# Estimating Site-Specific Crop Yield Response using Varying Coefficient Models

**Xiaofei Li**
**Department of Agricultural Economics**
**Mississippi State University**
**xiaofei.li@msstate.edu**

**Keith H. Coble**
**Department of Agricultural Economics**
**Mississippi State University**
**coble@agecon.msstate.edu**

**Jesse B. Tack**
**Department of Agricultural Economics**
**Mississippi State University**
**tack@agecon.msstate.edu**

**Barry J. Barnett**
**Department of Agricultural Economics**
**Mississippi State University**
**barnett@agecon.msstate.edu**

***Selected Paper prepared for presentation at the 2016 Agricultural & Applied Economics Association Annual Meeting, Boston, Massachusetts, July 31-August 2***

**Estimating Site-Specific Crop Yield Response using Varying Coefficient Models**


## Abstract

This study estimates the site-specific crop yield response function using varying coefficient models. It is widely recognized that the parameters of yield response function vary dramatically across space and over time. Previous studies usually capture this variability of response by using locational and time dummy variables. While that approach reveals the existence of the response variability, the exact pattern of the variability is unknown, and the capacity of *ex ante* prediction of such models are limited. This study takes a step forward to explicitly explain how the response varies with the actual site characteristic variables, such as soil, water, topography, weather, and other factors that are commonly available to producers. By using the varying coefficient model, the parameters of the response function are specified to change continuously with those site variables. Based on a simulation data set, the varying coefficient model is demonstrate to outperform the site-dummy model by creating better variable rate application (VRA) fertilizer prescriptions. We further propose to apply the model to large sample of high resolution production data, and create *ex ante* spatially explicit optimal VRA fertilizer recommendations. The ultimate goal is to develop a precision decision system which can statistically turn the soil testing and weather forecasting information into input application prescriptions for producers.



Keywords: Site-specific crop response, varying coefficient regression, *ex ante* prediction, precision agriculture, big data

**Estimating Site-Specific Crop Yield Response using Varying Coefficient Models**

## 1. Introduction

Precision agriculture is attracting more attention in recent years for its economic and environmental benefits. Advancing GIS technologies, computers, and farming equipment have developed the techniques and data for applying precision agriculture. However, a major barrier for the wide adoption of precision agriculture is obtaining the precision prescriptions of optimality of input applications.

The correct estimation of site-specific yield response to inputs is crucial for the success of generating optimal precision prescriptions. Due to the considerable heterogeneity in soil and weather, the optimal input rates vary both spatially and temporally. The identification of localized optimal input quantities (and thus profits) across heterogeneous within-field locations depends crucially on the ability to credibly estimate site-specific production relationships, or yield response functions. This is complicated both by the complexities of the physiological processes underlying crop growth and the lack of available data at such a fine scale.

In general, there are mainly two types of approach to estimate the yield response functions. The first one is the simulation models. They are mechanistic models which simulate the physiological process of crop growth such as photosynthesis, transpiration, nutrient update, etc. Those type of models have been widely used in agronomy, physiology, soil science, and agricultural engineering. Examples include the "School of de Wit" crop growth model (de Wit, 1978; Penning de Vries et al., 1989), DSSAT model (Tsuji et al., 1994), Hybrid-Maize (Yang et al., 2004), APEX model (Williams and Izaurralde, 2006), and Adapt-N (Melkonian et al., 2008). The fertilizer recommendation rates in many farming management support systems are based on simulation methods. Despite the popularity, due to the complexities of crop growth's interactions with various conditions, parameterization is a big issue. Those simulation models are usually too complicated to accurately estimate their parameters (Spitters, 1990). Many parameters are estimated based on strict assumptions, or even the modelers' subjective opinions. The model predicting error ranges are usually large. Therefore, those simulation models may serve better for the purpose of explanation rather than prediction (ten Berge et al, 1997).

The second approach is statistical regression. It works by empirically fitting yield data to input levels using single or multiple regressions. Regression models have been widely used in analyzing the agronomic experimental data. In particular, in recent years a growing body of literature has applied spatial econometric models to field-trial experimental data to empirically estimate site-specific crop response functions (SSCRFs). Studies include Anselin et al. (2004), Hurley et al. (2004), Lambert et al. (2004), Liu et al. (2006), Ruffo et al. (2006), and Lambert et al. (2006). The most common approach is to divide a field into sub-blocks or "sites", and then use discrete variables (i.e. block-specific dummies) to capture spatial variations of the response. Similarly, temporal variations in response are captured through interaction with year dummies. These approaches are useful for explaining yield variation and capturing ad-hoc forms of coefficient variation. However, the arbitrarily defined site dummy variables do little to help us better understand the causal influences of such variation. The relation between the response variability and site characteristics (soil, water, weather, etc) is not explicitly revealed. Consequently, the models are only capable of providing *ex post* evaluation of the same field's profit potential from variable rate application technology (VRA), while its capacity of *ex ante* prediction for other fields under various other agronomic and climatic conditions is very limited.

This study builds upon the existing literature of site-specific crop yield response functions, and takes a step forward in attempting to explicitly model the variation of yield response with respect to actual site-specific characteristics. As a starting point, we focus on estimating corn yield response to nitrogen (N) fertilizer. Following most studies, we specify a baseline nonlinear (quadratic) relationship. But instead of interacting response parameters with site-dummy variables, we allow the parameters to vary in a smooth and continuous manner across alternative site-specific soil, topography, and weather characteristics. Based on the estimated yield response function, the ultimate goal is to derive *ex ante* site-specific input quantity recommendations that achieves profit maximization in each specific location within field.

This study proposes to use observational producer data to estimate the comprehensive yield response functions. We start from collecting a cross-sectional data in 2014 from a group of corn producers spanning many locations in the United States. Observed variables include within-field

GIS data on seed rates, fertilizer application, and yield from producers' production records. Those production data are then matched to the public soil and weather databases such as the USDA Gridded Soil Survey Geographic (gSSURGO) 10-meter resolution maps and the 4-km resolution PRISM Climate Data. Our research data differ from most previous studies typically using on-farm experimental data. There are two advantages of producer data: (1) It has a much larger sample size. The total number of observed field spots are in millions or even billions. That can support the estimation of more comprehensive response functions with a large set of variables. (2) There is a rich amount of variation across inputs applications as well as agro-climatic conditions. It allows the simultaneous estimation of all kind of factors in impacting yield, which is normally impossible in experimental data where all other factors are controlled the same and only allow one factor to vary. Despite the downsides of non-random application rates and less accurate measurements, producer data better serve the purpose of estimating our comprehensive yield response functions.

The output of the estimation is a site-specific yield response function whose coefficients vary explicitly with soil and weather factors commonly available to producers (soil fertility and texture, elevation, slope, water, temperature, *etc.*). Based on the function, a spatially explicit profit surface can be derived after combining price information, and the optimal inputs rates in each specific site can be predicted by maximizing profit. That is, our estimation makes possible *ex ante* site-specific optimal inputs application maps, which can be predicted universally for all fields as long as the soil testing and weather information are available. Our study provides an approach to integrate the farming data as well as other weather and geography information into a precision decision supporting system for producers to achieve economically optimum fertilizer application in crop production practices.

## 2. Literature review

Agricultural economists have a long history of estimating output response to input applications (Bullock et al., 2002). In recent years, there has seen a growing body of empirical regression studies on the site-specific crop yield response functions. Those studies typically use spatial econometric models and high resolution yield data, and their major focuses are usually to

examine the spatial and temporal instability of the response functions, as well as to estimate the potential profitability of site-specific inputs application.

Mamo, Malzer, Mulla, Huggins and J. Strock (2003) uses sub-blocked experimental data to analyze the corn response to N rate with both spatial and temporal variations, and find the site-specific N response shows some temporal instability across the sample years of 1997 and 1999. Anselin, Bongiovanni and Lowenberg-DeBoer (2004) estimate the site-specific yield response functions of corn production to nitrogen (N) application in Argentina. Hurley, Oishi, and Malzer (2005) estimate the variable rate nitrogen applications using spatial autoregressive error model and geostatistical model, and find the results for the two models differ notably for the 1995 Southern Minnesota data. They further find the model performances are location specific. Lambert, Lowenberg-DeBoer and Malzer (2006) examine the spatial and temporal stability of corn and soybean response to nitrogen (N) and phosphorus (P) using a five-year corn-soybean rotation experiments in Minnesota over 1997—2000. Liu, Swinton, and Miller (2006) estimate the site-specific response of corn to nitrogen based on field trial data in Calhoun and Hillsdale counties, Michigan, during 1999–2001, and find the site-specific response function is unstable in dryland fields, while more stable in irrigated fields. Ruffo, Bollero, Bullock and Bullock (2006) estimate the site-specific corn production function for nitrogen fertilization in eight experimental fields in Central Illinois, and find soil and terrain attributes affect the responses. Bongiovanni, Robledo and Lambert (2007) analyze the site-specific response of wheat yield and grain quality to nitrogen in the semiarid region of Argentina. Tremblay et al. (2012) find the corn response to nitrogen is influenced by soil texture and weather by reviewing across 51 N rate treatment studies in North America during 2006 to 2009. The above list of site-specific response studies is not complete, and more studies are being conducted in this literature.

### 2.1 *Methodologies*

The basic research methodology in the site-specific response literature is to use spatial econometric models with a quadratic regression, and based on on-farm experimental data. The frequently used model specification is the quadratic functional form like:

$$Y_{ij} = \alpha_i + \beta_i N_{ij} + \gamma_i N_{ij}^{\,2} + \varepsilon_{ij},$$

where $Y$ is the crop yield, and $N$ denotes the nitrogen fertilizer application rate. The subscript $i$ denotes the "site" or "zone" or location within a field, and subscript $j$ denotes the spot within a site. Note that this specification allows intercepts and slopes to vary across space, where the parameters ($\alpha_i, \beta_i, \gamma_i$) are corresponding to specific site $D_i$. The field is divided into different sites, denoted by dummies ($D_1, D_2, \cdots D_m$). It assumes the crop growing conditions are homogenous within a site.

The quadratic function is theoretically appealing because they are interpretable as second-order approximations of any crop response function to inputs. Hurley et al. (2004) provide a detailed derivation of this feature based on Taylor series expansion. From an economic perspective, the quadratic function is also intuitive as it captures the diminishing marginal return of inputs. Mathematically, the concave functional form also allows a closed-form solution to optimal input rates. Those features promote the popularity of quadratic functional form in the regression specification of response function. But some also challenge the misspecification of quadratic form (Tumusiime et al., 2011). Some alternative functional forms such as linear plateau, quadratic plateau, Spillman-Mitscherlich, etc., are also considered (Liu et al., 2006; Lambert et al., 2006; Tumusiime et al., 2011). But the quadratic function is clearly more computational tractable, and is widely adopted by most of recent work estimating site-specific crop response functions.

Spatial econometric models are commonly used due to the existence of spatial correlated data. Early studies tend to use the classical ordinary least squares (OLS) approach to estimate site-specific response functions (for example, Mamo et al. (2003)). However, OLS assumes observations are independent in space. In the real world, the agronomic plots and on-farm experiments data are usually spatially correlated. In that case, OLS produces inefficient estimates, mainly due to the incorrect variance estimates, and consequently wrong statistical significance, inference and prediction (Anselin et al., 2004; Lambert et al., 2004). To account for the spatial correlation in the data, spatial econometric models are more appropriate and gain popularity in more recent studies. They are believed to outperform the OLS model by generating better fit of the data and more correct variance estimates.

Particularly, the predominantly used model is the Spatial Error model. As raised by Anselin et al. (2004), the source of spatial correlation in the yield data is from the unobservable micro-climate and subsoil characteristics that drive yield variation. Hurley et al. (2004) provide a mathematical demonstration of error spatial autocorrelation from the general production function. In addition, since heteroscedasticity is also prevalent in the field data, the most appropriate is proposed to be the spatial autoregressive error model with heteroscedasticity. Hurley et al. (2004) suggests a model that include soil, spatial and treatment strip heteroscedasticity and correlation.

Beyond the spatial econometric approach, other spatial statistical methods are also employed in analyzing response functions. Lambert et al. (2004) compare four different spatial regression models. Hurley et al. (2005) compare spatial econometric and geostatistical methods. It is found that the non-spatial and spatial models usually lead to different estimate results. While some studies find the variant spatial models results are generally similar, some find notable differences.

### 2.2 *Experimental data*

The estimations of site-specific response functions are mostly based on field experimental data. The experiments are usually conducted in one or several farms, using the classical agronomic experimental layout called randomized complete block design. The application of fertilizer is usually by long strip (for example, in Hurley et al. (2005), it is the strip of 274 meters long and 4.6 meters wide). The application rate of fertilizer (often called "treatment") within each strip is uniform. Several adjunct strips are grouped together as a "replication". The treatment levels are usually fixed, but the order of the six levels (strips) is random within each replication. The entire field is then comprised of several replications. The spatial units of the yield observations are grids divided from the strips. The size of the grids differ in different studies, such as the 15.2 × 4.6 meters grids (Hurley et al., 2005), 7 × 7 meters grids (Anselin et al., 2004), 15 × 15 meters grids (Lambert et al., 2006), and so forth.

One important advantage of using experimental data for regression analysis is that the explanatory variable (fertilizer rate) is completely exogenous. It rules out the endogeneity issue

that harasses most empirical economic analysis, and leads to clear causality relation from the regression. However, the sample size of experimental data is usually quite small due to the high costs, with at most hundreds of observations. Also, the experiments normally only allow for one or several factors to vary while all other factors are carefully controlled. Due to those limitations, the model built on experimental data cannot support a simultaneous estimate of a large set of factors and their interactions.

2.3 *Spatial heterogeneity of responses*

The major interest in site-specific response function studies is to examine how the yield response to fertilizer varies with site-specific characteristics over space, mainly due to interactions with different soil, topography, and water conditions. In the modeling, this spatial heterogeneity is mainly shown by the stability of the coefficients of response function. The general idea of testing the stability of coefficients is to interact the coefficients with the site variables.

Most studies use locational dummy variables to denote the sites, or simply divide the field into latticed sub-blocks or sites. Anselin et al. (2004) and Lambert et al (2004) divide the sample field into four "regimes" based on landscape position (Low East, Slope East, Hilltop, and Slope West). Hurley et al. (2004) divide the experimental field into 6, 48 or 102 lattice "sites". Lambert et al. (2006) partitioned the field into 69 sub-blocks. Only a few studies include actual site characteristic variables. For example, Liu et al. (2006) use organic matter, cation exchange capacity, leaching, water availability and sunlight reception. Ruffo et al. (2006) interact with terrain attributes. Mamo et al., (2003) interact with different soil types. But the list of site variables differ notably in different studies. A summary of the delineations of the "site" for the site-specific response studies are described in Table 1.

Most studies find the response varies across the sites. However, though the site dummy approach allows for modeling spatial variation of response function without knowledge of site characteristics (such as soil, topography, water, or any other agronomic variables which significantly affect the growth process of crops), on the other hand it only tells whether there is response heterogeneity over space, while the exact heterogeneity pattern is unclear.

2.4 *Temporal heterogeneity of responses*

In addition to heterogeneity over space, it is also well documented that the response function varies over time, mainly due to weather variations across growing seasons. Several efforts have been made using multiple years' data to estimate the temporal instability of response function. The typical modeling strategy is to interact with year dummies, or to estimate separately for different years and compare the coefficients.

Mamo et al. (2003) estimate site-specific response functions for three separate years, 1995, 1997, and 1999, and find significant temporal variations. Liu et al. (2006) expand the site-specific nitrogen response study from single-year to over-time analysis, mainly focusing on how water affects the optimal nitrogen rates. They estimate separately for years 1999, 2000, and 2001, and find the yield response functions vary dramatically across years, mostly due to weather and rainfall variations. Lambert et al. (2006) use 5-year corn-soybean rotation data to estimating response functions separately for each year from 1997 to 2001, and find that response of corn and soybean to P is temporally stable in some parts of the field, but unstable in other parts, while response of N was not temporally stable. The temporal heterogeneity is stronger in dryland fields. In irrigated fields, however, the response functions are found to be more stable across years (Liu et al., 2006).

Though the underlying reason for the temporal heterogeneity of response is the variation of weather (precipitation, temperature, wind, etc) across years, estimations that directly include the actual weather variables are rare. The year dummy approach suggests the response is instable across years, but the explicit form of instability with weather conditions is not revealed.

2.5 *Profitability of site-specific application*

The ultimate goal of estimating site-specific response function is to evaluate the profitability of site-specific application of inputs. Almost every study of site-specific response functions estimation also reports the final profitability estimate. But the estimated profitability results are dramatically different across studies, as well as across estimation models within a same study.

Most studies report net gains in return of VRA against uniform rate application strategy (VRA costs are mainly dependent on authors' choices). For example, Anselin *et al.* (2004) estimate the VRA profit gain against optimal uniform rate application is \$1.3 $ha^{-1}$ for OLS models and 3.5 $ha^{-1}$ for spatial error models. Hurley *et al.* (2004) find VRA could have increased profit by \$14.5 $ha^{-1}$ at one location of experiment, and \$48.3 $ha^{-1}$ at another location. Using the same data but different estimation models, Hurley *et al.* (2005) find the VRA profit gain is around \$8 to \$ 12 $ha^{-1}$. Other estimates of profit gain of VRA against uniform rate application include: \$3 to \$7 $ha^{-1}$ (Lambert et al., 2004), \$28 $ha^{-1}$ (Lambert et al., 2006), \$8 and \$23 $ha^{-1}$ (Mamo et al., 2003), and so on. But on the other hand, Liu *et al*. (2006) find the VRA gain only ranges from \$0.1 to \$3 $ha^{-1}$ for most estimation models they used, which is usually not sufficient to cover the VRA management cost.

Note that we should interpret those estimated dollar values of VRA with caution. First, those profitability calculations are all heavily depend on the crop prices chosen, which differ significantly across the authors. Thus, those numbers are not directly comparable in this sense. In that sense, it appears more desirable to compare yield instead of profit in the VRA gain estimation as that avoids the influence of crop price fluctuations. Second, the site-specific characteristics used in the response functions differ substantially across studies. Especially, the location dummies in the regressions vary in number, size and the way of definition.

2.6 *ex-post assessment*

A major drawback of the existing site-specific response estimates is that, they only demonstrate the existence of variability in response over space or time, but do not explicitly study the interaction of the response with site characteristics, and therefore provide little information on the exact pattern of the variability. Consequently, those estimation results essentially only provide assessment of site-specific applications from an *ex post* perspective, while they are unable to provide *ex ante* predictions of site-specific input rates for new fields or the same fields in different years. This substantially undermines the prediction ability of those models in real production practice.

To enable the prediction of *ex ante* site-specific application maps, the actual measurable site variables need to be explicitly incorporated in the modeling. Those include soil, topography, and weather characteristics which impact the crop growth and fertilizers' effect. This study is an exploratory effort toward that goal.

**3. Models**

This study focuses on estimating corn yield response to nitrogen (N) fertilizer. The response for other crops and to other inputs can be estimated in a similar manner. Following the literature of site-specific yield response estimate, we specify a quadratic response of yield to nitrogen fertilizer as the baseline model. Instead of interacting the coefficients of the response with the sub-block or site dummy variables which are defined on a lattice basis (or some other geographical basis), we allow for the coefficients to vary explicitly across measurable site-specific soil, topography and weather characteristics in a continuous manner. Formally, we employ the varying coefficient regression model:

$$Y_i = \alpha(X_i) + \beta(X_i)N_i + \gamma(X_i)N_i^2 + \varepsilon_i,$$

where $i$ denotes the site of the observation, which is usually a small gridded land area within a field. $Y_i$ is the yield in site $i$, $N_i$ is the nitrogen input rate in site $i$, and $X_i$ contains a set of actual site characteristics. The site variables in $X_i$ are those factors that are commonly raised in agronomy that either directly affect yield or interact with nitrogen's effect, and also factors which can be relatively easily monitored by producers in practice. The candidate variables include soil texture, soil water capacity, soil organic matter, pH, elevation, slope, precipitation, and temperature.

A fundamental feature of the varying coefficient model is that the coefficients ($\alpha, \beta, \gamma$) are no longer constant. Instead, they are changing as transition functions of site characteristics $X_i$. The choice of functional forms for $\alpha()$, $\beta()$ and $\gamma()$ highly depends on agronomic and physiologic knowledge. Given the complexity of the plant growth process and its interactions with soil and environment, it remains an open question as to what the appropriate functional forms for $\alpha()$, $\beta()$ and $\gamma()$. A large effort is required to test and compare several different candidate transition functions spanning both parametric and nonparametric alternatives. As a starting point, we use

the regime functions which divide observations into discrete groups based on the values of site variables. Note that this regime function is not strictly speaking continuous. But it can delineate a rough shape for the transition functions, which provides useful information for further imposing the appropriate parametric functional structures or for the refined nonparametric smoothing.

The response function takes a quadratic form, which is mathematically easy to estimate. However, the quadratic function may not be the best fit for yield response of corn production. Alternative response functional forms such as various plateau functions can be considered for comparison.

The yield data display strong spatial correlation, which suggests the incorporation of spatial effects in the modeling. The source of spatial correlation is mainly from the underlying soil characteristics that are unobserved in the model. For that reason, the spatial error model is appropriate (Anselin et al., 2004). Given the highly soil heterogeneity within a field, that is, the soil characteristics often vary dramatically in a short distance, an immediate 1st order autoregressive error structure is desirable for the spatial modeling. In addition, the errors are also highly likely to be heteroscedastic, and it motivates us to further correct for the heteroscedasticity in the spatial error model.

The ultimate goal for producers is to obtain profit maximization in the crop production. Given that the site-specific yield response function is correctly estimated, it is relatively straightforward to solve for the profit-maximizing nitrogen input rate for each specific site within the field. After combining the corn output price ($p$) and fertilizer cost ($r$), the spatially explicit profit maximization problem is:

$$\begin{aligned} \underset{N_i}{Max}\ \pi_i &= p\hat{Y}_i - rN_i \\ &= p[\hat{\alpha}(X_i) + \hat{\beta}(X_i)N_i + \hat{\gamma}(X_i)N_i^2] - rN_i, \end{aligned}$$

where $\pi_i$ denotes the profit at a specific site $i$, and the coefficients estimates $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$ are from the yield response model listed above. The optimal nitrogen rate for site $i$ is calculated as:

$$N_i^* = \frac{r - p\hat{\beta}(X_i)}{2p\hat{\gamma}(X_i)},$$

which depends on the values of site variables $X_i$. It provides a variable rate application (VRA) prescription of nitrogen fertilizer $N_i^*$. Based on that, the profitability gain of VRA can be further estimated.

## 4. Data

### *4.1 Production data*

Ideally, the data we need is a large comprehensive field experiments, with adequate variabilities in inputs and agro-climatic conditions. However, in reality this experiment is impossible.

In this study, we use actual production data for the estimate. The sources of data are from producers' production records of yield and input applications. Currently, many farming machines and software packages can automatically generate high resolution yield maps from yield monitors. The operations of input applications such as fertilizer inputs and seeding rates are also well digitalized. An example of the spatial explicit sample production data in one field may look like in Figure 1.

As the advancing of farming equipment and management software, those types of data are becoming more available. Many are stored in producers' computer hard drives, online clouds, or collected by agricultural companies (consulting, equipment makers, and supplier). As can be easily seen, one field could have thousands of observations of the spots, while if we can put together the data there will be billions or even trillions of observations. That is a very large spatial data set for analysis.

In addition to those production data, we also need soil, topography, and weather data. Producers usually do not have complete information about those data. Instead, we use the public databases about soil and weather. For the soil information we use the USDA Gridded Soil Survey Geographic (gSSURGO) 10-meter resolution maps. The soil property variables include soil water capacity, soil pH, soil texture, etc., which are relatively stable over time. For weather we

use the daily temperature and precipitation of the 4-km resolution PRISM Climate Data. The production data, including the yield maps and application maps, are then matched to the public soil and weather maps to compose a final highly spatially disaggregated large data set.

Comparing with experimental data, there are mainly two advantages of the observational production data: (1) Large sample size. Experimental data usually only have dozens of observations, or hundreds at most, due to the high expense to conduct controlled experiments. While this sample size is sufficient to estimate single regression of yield on one input factor, it is too small to estimate the comprehensive regression with multiple factors and also to capture the complicated nonlinear interactions between the factors. The degree of freedom is not enough. For production data, it is possible to gather billions or even trillions of observations. That allows to support the estimate of very comprehensive models. (2) Substantial variations in yield-impacting factors. The experiments are often designed to only allow the variation of one or a few factors, while all the other factors are controlled the same. While this design greatly guarantees the accuracy of response estimation, it also restricts the response only under a specific condition (usually claimed as the "optimal condition"). The observed production data have variations in all kinds of factors. As long as the data cover large regions, all kinds of soil conditions will be included. Weather be the same in one field, but varies if we can get different regions there will be enough variations too. Those rich variations allow the estimation model that allows simultaneous interactions of fall factors.

It should also be noted that there are limitations to using the producer data. (1) The first disadvantage is lower data quality. The measurement accuracy in actual production practices are usually much lower comparing to the carefully measured experimental data. Due to producers' skills and operation errors, as well as equipment accuracy, there are more errors and mistakes in the data. What is worse, the errors are sometimes systemic and cannot be treated as random noise. The public national soil and weather databases are also coarser in accuracy, comparing to the on-site soil sampling tests. There is no quick solution to that issue, but it may be mitigated by increasing sample size. (2) The second weakness is the non-randomness of variabilities. Unlike experimental design where the amount of application is random, in real production the input levels of fertilizer, seeding, and water are decided based on soil and weather conditions.

That is essentially an endogeneity issue. While there is no perfect solution for that, fortunately the well-developed empirical economics literature has provided adequate amount of econometric techniques to address the endogeneity issue. In addition, the finally estimated model can be assessed using experimental data to check the biasness.

Collecting and organizing the disaggregated production data is a challenging task. There are many practical difficulties in both the data collecting techniques and data ownership legal issues. Many companies and organizations have already started some efforts to integrate farmers' data (such as JohnDeere, Montanto, *etc.*). As a starting point, our research team works with a local agricultural consulting company and collects corn producers' data spanning many locations across Southern United States. The process is ongoing and a pilot database will be built shortly. For the purpose of this paper, we use the simulated data to test the feasibility of our comprehensive regression model estimation of yield response functions.

*4.2 Simulation data*

We use a simulated data set to illustrate the regression estimation of site-specific yield response functions and the generating of VRA prescriptions. The goal is to demonstrate the issues with the prevailing site-dummy estimation approaches, and to illustrate the improvement obtained by explicitly incorporating continuous site variables in the regression. A major advantage of simulation data is that it allows for solid comparison of different site-specific prescriptions since the true data generating process is known to us. The simulated data results can provide useful insights for the model building and testing.

We design the simulation data in the following way. Assume there is a large crop field which is divided into 100 by 100 grids. The soil properties vary across the grids. For simplicity, we describe the soil properties by a single soil index ranging from 0 to 80. The distribution of the soil index is illustrated in Figure 2(a). Not surprisingly, the soil is highly spatially autocorrelated. We also assume the only input for the crop production in nitrogen fertilizer. The application rate of nitrogen for each grid is completely random, as shown in Figure 2(b). The values of nitrogen rate span from 0 to 120. However, those numbers are merely for illustration purpose and do not have direct meaning.

The yield of the crop in each grid is based on a simplified 100-day crop growth process, following the prevailing mechanistic crop growth models but is simplified in many aspects. We assume the higher soil index is associated with stronger holding capacity of nitrogen, while the lower-indexed soil leaks nitrogen faster. The crop plan takes up nitrogen from the soil and grows in body biomass on a daily basis. Nitrogen taking up rate is positively related to soil nitrogen density and the body's need. Growth rate is increasing with body nitrogen density, but too much nitrogen also stresses plant growth. The final yield is a proportion of the body biomass at the $100^{th}$ day. Based on this simple growth model, the simulated yields for each grid of the field is as shown in Figure 2(c).

We proceed to use the simulated 10,000 observations of field grids to estimate the site-specific crop yield response functions, and subsequently compare the performances of different variable rate application prescriptions. We only look at the yield at this point, but it is relatively straightforward to extend the analysis to profit after adding price information.

## 5. Result

### 5.1 *Yield response functions*

We use the simulated soil index, nitrogen rates, and yield data to estimate the varying coefficient model. Here there is only one site variable, the soil index, and therefore $X_i=S_i$. The exact functional forms of $\alpha()$, $\beta()$, and $\gamma()$ are not clear here. To get started we use a discrete transition function which simply creates six regimes based on the value of soil index. Essentially this can be viewed as a simple non-parametric model. The estimated yield responses to nitrogen rate is shown in Figure 3, where each curve is corresponding to a response in a soil regime. It shows a clear pattern that the response differs with soil index values. For lower soil index value, the response curve is located farther to the right, and also slightly flatter. It suggests the optimal nitrogen rate is higher for lower soil index value, and lower for higher soil index value. The difference in the optimal rates is notably large, with the highest rate as of 88 for the soil regime (0, 20], and the lowest rate as of 51 at the soil regime (60, 80]. Note that the response change is

not strictly speaking continuous from the soil regime model, but it approximately delineates the pattern of the change with respect to soil index.

To compare with the soil regime estimate, we also follow the prevailing approach of estimating the site-specific response based on lattice site dummy variables. The field is divided into 16 (4 by 4) regular sub-blocks, as shown in Figure 4(a). Each site is represented by a dummy variable, and interacts with the parameters of response function. The estimation results are shown in Figure 4(b). As can be seen, the responses are quite different across the sites. The underlying reason for this difference is the soil variation. However, the site-dummy approach does not explicitly take into account the soil information since the regular sites are delineated rather arbitrarily.

Finally, a uniform response function is also estimated by simply pooling all the grid data together, regardless of the soil index. The result is shown in Figure 5. That estimated response curve can be viewed as the average response for the entire field.

5.2 *Variable rate prescriptions*

Based on the estimated yield responses to nitrogen, the ultimate goal is to generate spatially explicit variable-rate-application (VRA) nitrogen prescriptions to achieve best output in each grid. Especially, we are interested to test whether the prescription created by the varying coefficient model has any advantage.

The quadratic functional form of response makes the derivation of the optimal nitrogen (N) fairly straightforward. For the uniform yield response, the optimal N rate at grid *i* is simply:

$$N_i^* = -\frac{\hat{\beta}}{2\hat{\gamma}} = -\frac{41.6}{2\times 0.302} = 68.8\,.$$

Note that because the parameters $\hat{\beta}$ and $\hat{\gamma}$ are constant across space, it is actually the uniform-rate-application (UAR) prescription. The N prescriptions from the site dummy regression and soil regime regression can be obtained in a similar way, except that the parameters $\hat{\beta}$ and $\hat{\gamma}$ vary with the 16 sub-blocks or the 6 soil regimes. The visualization of the nitrogen prescriptions generated by the three estimated response functions are shown in Figure 6(a).

Given that the data generating process is known by us, it allows to compare the performances of the three VRA prescriptions. For simplicity, we do not consider output price and input cost at this moment, and only compare the physical yield achieved. The comparison is conducted by applying the VRA nitrogen rates to the field, and re-simulate the crop growth process to obtain the new yield maps. The new yields from the three VRA prescriptions are shown in Figure 6(b). Through visual comparison, the soil regime VRA prescription yield is the higher than the rest two, and the uniform prescription yield is the lowest. The aggregate yield from the three VRA's are compared in Table 2. The soil regime model improves the total yield level by 6%, while the site dummy model improves 3%. Note that those numbers are determined by the simulation model parameters we choose, and are only for illustration purpose here. The actual magnitudes of improvement from VRA need to be assessed using actual production data.

The results demonstrate the potential yield improvement from using varying coefficient (soil regime) model. At the presence of soil heterogeneity, VRA of nitrogen can increase yield. Even from the *ex post* perspective VRA from explicitly incorporating soil information achieves higher yield than VRA from arbitrarily delineated sites. Furthermore, the varying coefficient (soil regime) model is able to generate *ex ante* VRA prescriptions to be applied to new fields, which is a large advantage against the dummy site-specific response model.

## 6. Conclusions and future research

This study explores the development of a new methodology to estimate the site-specific crop yield response function. Instead of using dummy variable to represent the sites, this study explicitly incorporates the actual site characteristic variables to explain the changes in the response. It uses the varying coefficient model where the parameters of the response changes as a function of the site conditions such as soil, water, topography, weather, and other factors that are commonly available to producers.

Based on a simulated data set, the varying coefficient model is estimated where the response changes with respect to the regimes of soil index values. The variable rate application (VRA) of

fertilizer prescription is generated based on the model estimate result, which provide site-specific optimal input rate in an *ex ante* perspective. The simulation results also demonstrate the varying coefficient model VRA prescription performs better than the dummy site-specific VRA prescription. This finding suggests evidence for the potential improvement by using the varying coefficient model in yield response estimation.

The next step is to collect high resolution crop production data for the model estimates, and incorporating more actual agro-climatic variables in the varying coefficient model. A limitation of actual production data is that we can no longer test the performance of the estimated response functions as well as their associated VRA prescriptions, since the true data generating process is unknown. We propose to design a validation algorithm of the prescriptions on the basis of mean-squared-error measure of the departure of model predicted yields from actual yields.

Furthermore, market fluctuations and price uncertainty ought to be included in the analysis to obtain profit-maximizing input rates. The ultimate goal is to build a precision input rates decision system, which can turn the soil testing and weather forecasting information into recommendations of spatially explicit economically optimal inputs application, and assist the producers to increase their output in precision farming.

## References:


Anselin, L., Bongiovanni, R., & Lowenberg-DeBoer, J. (2004). A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics*, 86(3), 675-687.

Basso, B., Cammarano, D., Fiorentino, C., & Ritchie, J. T. (2013). Wheat yield response to spatially variable nitrogen fertilizer in Mediterranean environment. *European Journal of Agronomy*, 51, 65-70.

Bongiovanni, R. G., Robledo, C. W., & Lambert, D. M. (2007). Economics of site-specific nitrogen management for protein content in wheat. *Computers and Electronics in Agriculture*, 58(1), 13-24.

Bouman, B. A. M., Van Keulen, H., Van Laar, H. H., & Rabbinge, R. (1996). The 'School of de Wit' crop growth simulation models: a pedigree and historical overview. *Agricultural systems*, 52(2), 171-198.

Bullock, D. S., Lowenberg-DeBoer, J., & Swinton, S. M. (2002). Adding value to spatially managed inputs by understanding site-specific yield response. *Agricultural Economics*, 27(3), 233-245.

de Wit, C.T. (1978), Simulation of Assimilation, Respiration and Transpiration of Crops. Simulation Monographs, PUDOC, Wageningen, The Netherlands.

Hurley, T. M., Malzer, G. L., & Kilian, B. (2004). Estimating site-specific nitrogen crop response functions: A conceptual framework and geostatistical model. *Agronomy Journal*, 96(5), 1331-1343.

Hurley, T. M., Oishi, K., & Malzer, G. L. (2005). Estimating the potential value of variable rate nitrogen applications: A comparison of spatial econometric and geostatistical models. *Journal of Agricultural and Resource Economics*, 30(2), 231-249.

Lambert, D. M., Lowenberg-Deboer, J., & Bongiovanni, R. (2004). A comparison of four spatial regression models for yield monitor data: A case study from Argentina. *Precision Agriculture*, 5(6), 579-600.

Lambert, D. M., Lowenberg-Deboer, J., & Malzer, G. L. (2006). Economic analysis of spatial-temporal patterns in corn and soybean response to nitrogen and phosphorus. *Agronomy Journal*, 98(1), 43-54.

Liu, Y., Swinton, S. M., & Miller, N. R. (2006). Is site-specific yield response consistent over time? Does it pay? *American Journal of Agricultural Economics*, 88(2), 471-483.

Mamo, M., Malzer, G. L., Mulla, D. J., Huggins, D. R., & Strock, J. (2003). Spatial and temporal variation in economically optimum nitrogen rate for corn. *Agronomy Journal*, 95(4), 958-964.

Melkonian, J. J., van Es, H. M., DeGaetano, A. T., & Joseph, L. (2008, July). ADAPT-N: Adaptive nitrogen management for maize using high resolution climate data and model simulations. In Proceedings of the 9th international Conference on Precision Agriculture.

Penning de Vries, F. W. T., Jansen, D. M., ten Berge, H. F. M. & Bakema, A. (1989). Simulation of ecophysiological processes of growth in several annual crops. Simulation Monographs, PUDOC, Wageningen, The Netherlands.

Ruffo, M. L., Bollero, G. A., Bullock, D. S., & Bullock, D. G. (2006). Site-specific production functions for variable rate corn nitrogen fertilization. *Precision Agriculture*, 7(5), 327-342.

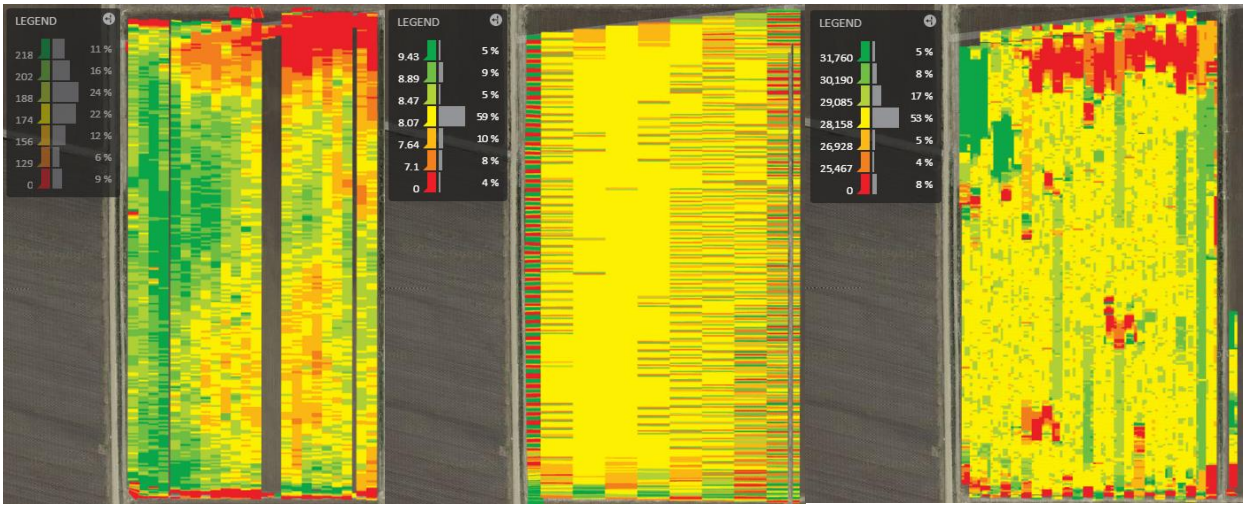
Spitters, C. J. T. (1989, August). Crop growth models: their usefulness and limitations. In VI Symposium on the Timing of Field Production of Vegetables 267, (pp. 349-368).

ten Berge, H. F. M., Thiyagarajan, T. M., Shi, Q., Wopereis, M. C. S., Drenth, H., & Jansen, M. J. W. (1997). Numerical optimization of nitrogen application to rice. Part I. Description of MANAGE-N. *Field Crops Research*, 51(1), 29-42.

Tsuji, G. Y., Uehara, G. & Bala, S. (eds) (1994). DSSAT Version 3.0. DSSAT approach to input management. University of Hawaii, Honolulu, Hawaii, USA.

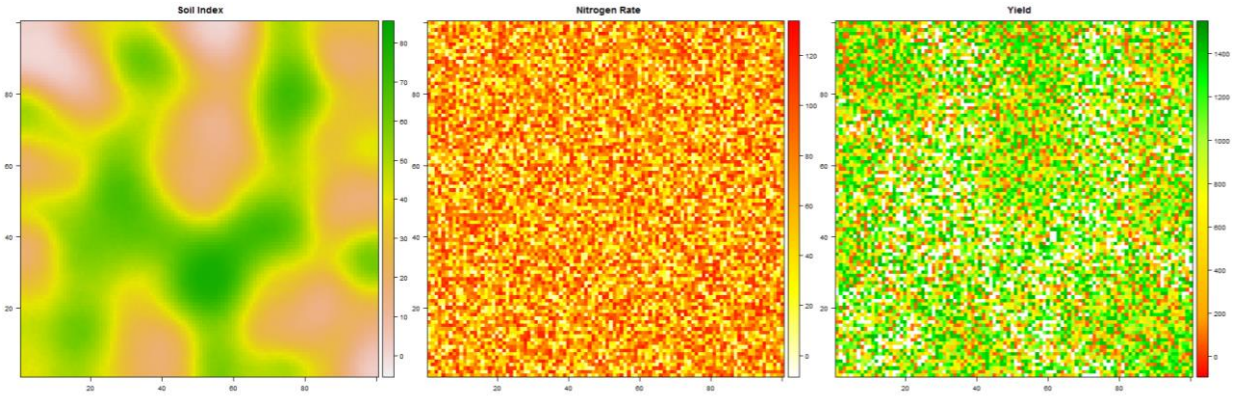
Williams, J. R., & Izaurralde, R. C. (2006). The APEX model. *Watershed Models*, 437-482.

Yang, H. S., Dobermann, A., Lindquist, J. L., Walters, D. T., Arkebauer, T. J., & Cassman, K. G. (2004). Hybrid-maize—a maize simulation model that combines two crop modeling approaches. *Field Crops Research*, 87(2), 131-154.




Yield Fertilization Seeding

Figure 1. An Example of Farm Production Data




(a) (b) (c)

Figure 2. A Simulate Sample Field.

- Size: 100 by 100 grids, with total observations of 10,000.
- Soil index ranges from 1 to 80. The higher the number, the stronger nitrogen holding capacity.
- Assume nitrogen rates are completely random. More realistic non-random nitrogen rate cases will be addressed later.
- White colored grids in yield map represent zero yield.

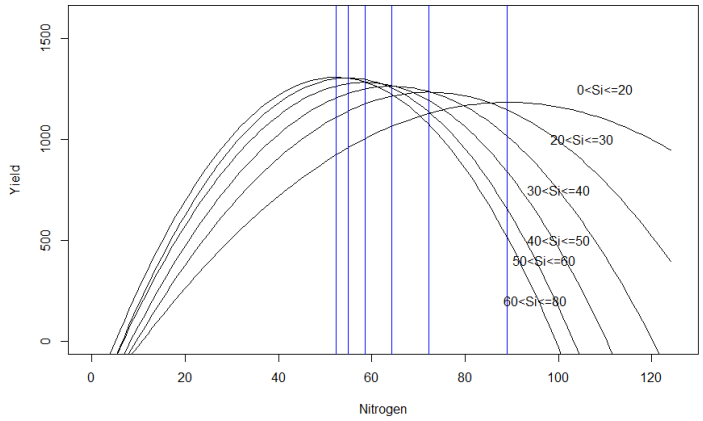



Figure 3. Estimated yield response to nitrogen by varying coefficient model

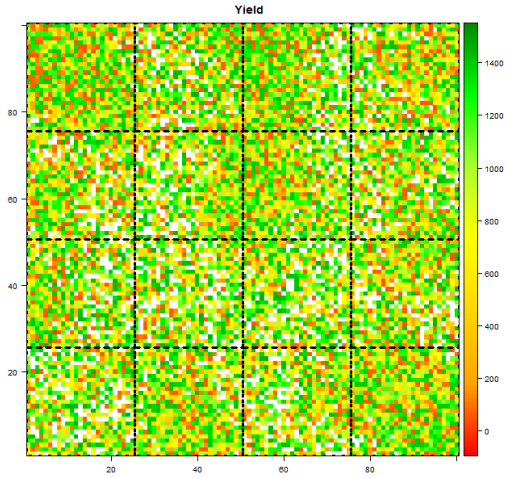
Note: The delineation of regimes by soil index value is as follows:

$$Y_i = \alpha(S_i) + \beta(S_i)N_i + \gamma(S_i)N_i^2 + \varepsilon_i$$

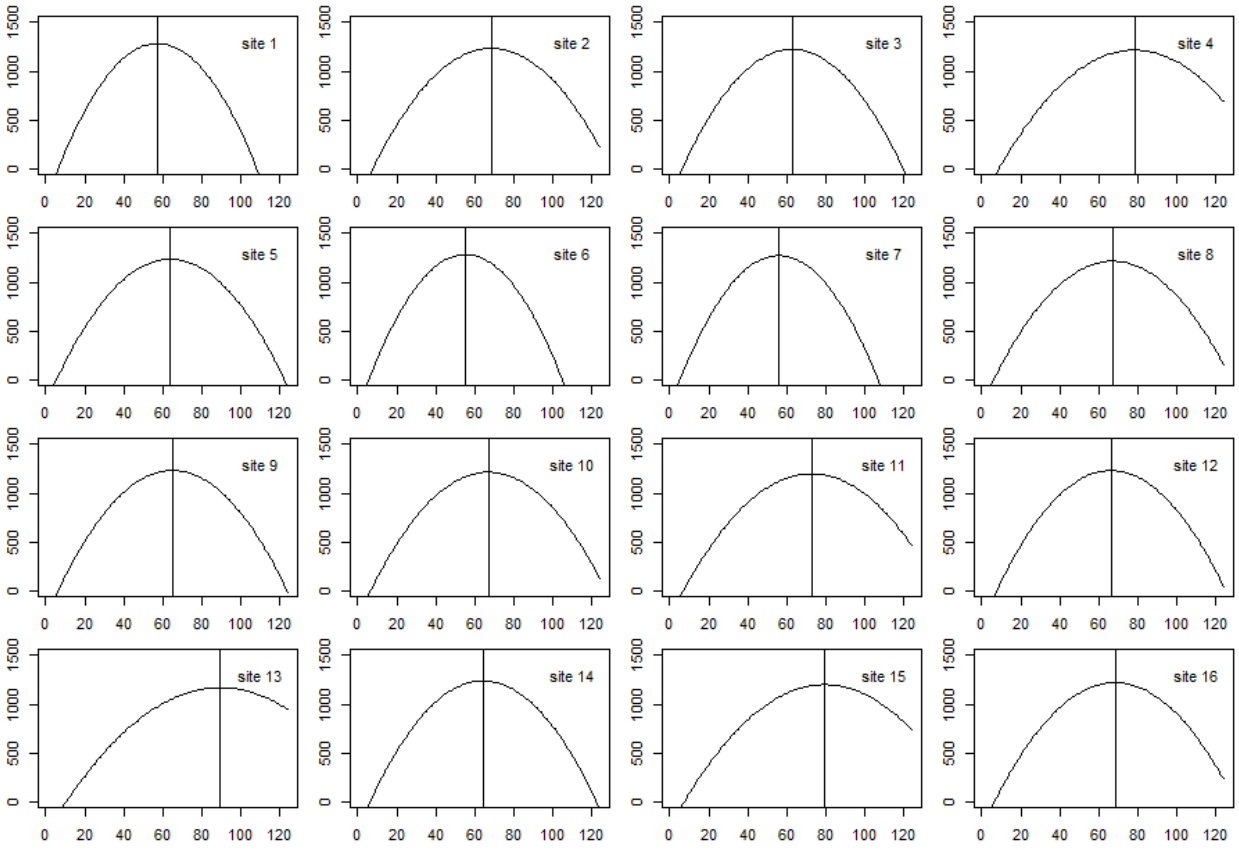
$$\alpha(S_i) = \begin{cases} \alpha_1, & \text{if } 0 < S_i \le 20 \text{ (Regime 1)} \\ \alpha_2, & \text{if } 20 < S_i \le 30 \text{ (Regime 2)} \\ \alpha_3, & \text{if } 30 < S_i \le 40 \text{ (Regime 3)} \\ \alpha_4, & \text{if } 40 < S_i \le 50 \text{ (Regime 4)} \\ \alpha_5, & \text{if } 50 < S_i \le 60 \text{ (Regime 5)} \\ \alpha_6, & \text{if } 60 < S_i \le 80 \text{ (Regime 6)} \end{cases}$$

$$\beta(S_i) = \begin{cases} \beta_1, & \text{if } 0 < S_i \le 20 \text{ (Regime 1)} \\ \beta_2, & \text{if } 20 < S_i \le 30 \text{ (Regime 2)} \\ \beta_3, & \text{if } 30 < S_i \le 40 \text{ (Regime 3)} \\ \beta_4, & \text{if } 40 < S_i \le 50 \text{ (Regime 4)} \\ \beta_5, & \text{if } 50 < S_i \le 60 \text{ (Regime 5)} \\ \beta_6, & \text{if } 60 < S_i \le 80 \text{ (Regime 6)} \end{cases}$$

$$\gamma(S_i) = \begin{cases} \gamma_1, & \text{if } 0 < S_i \le 20 \text{ (Regime 1)} \\ \gamma_2, & \text{if } 20 < S_i \le 30 \text{ (Regime 2)} \\ \gamma_3, & \text{if } 30 < S_i \le 40 \text{ (Regime 3)} \\ \gamma_4, & \text{if } 40 < S_i \le 50 \text{ (Regime 4)} \\ \gamma_5, & \text{if } 50 < S_i \le 60 \text{ (Regime 5)} \\ \gamma_6, & \text{if } 60 < S_i \le 80 \text{ (Regime 6)} \end{cases}$$




(a) Division of the field into 16 sites




(b) Response curves for each site

Figure 4. Estimated yield response by site dummy model

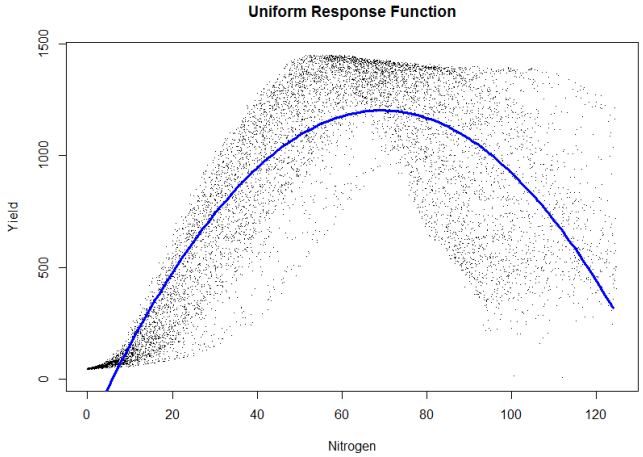
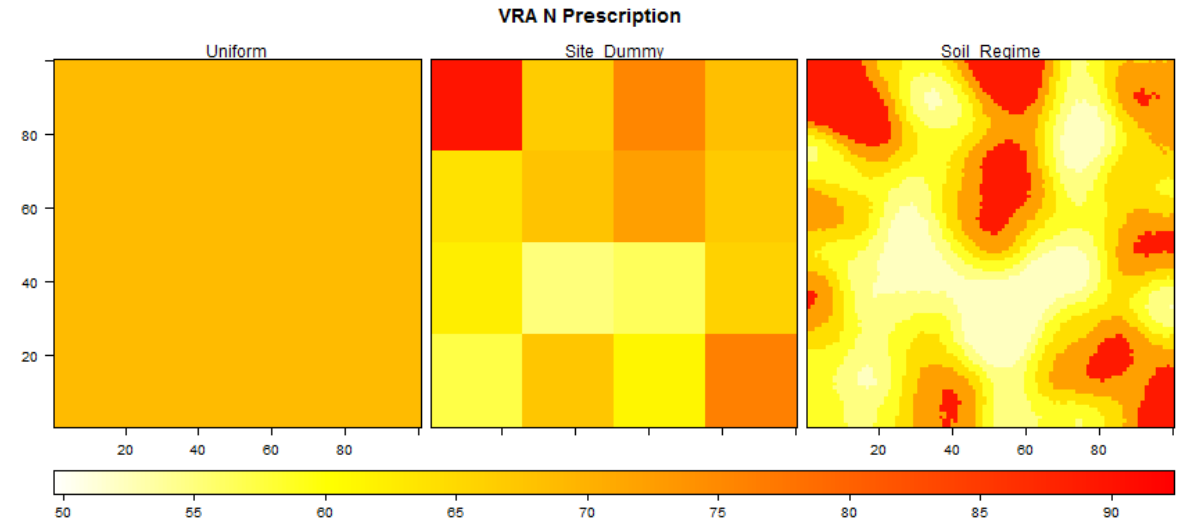
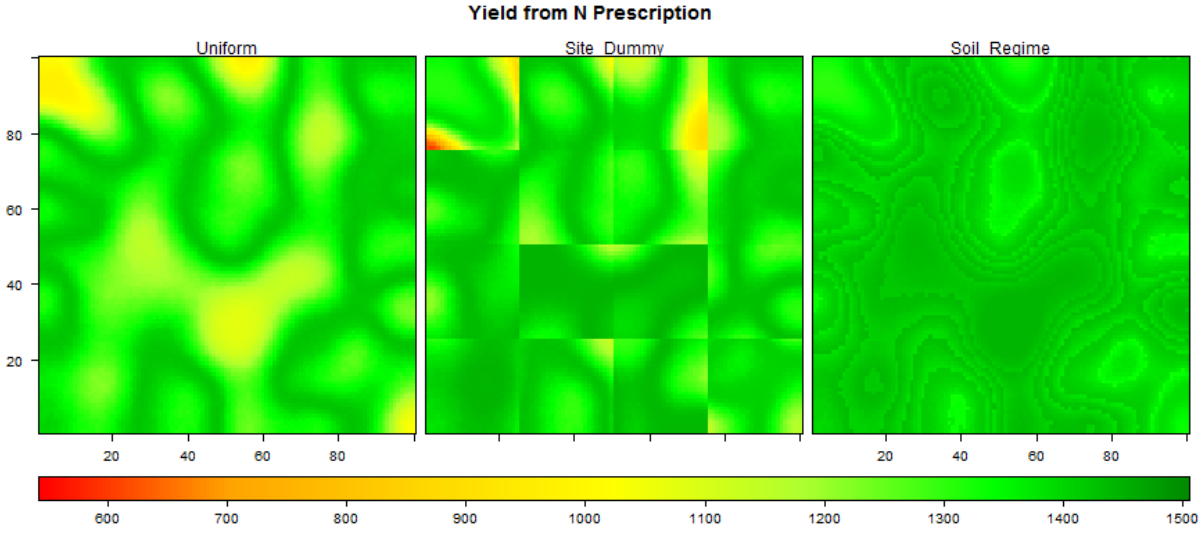



Figure 5. Estimated uniform yield response to nitrogen

$$\hat{Y}_i = -230.9 + 41.6N_i - 0.302N_i^2$$




(a) VRA nitrogen prescriptions generated from different yield response functions




(b) Optimal yields from the different VRA nitrogen prescriptions

Figure 6. VRA Performance Comparison from the Different Yield Response Functions

Table 1. "Site" Variables used in the Site-Specific Response Estimates

| Study | Site variables |
|---|---|
| Liu et al. (2006) | Organic matter, cation exchange capacity, leaching, water availability and sunlight reception |
| Ruffo et al. (2006) | Terrain attributes |
| Mamo et al. (2003) | 7 different soil types |
| Anselin et al. (2004), Lambert et al (2004) | 4 regimes, according to the slope of land |
| Hurley et al. (2004) | 6, 48 or 102 lattice "sites" |
| Long (1998) | Elevation, vegetation index: 2 blocks |
| Bongiovanni et al. (2007) | 2 types of soil/landscapes: Hilltop and Lowland; and 2 antecessor crops: corn and soybean |
| Lambert et al. (2006) | 69 sub-blocks |

Table 2. Yield comparison of VRA prescriptions

| VRA N prescription | Total yield | % |
|---|---|---|
| Uniform | 13215656 | 100 |
| Site Dummy | 13644886 | 103 |
| Soil Regime | 14050461 | 106 |