even genotypic sensitivity in relation to environmental circumstances is a crucial factor. The Finlay-Wilkinson (FW) regression is a well-known method for analyzing this type of stability. The target variable is the observed yield for the certain genotype in that environment, and the explanatory variable is the mean phenotype (e.g., yield) for that environment (location-year). Although the slope of such a regression should always be positive, the magnitude of the slope is important since it reflects sensitivity to the environment. If the slope of this regression line is steep, the genotype may be sensitive to environmental quality, and vice versa if the slope has an insignificant value.

When utilizing prediction models such as the above-mentioned FW regression to estimate values at an unsampled location, error or uncertainty associated with the model cannot be avoided. Uncontrolled variables such as environment & also controlled factors such as soil fertilizer, and soil water content make agricultural yield mapping significantly vary in space.

Thus, it is important to consider spatial variability when estimating for crop yields in high uncertainty/error ranges in crop prediction models. Geostatistical models such as kriging consider the spatial interdependence of data points and the kriged estimate at the unsampled location by incorporating the standard deviation of the prediction error (Castoldi et al. 2009).

Our research aims to propose a complementary approach to the FW regression model that incorporates geostatistical modeling along with regression to estimate crop yield and its uncertainty at any given location using soybean crop yield dataset collected from the midwestern

region of the US. The estimated crop yield values will then be utilized by crop-breeders to understand the crop varieties that are sensitive to a location and that achieve stable or desirable yield at any given location, assuming negligible fluctuations in uncontrolled factors associated with weather conditions/climate change.

This method can be used when phenotype data is collected over a long period of time, which is likely in any breeding effort. It is not required to observe each genotype for several years; rather, some genotypes are observed for several years, and some places are planted for several years. This allows us to examine genotype plasticity in terms of location and year effects separately. We believe this is particularly significant from a producer's standpoint because a portion of the environmental effect is constant and well-known for each producer, hence it may be more vital for the producer to select genotypes that are more resistant to unpredictably changing circumstances each year (e.g., excessive rain during planting season, drought during the growing season, and too little or too much heat during critical stages).

In this thesis, we look at a method for separating the environmental effect into i) a location effect and ii) either a simple year-effect or a location-year effect. The idea is that a location effect (the average value for a certain location over a period of a year or not accounting for years altogether) could be estimated, in long-term environmental factors like soil type and temperature. While year effect could be preserved to be less predictable due to the variation in weather over the years.

This location-year effect could be viewed as the departure from the expected environmental influence each year. We would like to propose a method like the FW approach for estimating location and year effects, allowing the results to be easily comparable to the standard FW regression.

Previous Work

In their research, Sellam and Poovammal (2016) implement regression analysis is to find the relationship between the dependent variable which is the rice yield and ecological factors namely Annual Rainfall (AR), Area Under Cultivation (AUC), and Food Price Index (FPI). The results indicate a R^2 value of 0.7, showing any slight changes in the rainfall and in AUC to have a significant effect on the rice crop yield.

In their research, Gonzalez-Sanchez et al., 2014 utilize and compare linear and non-linear techniques to predict crop yields based on the best attribute subsets created for each of the techniques. The optimal subsets are determined using a repetitive or a recursive technique, by obtaining combinations of different features and then building a regression model on these obtained subsets to determine the prediction accuracy of the models on the subsets. The regression models utilized in this research were MLR (Multiple Linear regression), perception multilayer linear regression, stepwise linear regression, and ANN (Artificial Neural Networks) . On further assessing the models for their accuracies, accuracy metrics such as RMSE, RMAE, and R were used. The results indicate ANN to have obtained the best attribute subset. The results obtained in this paper cannot directly be obtained on a different crop-breeding dataset due its high dependent nature.

In the research conducted by Imran et al. (2015), researchers used GWRK (Geographically weighted regression kriging) to predict sorghum crop yields in Burkina Faso. The study then used kriging to interpolate the GWR residuals and geographically weighted regression (GWR) to model the local variation in the data. The accuracy of predicting sorghum crop yields and measuring the uncertainty of such projections were compared. MAE, MSE, and R^2, are used to assess the accuracy of crop yield prediction. The prediction error variances and

the RMSE acquired during cross validation of the model were used to assess the correctness of the uncertainty estimates. GWRK outperforms KEDLN and RK in terms of overall performance. In GWRK, the prediction uncertainty was lowered considerably. GWRK had a prediction error variance of 20, whereas RK had 31 and KEDLN had 39.

However, the above-mentioned research uses certain co-variates to build these prediction models and further analyze the success and limitations of each of these models based on the derived relationships between these co-variates and the crop yield predictions. For example, Sellam and Poovammal (2016) utilize variables such as Annual Rainfall (AR), Area Under Cultivation (AUC), and Food Price Index (FPI) to estimate rice yield, while Gonzalez-Sanchez et al., 2014 conduct optimal selection of attributes namely SP(location where crop was sowed), IWD (Irrigation Water Depth), MaxT(Maximal Temperature), MinT(Minimal Temperature), etc. Imran et al. (2015), in their research of using Geographically weighted regression kriging to estimate crop yields, build their model in two steps, where the first step builds either an MLR (Multi-Linear Regression) or a Linear Regression (LR) using external variables such as NDVI (Normalized Difference Vegetative Index), rainfall, population density, poverty head count ratio to establish the correlation between the independent variables and the sorghum yield. Using the grid cell values of the external variables at those sites, the coefficients of these regression models were used to predict the yield in unvisited locations. There hasn't been much research conducted in-terms of analyzing and assessing regression models that solely use variables such "Location", "Year", and a combined location-Year effect to estimate crop yield.

In this thesis, we aim to build models that have the same structure as the FW regression, but that primarily utilize the above-mentioned variables, or a combination of the variables estimate the crop yields and then utilize the models to find sensitivities of certain varieties in

certain locations. In the following write-up, we will explain the premises of the two primary models (FW Regression & Kriging) to illustrate their general structures and their relevance to the four proposed models in chapter 2.

FW Regression Model

A well-known regression model in the field of crop-breeding is FW regression, used to analyze stability of different crop varieties planted in different environments (Lian & de los Campos, 2016). FW regression aims to estimate how the expected performance of the crop (yield) varies as a function of environment effect E, which is known to be the main trait effect for a particular environment. For example, an. Environmental effect could be the mean yield of all the crop varieties planted in that environment (Gauch Jr. & Zobel, 1997).

The primary model for averaged value of genotype *i* in environment *j* is given by

$$Y_{ij} = \mu + G_i + E_j(1 + \beta_i) + \delta_{ij} \quad \text{(i)}$$

β_i gives response of genotype *i* to the changing environment, $Y_{ij}$is the measured or the observed yield of the i^{th} genotype at j^{th} location/environment, μ is overall mean of the ith genotype, δ_{ij} represents the mean error associated with observed Y_{ij}

Kriging Model

The Kriging algorithm utilizes the concept of autocorrelation. Correlation refers to the tendency for two types of variables to be associated with each other. This is the fundamental in the field of geo-statistics: objects that are closer together are more similar than those that are

farther apart (Legendre, 1993). For autocorrelation in Geo-statistics, all the spatial locations are assumed to have some sort of relationship with each other and are used to calculate distances between the spatial observations and to build an autocorrelation model as a function of distance (*ArcGIS Desktop Help 9.2 - Understanding Different Kriging Models*, n.d.).

The premise of Kriging model is that it is highly dependent on spatial model (also known as the variogram) between datapoints to make predictions at unknown locations. Kriging is known to consider the spatial structure (direction & magnitude) of the given datapoints by comparing spatial distances of two datapoints at a time. This is done to plot a variogram and understand how datapoints are in relation with each other in terms of different "lag distances". Once the variogram has been plotted, a kriging model calculates spatial weights for the observed datapoints (Imran et al., 2015).

Understanding the Kriging Model

Kriging consists of a two-step process

1. Building a spatial autocorrelation model known as variogram for estimating statistical dependence.
2. Predicting value at an unknown locationVariogram

The primary purpose of building a variogram is to give a close autocorrelation structure of the measured points. The semi-variogram function $\gamma(d)$ is defined as half of the averaged square of differences between measured points separated by distance d (*ArcGIS Desktop Help 9.2 - Understanding Different Kriging Models*, n.d.)

$$\gamma(d) = \frac{1}{2|N(d)|}\sum\left(z_i - z_j\right)^2 \quad \text{(ii)}$$

Where $N(d)$ the number of all the pairwise Euclidian distances d (Magnitude only) at measured locations i and j respectively, and z_i and z_j are the observed values at those locations. After calculating the differenced square between these paired locations, these differences are then binned into lag bins.

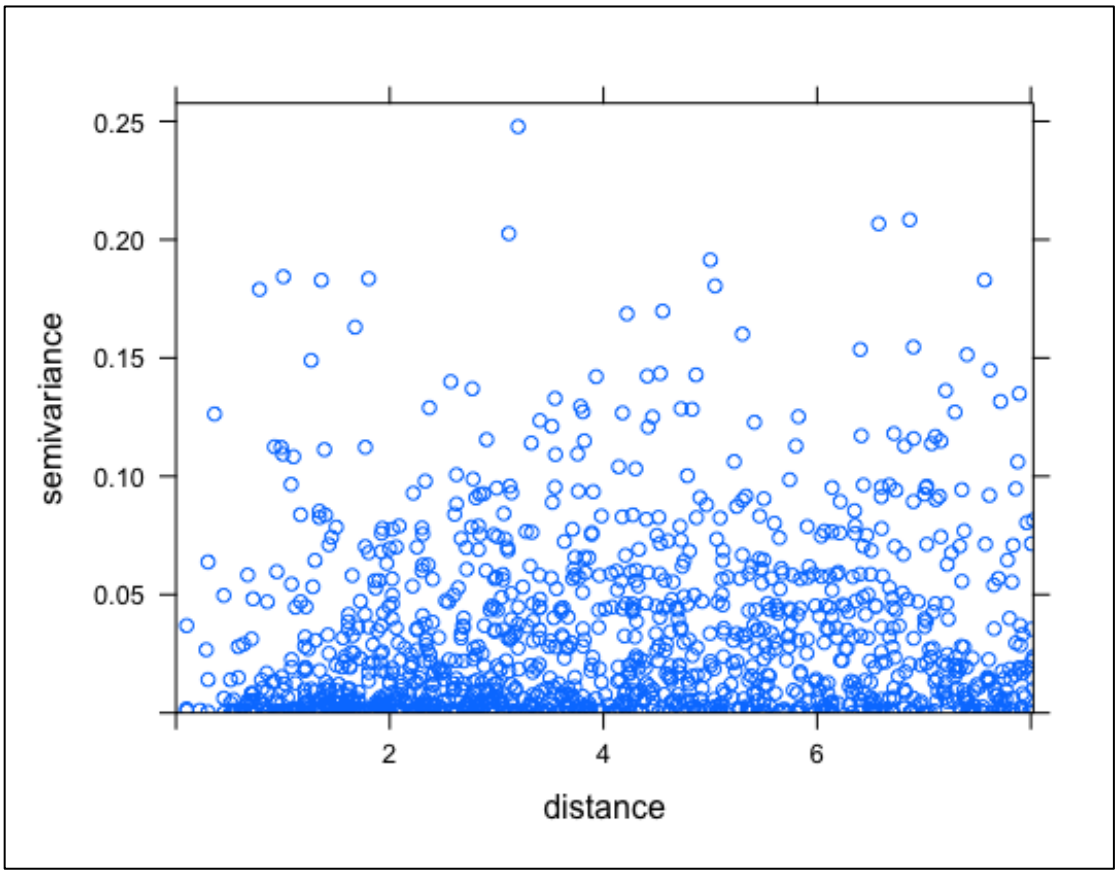

Figure 1.1. Illustrates each blue dot in the diagram (for year 2018 with 58 datapoints) represents the empirical semi variogram value plotted against the distance between

As the pair of locations become further away from each other, they have a higher $\gamma(d)$ value and become more dissimilar than the pairs on the left of x-axis. The next step is to fit a model that is a continues function to the semi variogram, this model is what influences the estimation of values at unknown locations, i.e., the shape of the curve near the origin will

influence the effect the closest neighboring points will have on the prediction (*How Kriging Works—Help | ArcGIS for Desktop*, n.d.).

The fitted model is now used to estimate value at the desired location, kriging assigns weights of the measured values at the neighboring locations to predict values at unknown locations.

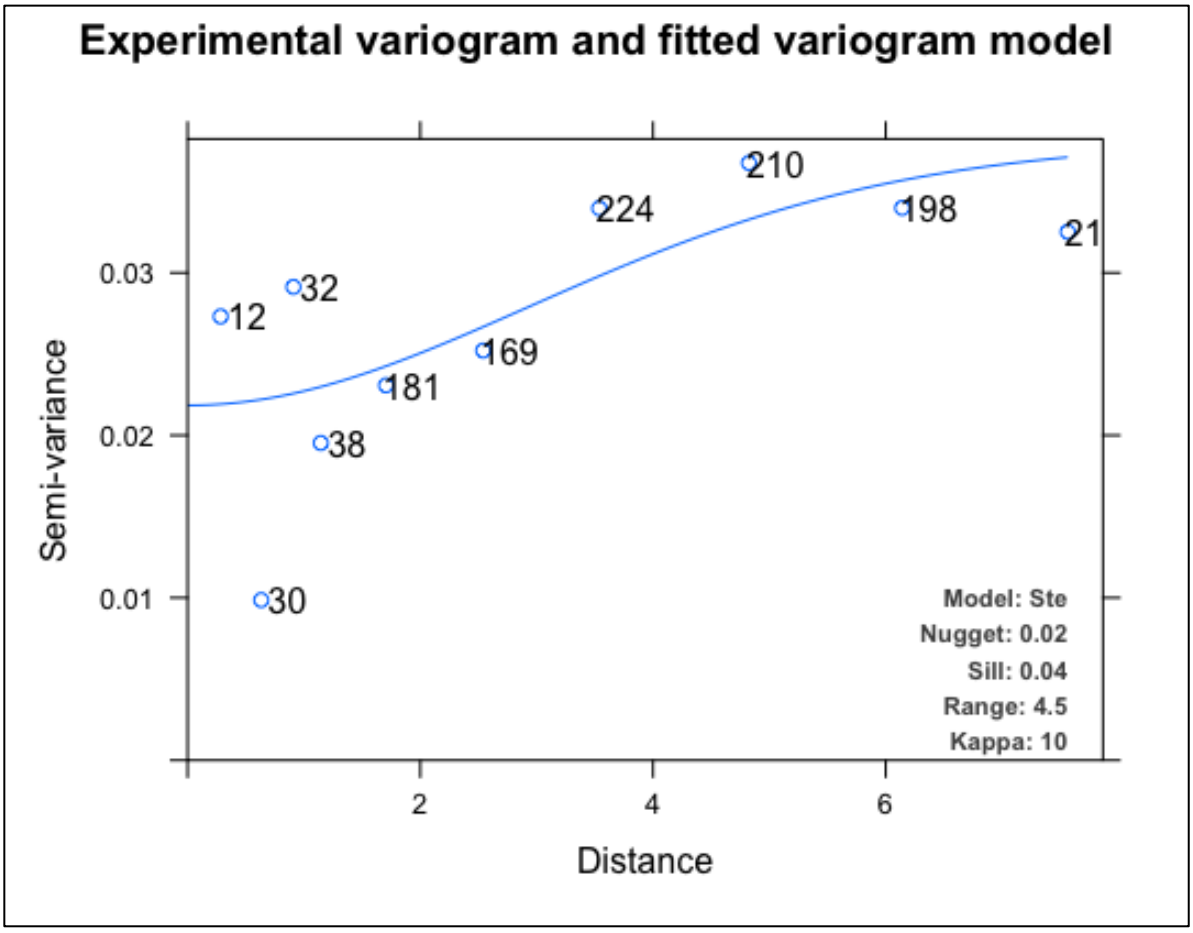

Figure 1.2. Illustrates a fitted variogram (for year 2018 with 58 datapoints) with the number of pairs that went into calculating each variogram lag

CHAPTER 2. METHODOLOGY

Let y_{ij} denote the phenotypic response of genotype i in environment j. The classic approach is to consider this response as a linear model, combining the genotype main effect g_i, the environment main effect h_j, and an interaction between the two. We will follow this convention and assume a linear model

$$y_{ij} = \mu + g_i + h_j + b_i h_j + \varepsilon_{ij} \quad (1)$$

Here $e_{ij} \sim N(0,1)$ is as usual, a normally distributed error term.

The Finlay-Wilkinson Model

We will introduce standard Finlay Wilkinson Regression as a benchmark model to compare our newly proposed approach with. Classic Finlay-Wilkinson, which uses a two-stage approach to estimate a single sensitivity to environmental quality for each genotype. This approach does not distinguish between year and location effects, so we consider two direct extensions that separate those effects in two diverse ways. For consistency we stay with the two-stage approach, where we first estimate environmental effects (or separate location/year effects), and then estimate the sensitivity to the environment.

Procedure:

The standard Finlay-Wilkinson (FW) approach is a two-step process, where in the first step we estimate the environmental effect of h_j and in the second step we estimate the slope b_i of each genotype for the environmental factor.

Step 1. First estimate the environmental effect from a simple main-effect model:

$$y_{ij} = \mu + g_i + h_j + \varepsilon_{ij} \tag{2}$$

Step 2. Substituting the estimate $\hat{h}_j$ into the main model results in

$$y_{ij} = \mu + g_i + \hat{h}_j + b_i\hat{h}_j + \varepsilon_{ij} \tag{3}$$

And the desired slopes b_i for each genotype are estimated from this model.

Note that for a fixed genotype, the genetic main effect is constant, so we could write $\tilde{\mu} = \mu + g_i$ as the single regression model constant, which makes it more transparent that for a specific genotype this is a simple linear regression equation that finds the sensitivity of the phenotype to the environmental effect, namely:

$$\tilde{y_j} = \tilde{\mu} + (1 + b)\hat{h}_j + \varepsilon_j \tag{4}$$

Thus, this model tells us how sensitive the phenotype is to the environmental.

Location-Environment Model (LE)

If the phenotype data is observed over a single year, then the environmental effect in the FW regression is simply due to location. However, if the data is observed over multiple years, then this effect is due to some combination of location (*l*) and year (*k)*. We are interested in understanding the environmental effect as the combination of the location effect and the location-

year effect, which can be different in each location; that is, $h_{lk} = h_l^1 + h_{lk}^2$. What this does is to separate out the effect of environmental factors that depend only on the location, namely h_l^1, versus those that depend on various fluctuations from year-to-year, namely h_{lk}^2 (of course, those could *also* vary by location).

We now have a base model

$$y_{ilk} = \mu + g_i + h_{lk} + b_i^1 h_l^1 + b_i^2 h_{lk}^2 + \varepsilon_{ilk} \quad (5)$$

$$h_{lk} = h_l^1 + h_{lk}^2$$

We need to estimate both h_l^1 and h_{lk}^2 . Note that the number of h_j^2 terms is the same as the main environmental effect in the standard model, whereas there will generally be much fewer h_j^1 terms, depending on the relative number of locations versus years. In terms of additional estimation effort, the difference between the proposed model and the classic FW model is simply the estimation of the h_j^1 terms. Our proposed procedure is like the FW approach but requires a three-step process to first estimate the h_j^1 terms.

Procedure:

Step 1: We estimate h_j^1 by treating all the environments that include the same location (but multiple years) as a single environment, and then estimate h_j^1 from a linear no-interaction model.

$$y_{ij} = \mu + g_i + h^1{}_j + \varepsilon_{ij} \quad (6)$$

In this model we have genotype and **location** j' as predictor variables, but do not use year.

Step 2. We estimate h_j^2 from the linear model

$$y_{ij} = \mu + g_i + \hat{h}^1{}_j + h^2{}_j + \varepsilon_{ij} \quad (7)$$

In this model we use genotype and **environment** as predictor variables, namely we have combined location-year to define an environment j.

Step 3. Finally, we substitute $\hat{h}^1{}_j$ and $h^2{}_j$ into the original model, and estimate both the slopes from the interaction model

$$y_{ij} = \mu + g_i + (1 + b_i^1)\hat{h}^1{}_j + (1 + b_i^2)\hat{h}^2{}_j + \varepsilon_{ij} \quad (8)$$

We will refer to this as the Location-Environment (LE) model and will compare its use to the standard FW model, as well as one more variant described below.

Location-Year (LY) model

A simpler model would assume that there is a location effect and a year effect and that these effects are independent. Thus, we have some location effect h_j^1 as before (does not depend on year), and a year effect h_j^2 that does not depend on the location. Like before, we have a model:

$$y_{ij} = \mu + g_i + (1 + b_i^1)h^1{}_j + (1 + b_i^2)h^2{}_j + \varepsilon_{ij} \quad (9)$$

In Step 1 we can estimate the location and year effects simultaneously. Since the location and year effects are assumed, independent there is no reason to estimate them separately, and Step 2 is identical to Step 3 in the previous section, that is, both slopes are estimated from the interaction model. We will refer to this as the Location-Year (LY) model and will compare its use to that of the LE model described above and the classic FW model.

New sensitivity Model: Deviation from Expected Environment Factor

We argued in the introduction that it might be of particular interest to growers to understand the stability of a genotype to yearly fluctuations, namely: how will a genotype perform if a specific location severely underperforms (or overperforms) its expected performance. Some of this might be indirectly observed from the LY and LE models, but in this section, we propose a model that addresses this question directly, that is, for each genotype it estimates a slope of its phenotypic response with respect to the deviations (positive or negative) from the expected environment effect.

We start by estimating the environmental effects from a simple linear model, using only location and year:

$$h_{lk} = \mu + {h^1}_l + {h^2}_{lk} + \varepsilon_{lk} \quad (10)$$

Using the estimated environmental effects, we then define a model where the genotype-environmental interactions of interest are a genotype's reaction to deviation from the expected environmental performance:

$$y_{ilk} = \mu + g_i + h_{lk} + b_i\left(h_{lk} - \hat{h}_{lk}\right) + \varepsilon_{ilk} \tag{11}$$

The estimation procedure involves three steps:

Procedure

Step 1: Estimate the expected performance of an environment from the linear model:

$$h_{lk} = \mu + h_l^1 + h_{lk}^2 + \varepsilon_{lk} \tag{12}$$

Step 2: Calculate the difference between observed and expected performanced in each environment, where the expected performance $\hat{h}_{lk}$ is found in Step 1 above.

$$d_{lk} = h_{lk} - \hat{h}_{lk} \tag{13}$$

Step 3: Estimate the slope from the interaction model

$$y_{ilk} = \mu + h_{lk} + b_i d_{lk} + \varepsilon_{ilk} \tag{14}$$

Kriging-DE Model: Deviation from Expected Environment Factor

In this section, we would like to propose an extension of the DE model that incorporates Kriging. The Kriging-DE model is almost the same as the DE model except for how it estimates average yield at a particular location utilizing Kriging.

Procedure:

Step 1: Create separate datasets for each of the years. Each year would consist of unique locations. In a certain year, calculate average yield for all the varieties planted at a particular location for that particular year. Thus, each year's dataset would consist of unique locations with average yield values. Combine these subset datasets and run an auto-kriging model using leave one out cross-validation to predict yield values at each of the location $\widehat{h_{lk}}$.

Step 2: Calculate the difference between observed and expected performanced in each environment, where the expected performance $\widehat{h_{lk}}$ is found in Step 1 above.

$d_{lk} = h_{lk} - \widehat{h_{lk}}$ is calculated as for the DE model in equation 13

Step 3: Estimate the slope from the interaction model (same as eqaution 14)

$$y_{ilk} = \mu + h_{lk} + b_i d_{lk} + \varepsilon_{ilk}$$

We will refer to this model as Kriging-DE model and will compare it to the LY, LE and FW models.

CHAPTER 3. DATA

The data for this study came from a commercial soybean breeding program. In the program, we begin with a single late-stage experiment focusing on the midwestern experiment. There are 28 soybean varieties in this experiment, all of which were planted in the same year in the same 27 locations. All the types had been planted in the previous two years, but not in the same locations. The varieties are compared to one another (and to other types) to see which ones should be developed further and eventually commercialized. Breeders are usually only interested in the varieties giving good yields. Table 1 displays a ranking of the top 5 kinds out of 28 (ranking based on yield = bushel per acre). We would like to advance the best varieties, but other factors to consider include the yield's sensitivity to environmental quality, which will be discussed in Section 3 below.

Table 3.1. The average yield of the top varieties in the first experiment (Midwest)

Variety	Yield
V126389	65.42
V14243	65.25
V44771	65.14
V126793	64.37
V75340	64.24

The regression models incorporate extra variables in addition to using the data from the first experiment directly. We identified all areas where these varieties had been planted in any year given the set of varieties (see Figure 3 for a map of the approximate locations).

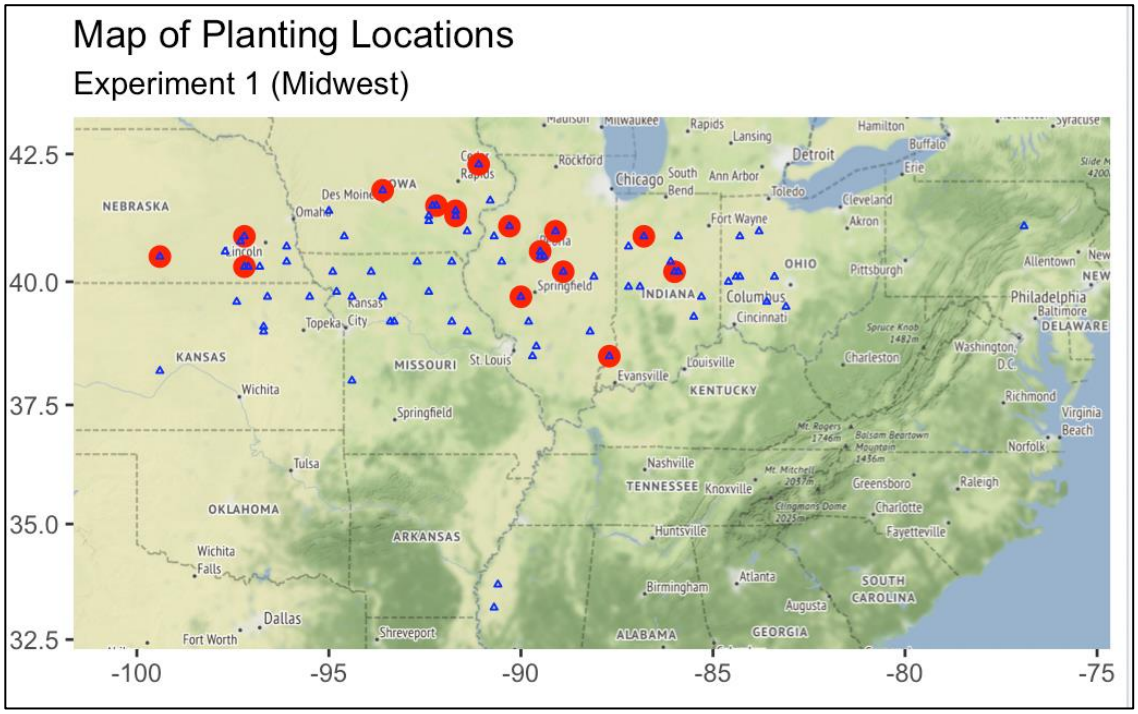

Figure 3.1.Shows map of the 8 planting locations studied in the second experiment (red dots), as well as the 25 locations used by the regression models, where the target types were planted (blue diamonds)

Then we retrieve all the data that was collected at these locations including observations of additional types and years in which the initial varieties were not planted, which helps in the estimation of environmental effects (both overall and location and year effects). Table 2 summarizes the features of the data used by the regression models. The characteristics of the initial experiment, namely the field experiment that is planted in order to compare, are also shown in this table. We should note that the regression model allows us to include observations of other varieties from different years and locations, resulting in over one million observations used by the regression models to estimate primarily the environmental effects, which are then useful for estimating the stability of these specific 28 varieties.

Table 3.2.shows the characteristics of the the original field experiment, and the sampled data

	Field Experiment	**Original Data**	**Sampled Data**
Observations	1,549	1,030,826	12,098
Varieties	28	177,039	8,463
Locations	27	76	76
Years	1	10	10
Environment	27	279	269

Since the dataset can be difficult to work with, we built a sampled dataset (Table 1). To create this sampled data, first, take 10,000 random samples from the original data, then add any observations from the experimental varieties that were not included in the sample. The sampling is stratified, with 1000 samples collected for each year between 2009 and 2018. As a result, there are 12,092 observations of 8,463 varieties in the dataset. The sampled dataset includes all the original locations, bringing the total number of environments to 269 throughout a 10-year period. There are 1000 observations for each year from 2009 to 2015, 1119 for 2016, 1279 for 2017, and 2694 for 2018. The number of observations in each location ranges from three to sixteen hundred and thirty-nine, with a mean of 159 observations.

CHAPTER 4. RESULTS AND DISCUSSION

In this section we analyze and compare the above-mentioned regression models with our novel approach on the mid-west dataset generated from the commercial breeding program. The results are then used to run two separate analyses demonstrating two distinct applications of the new proposed model. The first analysis can be used to identify soyabean varieties that are insensitive to environmental quality using the Finlay-Wilkinson approach but are in fact sensitive to deviation from expected performance in each location. In the second analysis we observe genotypes where the oppositive is true, that is, a classic analysis indicates genotypes that are highly sensitive to environmental quality, but our new model demonstrates how these varieties yield as expected in each location. In other words, we aim to identify stable yielding varieties particular to each of the locations.

Regression Model Comparison

We start by evaluating the fit of different regression models, including Finlay-Wilkinson (FW), Location-Environment (LE), Location-Year (LY) and Deviation-from-Expectation (DE) model, and Kriging-DE model (Kriging-DE). The primary goal of our analysis is to understand phenotypic plasticity, or the sensitivity of the performance of each genotype in a particular location with respect to some changes pertaining to the environment.

It is thus the slopes that are introduced in the four models described in Section 2 that are of the primary interest, but it is also important to evaluate how well do the 4 models fit the data, to use the optimal model, which will then be utilized to evaluate sensitives of crop-varieties in certain locations. The quality of that fit is compared in the Table below. FW Regression, DE, and Kriging-DE provide the closest fit.

Table 4.1.Comparison of model fitness

	FW	LE	LY	DE	Kriging-DE
Residual SE	6.226	6.998	7.101	6.300	6.278
Multiple R2	0.7511	0.6831	0.6793	0.7561	0.7574
Adj R2	0.7434	0.6697	0.6662	0.7376	0.7390
F-Statistic	112.9	50.87	51.79	40.84	41.14
Degrees of Freedom for SE	2057	1959	2029	1963	1963

Sensitivity Comparison

In this section we evaluate the genotype plasticity or stability of the 28 late-stage soybean varieties. As, mentioned in the data section, these 28 soybean varieties in the mid-western experiment were planted in the same year in the same 27 locations. We do this evaluation utilizing each of the four models i.e., the Finlay-Wilkinson (FW), Location-Environment (LE), Location-Year (LY), Deviation-from-Expectation (DE) and kriging-DE model for analyzing the sensitivity of the genotype to the environment. This evaluation also helps understand limitations and benefits of each of these models better.

The new regression models were primarily built to increase our ability to find genotypes (soybean varieties) whose phenotype (yield) is sensitive to year-to-year changes in a specific location. The ability of these models to discover significant interaction effects between genotypes and some form of year-effect may thus be used to assess their success.

For the LE model, this is the slope with respect to the location-year effect given the independent location effect; for the LY model is the slope with respect to the year effect; and for

the DE model it is the slope with respect to the deviation from expected yield in an environment. Table 4 summarized the number of significant effects found by each approach.

All the models have a main genotypic (G) effect. Each of the models also have different variants of how the environment for a particular model is accounted, i.e., as the environment in its entirety (E), the location (L) and year (Y) separately, or the location-year given an independent location effect (E|L).

To reinstate the goal of study, our main goal is to detect sensitivities to year-by-year changes for a certain location. As indicated by these results, the LE, DE & Kriging-DE models do well in-terms of identifying such sensitivities.

Table 4.2.Shows significant effects for the top 5 varieties for each of the models

		Sensitivity (slope)						
		FW-Model	**LE-Model**		**LY-Model**		**DE-Model**	**Kriging-DE**
Variety	**Yield**	**Env**	**Loc**	**Env**	**Loc**	**Year**	**Dev**	**Dev**
V126389	65.42	0.04	0.01	-0.10	0.04	-0.07	0.25	0.345
V14243	65.25	0.08	0.10	-0.29**	0.03	-0.13	0.79**	0.91*
V44771	65.14	0.15	0.15	-0.11	0.17	0.36	0.58	0.68.
V126793	64.37	0.16	0.18	-0.11	0.17	0.33	0.72*	0.90*
V75340	64.24	0.04	0.06	-0.16	0.04	0.20	0.38	0.41
**Significant at 0.1 level; **significant at 0.05 level; . significant at 0.1 level*								

For plant breeders, the top varieties are of main interest since they will usually have the potential to be advanced in breeding programs to become commercial varieties. Thus, we take look at the top five yielding varieties in the above table, FW model have slopes closer to zero illustrating that all the five varieties are unaffected to environment.

However, the other models consist of separate location and year effects indicating sensitivities of the models. For example, the LE model finds that there is a significant environmental interaction for V14243 when the location has already been accounted for and we estimate the location-year effects given the location effects. Thus, it can be said that there is sensitivity associated with the weather that each year experienced in each location. Thus, V14243 variety is not suitable as its yield performance is less stable year-to-year in each specific location. LY model, like FW regression is also unable to deduce any significant slopes. As described in Section 2 above, LY model assumes independent location and year effects, which explains the reason why it cannot find a significant year effect for V14243.

Finally, the DE & Kriging-DE models find that both V14243 and V126793 have a significant sensitivity to the deviation of the mean yield in a planting location from its expected mean yield (given a linear model).

CHAPTER 5. CONCLUSION

DE as well as Kriging-DE seem to show the desired sensitives and better fits, but to further understand the usefulness of the models to the crop breeders & farmers, let us consider an example crop variety. Variety V14243 has the second highest yield in this experiment, which certainly indicates that this is a high-performing variety; but on the other hand, it has a significant slope (sensitivity) with respect to deviation from expected mean yield in the planting location, which could make it less desirable, if were to look for varieties that are stable from year-to-year. On the one hand, all 28 varieties are planted in the same locations (and the same year), so the yield comparison should be fair; but on the other hand, the GxE interactions together with a limited set of observed locations make it difficult to be certain of that claim. Specifically, if the locations planted were particularly favorable this year, then we know that a variety would perform well in that location. Thus, although Kriging-DE model has proved its usefulness in terms of fit and in finding desired varieties for a particular location, we still need to conduct planting experiments that generate location datasets in-order to assess the model further for its reproducibility.

REFERENCES

1. *ArcGIS Desktop Help 9.2—Understanding different kriging models*. (n.d.). Retrieved March 12, 2022, from http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Understanding_different_kriging_models

2. Bose, D. & Author. (2020). *Big data analytics in Agriculture*.

3. Bostan, P. A., Heuvelink, G. B. M., & Akyurek, S. Z. (2012). Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey. *International Journal of Applied Earth Observation and Geoinformation*, *19*, 115–126. https://doi.org/10.1016/j.jag.2012.04.010

4. Colombari Filho, J. M., de Resende, M. D. V., de Morais, O. P., de Castro, A. P., Guimarães, É. P., Pereira, J. A., Utumi, M. M., & Breseghello, F. (2013). Upland rice breeding in Brazil: A simultaneous genotypic evaluation of stability, adaptability and grain yield. *Euphytica*, *192*(1), 117–129. https://doi.org/10.1007/s10681-013-0922-2

5. Fan, X.-M., Kang, M. S., Chen, H., Zhang, Y., Tan, J., & Xu, C. (2007). Yield Stability of Maize Hybrids Evaluated in Multi-Environment Trials in Yunnan, China. *Agronomy Journal*, *99*(1), 220–228. https://doi.org/10.2134/agronj2006.0144

6. Finlay, K., & Wilkinson, G. (1963). The analysis of adaptation in a plant-breeding programme. *Australian Journal of Agricultural Research*, *14*(6), 742. https://doi.org/10.1071/AR9630742

7. Gauch Jr., Hugh. G., & Zobel, R. W. (1997). Identifying Mega-Environments and Targeting Genotypes. *Crop Science*, *37*(2), cropsci1997.0011183X003700020002x. https://doi.org/10.2135/cropsci1997.0011183X003700020002x

8. *Genotype by Environment Interaction for Grain Yield and Association among Stability Parameters in Bread Wheat (Triticum aestivum L.)*. (n.d.). Retrieved March 23, 2022, from https://www.scirp.org/(S(lz5mqp453edsnp55rrgjct55.))/journal/paperinformation.aspx?paperid=97550

9. Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. *The Scientific World Journal*, *2014*, 1.

10. *How Kriging works—Help | ArcGIS for Desktop*. (n.d.). Retrieved March 13, 2022, from https://desktop.arcgis.com/en/arcmap/10.4/tools/3d-analyst-toolbox/how-kriging-works.htm#ESRI_SECTION1_C1D6C1E74B8F468D94CAA30F68199CDC

11. Imran, M., Stein, A., & Zurita-Milla, R. (2015a). Using geographically weighted regression kriging for crop yield mapping in West Africa. *International Journal of Geographical Information Science*, *29*(2), 234–257. https://doi.org/10.1080/13658816.2014.959522

12. Imran, M., Stein, A., & Zurita-Milla, R. (2015b). Using geographically weighted regression kriging for crop yield mapping in West Africa. *International Journal of Geographical Information Science*, *29*(2), 234–257. https://doi.org/10.1080/13658816.2014.959522

13. Kang, M. S. (1993). Simultaneous Selection for Yield and Stability in Crop Performance Trials: Consequences for Growers. *Agronomy Journal*, *85*(3), 754–757. https://doi.org/10.2134/agronj1993.00021962008500030042x

14. Kusmec, A., Srinivasan, S., Nettleton, D., & Schnable, P. S. (2017). Distinct Genetic Architectures for Phenotype Means and Plasticities in Zea mays. *Nature Plants*, *3*(9), 715–723. https://doi.org/10.1038/s41477-017-0007-7

15. Legendre, P. (1993). Spatial Autocorrelation: Trouble or New Paradigm? *Ecology*, *74*(6), 1659–1673. https://doi.org/10.2307/1939924

16. Lenaerts, B., Collard, B. C. Y., & Demont, M. (2019). Review: Improving global food security through accelerated plant breeding. *Plant Science*, *287*, 110207. https://doi.org/10.1016/j.plantsci.2019.110207

17. Lian, L., & de los Campos, G. (2016). FW: An R Package for Finlay–Wilkinson Regression that Incorporates Genomic/Pedigree Information and Covariance Structures Between Environments. *G3 Genes|Genomes|Genetics*, *6*(3), 589–597. https://doi.org/10.1534/g3.115.026328

18. Nbsp, V. S. and E. P. (2016). Prediction of Crop Yield using Regression Analysis. *Indian Journal of Science and Technology*, *9*(38), 1–5. https://doi.org/10.17485/ijst/2016/v9i38/91714

19. van Dijk, A. D. J., Kootstra, G., Kruijer, W., & de Ridder, D. (2021). Machine learning in plant science and plant breeding. *IScience*, *24*(1), 101890. https://doi.org/10.1016/j.isci.2020.101890

20. *Variogram.pdf*. (n.d.). Retrieved March 12, 2022, from http://faculty.washington.edu/edford/Variogram.pdf