How to Use Yield Monitor Data to Determine Nitrogen Recommendations

: Bayesian Kriging for Site-Specific Parameter Estimates

Eunchun Park, B. Wade Brorsen, and Xiaofei Li

Selected Paper prepared for presentation at the 2018 Agricultural & Applied Economics Association Annual Meeting, Washington, D.C., August 5-7, 2018

Eunchun Park is an assistant professor in the Department of Agricultural Economics at Mississippi State University, B. Wade Brorsen is a regents professor and A.J. and Susan Jacques Chair, Department of Agricultural Economics, Oklahoma State University, Stillwater, Oklahoma, Xiaofei Li is an assistant professor in the Department of Agricultural Economics at Mississippi State University.

In the big data era, agricultural producers and crop consultants are being inundated with data. Modern GIS and computer technologies have developed data for applying precision agriculture. One of the largest sources of the data comes from yield monitors. Perhaps the most critical place yield monitor data could be used is in establishing nitrogen fertilizer recommendations. Due to considerable site-specific heterogeneity in farming fields, such as soil type, weather, and moisture, optimal input recommendation can vary spatially. Therefore, optimal input recommendations across heterogeneous locations ultimately rely on the ability to establish well-identified site-specific yield response functions. In this regard, a methodological development of obtaining the site-specific recommendations will fuel the adoption of precision agriculture.

Several types of regression models have been employed to obtain optimal input recommendations. These models generally fit yield data in response to the different level of inputs in the agronomic experimental data. Particularly, many studies have employed spatial econometric models due to the existence of spatial correlation of crop yield outcomes. An ordinary least squares (OLS) model that assumes spatial independence will produce inefficient estimates when there exists the spatial correlation, mainly due to the incorrect variance estimates (Anselin et al., 2004; Lambert et al., 2004). To address the spatial correlation in the data, therefore, spatial econometric models are mainly used in recent studies. The studies include Anselin et al. (2004), Lambert et al. (2004), Hurley et al. (2004), Liu et al. (2006), and Lambert et al. (2006). A typical approach of the studies for site-specific recommendations is defining sub-districts and impose dummy variables on the districts to account spatial heterogeneity. However, the arbitrarily defined district dummies do little to explain the causality of the site heterogeneity since the relationship between the input response variability and location characteristics, such as soil type, moisture rate, is not thoroughly considered when comparting the districts. More importantly, the approach cannot provide the different level of input response coefficients of every site in the farming field.

The spatial econometric models for the yield response functions are mostly based on the classical statistical inference, which assumes the non-stochastic parameter and random input variables. Therefore, the estimation results of such methods only provide a representative yield response to the input level changes of arbitrarily defined sub-districts. To obtain input response coefficients for every site, therefore, we should regress the model independently by sites. For instance, if we have N number of the sites, then we should regress N times separately by using each site's own observations. If we have sufficiently large enough number of historical outcomes for each site with variable input uses, we might get good estimator of the input response coefficient. However, current field monitoring data has small historical observations and uniform input rates.

In this regard, insufficient historical observations of the monitoring data can be the most critical obstacle when estimating site-specific parameter estimates. Although the field monitoring data includes large locations with high-quality spatial information, most datasets have less than ten years observations in each location. Therefore, we cannot expect accurate site-specific coefficient estimates from the general spatial econometric models. Moreover, a uniform rate of nitrogen has been applied to most fields. Therefore, we cannot expect sufficient variation of input uses for single locations over time as well. This can be another obstacle to apply the general spatial econometric methods to fit the field monitoring data.

Ultimately of interest is how can these yield monitor data be used to estimate some of the parameters of a production function in order to make variable rate nitrogen recommendations. Variable rates can now be applied at low cost, but the problem of how best to determine the rate applied has not been solved. We address the problem by using Bayesian Kriging (Gaussian process regression) method. The method assumes that the input response coefficients are stochastic and spatially correlated. The method produces spatially smoothed site-specific coefficients based on the latent spatial correlation structure. We adopt the method because crop yields in the field monitors show strong spatial correlation. Since Bayesian statistic assumes stochastic parameters, spatially correlated stochastic coefficients can describe the spatial correlations in crop yields. The spatial structure is updated under the Bayesian inference by using the yield outcomes and site-specific information (locational distance, soil type, etc). We assume a linear stochastic plateau production function as in Tembo et al. (2008) and estimate the plateau and its variance for each point in the field. To evaluate the accuracy of the proposed model, we conduct a Monte-Carlo simulation under the three different scenarios on the spatial correlation structure: (a) strong spatial correlation on the plateau, (b) randomly distributed plateau, and (c) uniformly distributed plateau over space. Each scenario has datasets with a different number of locations and historical observations. We then compare the accuracy of the proposed model with the maximum likelihood estimation (MLE) method. The root mean squared error (RMSE) is used as a tool to measure the accuracy. The results are strongly favorable to the new method in all combinations in each scenario with more than 50 locations. The new method shows great accuracy even when each location has 5 observations if the dataset has sufficient number of locations (more than 50 locations). Specifically, the new method more significantly outperforms to the general MLE when there exists a strong spatial correlation. Recommendations are then made using Bayesian decision theory.

The objective of the study is to answer the question of how to use yield monitor data to determine nitrogen recommendations. To address the issue, we develop a new estimation routine that provides site-specific coefficient estimates based on the Bayesian Kriging method. The estimated site-specific coefficient are used to suggest variable nitrogen recommendations to maximize an expected yield and profit of farms.

Data

The data we use for this analysis are collected from about 491 acres of corn production fields in Mississippi. Those fields are adjacent to each other and provide a very good sample for spatial analysis. The major soil types are Dubbs silt loam and Dundee silt loam, with around 0 to 2% slopes. A small portion of the land area is covered by Dundee silty clay loam. Consecutive corn was planted in those fields from 2011 to 2016. However, only four years have complete input data records (2012, 2014, 2015, and 2016), which are used in the modeling of this study.

The original yield data are geospatially referenced point records collected from yield monitor. The corn harvesting usually happened between August 1 and 23 during the sample years. The yield monitor was installed on the John Deere corn harvester, and with 6 sensors in a row mounted on the header of the harvester. The distance between those sensors is about 2 meters. The sensors recorded the instantaneous yield and moisture data at 1 second intervals during harvesting. The length of the intervals depends on the speed of the driving. At an average speed of about 7 km/h, the length is about 2 meters. Therefore the original yield data are geo-referenced points about 2 meters apart from each other. An illustration of the yield point data is shown in Figure 1. The farm also kept records of seeding rates and nitrogen fertilizer application data. Those are also geospatially referenced point data collected by planting machine and fertilizer sprayer, which are in similar data format as yield. The seeding rate points are about 0.45 meter apart. Fertilizer application rate points are 3 meters to 12 meters apart, depending on the types of sprayer used.

We then aggregate the point data (yield, seeding rates, and nitrogen fertilizer rates) into grids with a square cell size of 100 meters. The value for each grid cell is the mean of the original point readings falling within the 100 meters by 100 meters grid cell. The border areas of the field usually have some abnormal values in the original point readings. Those are likely due to the turning and speed changing of farming machines during operation. While it is very cumbersome to correct for those data errors, we simply discarded all points within 20 meters of the border line. Certain portions of the field had missing data in some years. Those portions were discarded as well. In the end we obtained a balanced panel of 160 grids covering 4 years (2012, 2014, 2015, and 2016). The locations of those grids are visually presented in Figure 2 (using yield as an example).

The average yield was relatively stable over time. The pooled average yield for all grids of all years is 216.2 bu/acre. Year 2012 had the highest average yield of 229.9 bu/acre, while year 2015 had the lowest of 207.8 bu/acre. The across space variability in yield was much more substantial. In the pooled data, the highest-yield grid was 274 bu/acre, while the lowest was 163.5 bu/acre. The spatial variability in corn yield can be easily detected as in Figure 2. Seeds were usually planted during March 26 and April 4 in those sample years. Different varieties and target seeding rates were applied, ranging from 28k to 36k seeds per acre. The farm applied uniform rate seeding at field level (or a large part of the field). But during the operation the actually applied rates may deviate from the target rates. Those operation “errors” created important sources of variations in seeding rates for the analysis. In the sample the lowest grid level seeding rate is 25.3k seed/acre, and the highest is 36.5k seed/acre. Several types of liquid nitrogen fertilizer were applied in the fields, including 28-0-0-5, 30-0-0-2, and 32-0-0. The split applying pattern was used, with starters usually applied at the seeding (March 26 to April 4), and sidedress applied in early May (May 01 to May 05). The total nitrogen amount was calculated by summing up the actual nitrogen pounds in the different nitrogen types. It should be noticed that the timing of nitrogen fertilizer application also critically impacts corn yield. But the timing issue is by itself a complicated research question, and large number of studies have looked into it yet without consensus conclusions. Some studies found no effects of split nitrogen application while some found positive effects. It is beyond the scope of this paper to explore the timing effect. We only look at the effect of total nitrogen amount at this point, and leave the timing effect for future research. Similar to seeding rates, the farm applied uniform rate nitrogen fertilizer at field level. But in the actual application operations some random variations happened in the applied rates from the target rates. Because those variations can be regarded as operation errors, most likely they can be treated as random variations. The descriptive statistics of the major variables are presented in Table 1.

Bayesian Modeling Framework

Crop response function

Many agronomic studies on crop yield response to nitrogen input have suggested using a plateau function, but the plateau can vary across locations and years. In this regard, we adopt a stochastic linear response plateau function (Tembo et al. 2008; Ouedraogo and Brorsen 2017). They introduce methods of estimating a response function with a stochastic plateau that addresses field and year random effects based on classical (Tembo et al. 2008) and Bayesian econometric methods (Ouedraogo and Brorsen 2017). Unlike their models, we assume that the plateau means $\boldsymbol{P}$ and variation $\boldsymbol{\epsilon}$ of fields are spatially correlated and vary across locations. Our model assumes a single nitrogen input, a linear response, and normality. The proposed site-specific stochastic plateau function can be defined as

$$(1) \qquad y_{it} = \min(\alpha + \beta x_{it}, P_i + \epsilon_i) + \varepsilon_{it}$$

where x_{it} and y_{it} are the level of the nitrogen input and response yield on ith plot at year t, $i = 1, \dots, N$ and $t = 1, \dots, T$, the plateaus P_i are assumed to be spatially correlated by distance and multivariate normally distributed, $\boldsymbol{P} \sim MVN(\overline{\boldsymbol{P}}, \Sigma_P)$, where $\boldsymbol{P} = [P_1, \dots, P_N]'$, $\Sigma_P = \rho_P e^{-D_{ij}/\theta_P}$, ϵ_i is a spatially correlated plateau error (plateau variance), $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, \Sigma_\epsilon)$, where $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_N]'$, $\Sigma_\epsilon = \rho_\epsilon e^{-D_{ij}/\theta_\epsilon}$ and ε_{it} is an independently identically distributed error term $\boldsymbol{\varepsilon_t} \sim MVN(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_{NT})$, where $\boldsymbol{\varepsilon_t} = [\varepsilon_{1t}, \dots, \varepsilon_{Nt}]$.

The model allows spatially varying plateau variance. Tembo et al. (2008) assume a time random effect for the plateau, which implicitly assumes spatial correlation equal to one. Yield monitor data typically have a short time series, but a large number of cross-section observations. Since we are estimating different plateau parameters for every area of the field, Bayesian Kriging offers a way to use the spatial information in estimating the plateau parameters.

Hierarchical structure

The Bayesian Kriging regression model (Gaussian process regression) in the study has three hierarchical layers. First, in the likelihood layer, the crop yields are assumed to follow a normal distribution. Second, the process layer models the spatial structure of the density parameters. This second layer produces spatially varying (site-specific) parameters (plateau and its variance for each site). The model updates the spatial structure within the Markov Chain Monte Carlo (MCMC) procedure. The third layer imposes priors for the explanatory variables and Kriging parameters to conduct spatial smoothing, called hyper priors.

The hierarchy we use can be structured as,

$$\begin{aligned} &\boldsymbol{Y}|\, \boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta} \sim p_1(\boldsymbol{Y}|\, \boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta}) \\ &\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t|\, \boldsymbol{\Theta} \sim p_2(\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t|\, \boldsymbol{\Theta}) \\ &\boldsymbol{\Theta} \sim p_3(\boldsymbol{\Theta}) \end{aligned} \tag{2}$$

where p_1, p_2, and p_3 are the densities associated with each layer of the hierarchy, likelihood layer, process layer, and prior layer, $\boldsymbol{Y}$ is a $N \times T$ matrix of crop yield observations that spans all sites $(n = 1, \dots, N)$ and years $(t = 1, \dots, T)$, $\boldsymbol{\mu}$ is the input response of the crop yields, $\boldsymbol{\mu} = \alpha + \beta \boldsymbol{X}$, $\boldsymbol{X}$ is a $N \times T$ matrix of nitrogen inputs spanning all sites and years, $\boldsymbol{P}$ is a vector of plateau process for each site, $\boldsymbol{P} = [P_1, \dots, P_N]'$, $\boldsymbol{\epsilon}_t$ is spatial correlated plateau error (variance) process, $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_N]'$, $\boldsymbol{\varepsilon}_t$ is an independently identically distributed error, $\boldsymbol{\varepsilon}_t = [\varepsilon_{1t}, \dots, \varepsilon_{Nt}]$, and $\boldsymbol{\Theta}$ is a vector of hyper priors, $\boldsymbol{\Theta} = [P_i\, , \epsilon_i\, , \rho_P, \rho_\epsilon, \theta_P, \theta_\epsilon, \sigma^2]'$.

By Bayes' theorem, the joint posterior distribution of the model is

$$p(\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta} \mid \boldsymbol{Y}) \propto p_1(\boldsymbol{Y} \mid \boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta}) p_2(\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t \mid \boldsymbol{\Theta}) p_3(\boldsymbol{\Theta}). \tag{3}$$

Therefore, the joint posterior density of the model $p(\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta} \mid \boldsymbol{Y})$ is proportional to the multiplication of the three layers, which are specified in the following subsections.

Likelihood layer

A likelihood function of the crop yield distribution forms the first layer of the model. Let $\boldsymbol{y}_t$ is a vector of crop yield at year t that spans all locations, $\boldsymbol{y}_t = [y_{1t}, \dots, y_{Nt}]'$. Since the model assumes normality of crop yield distributions, the first layer of the model in equation (2) is,

$$\begin{aligned} & p_1(\boldsymbol{Y} \mid \boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t, \boldsymbol{\Theta}) \\ & = \frac{1}{\sqrt{(2\pi)^N \sigma^2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left[(\boldsymbol{y}_t - \boldsymbol{\mu})' \boldsymbol{\Psi} (\boldsymbol{y}_t - \boldsymbol{\mu}) + (\boldsymbol{y}_t - \boldsymbol{P})' (\mathbf{I} - \boldsymbol{\Psi})(\boldsymbol{y}_t - \boldsymbol{P}) \right] \right) \end{aligned} \tag{4}$$

where $\boldsymbol{Y}$ is a matrix of historical yield outcomes spans all locations and years, $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_T]$, $\boldsymbol{\Psi}$ is a $N \times N$ diagonal matrix where the nth diagonal elements are 1 if $\mu < P_i + \epsilon_i$, 0 otherwise, $\mathbf{I}$ is an $N \times N$ identity matrix, and σ^2 is a variance of the i.i.d error $\boldsymbol{\varepsilon}_t$ of the model.

Process layer

The process layer is the key part of the model. The process layer models spatial structure of the site-specific coefficients. The theoretical background to produce the parameters is based on a Gaussian spatial process[1]. The process assumes that the location specific coefficients are spatially correlated. The correlation structure is determined by the Kriging parameters (range θ and sill ρ) and Euclidean distances (D_{ij}) among locations[2].

We transform the stochastic plateau function in equation (1) to the process layer form of the Bayesian Kriging regression model,

$$\boldsymbol{Y_t} = \min(\boldsymbol{\mu}, \boldsymbol{P} + \boldsymbol{\epsilon}) + \boldsymbol{\varepsilon_t}. \tag{5}$$

where $\boldsymbol{\mu}$, $\boldsymbol{P}$, $\boldsymbol{\epsilon}$, and $\boldsymbol{\varepsilon}$ are input response yield, site-specific plateau, spatially correlated plateau error (variance), and i.i.d. error which are defined in equation (2) such that,

$$\begin{aligned} &\boldsymbol{\mu} = \alpha + \beta \boldsymbol{X}, \\ &\boldsymbol{P} \sim GP(\overline{\boldsymbol{P}}, \Sigma_P) \\ &\Sigma_P = \rho_P e^{-D_{ij}/\theta_P}, \\ &\boldsymbol{\epsilon} \sim GP\big(\boldsymbol{0},\ \Sigma_\epsilon\big), \\ &\Sigma_\epsilon = \rho_\epsilon e^{-D_{ij}/\theta_\epsilon}, \\ &\boldsymbol{\varepsilon}_t \sim MVN\big(\boldsymbol{0},\ \sigma^2 \boldsymbol{I}\big) \end{aligned} \tag{6}$$

where $\boldsymbol{X}$ is a $N \times T$ matrix of uniform nitrogen inputs spanning all sites and years, $\boldsymbol{P}$ is a vector of plateau levels for each site that is assumed to follow a Gaussian spatial process with covariance matrix[3] Σ_P, the spatially correlated plateau error $\boldsymbol{\epsilon}$ is also assumed to follow the Gaussian spatial process with a covariance matrix Σ_ϵ, D_{ij} is the Euclidean distance between two locations i and j, and $\boldsymbol{\varepsilon}_t$ is an independently identically distributed (i.i.d) error that is assumed to follow multivariate normal distribution with zero mean and constant variance over space with no spatial correlation[4].

The discussion here is to model the input response yields $\boldsymbol{\mu}$ to match the empirical model. Ideally, if the data are cross-section time-series data with multiple plots across sites and years then variable stochastic responses over sites and years would be needed. Instead of the stochastic response, we model that the deterministic input response in all sites and years to reflect the reality of the empirical dataset uniform nitrogen application[5]. We estimate a stochastic plateau model to find good priors of the $\boldsymbol{\mu}$.

The two stochastic spatial processes for the plateau (mean $\boldsymbol{P}$ and variance $\boldsymbol{\epsilon}$) are assumed to be independent each other in the MCMC. Therefore, the second part of equation (3), which is the process layer of the model is the multiplication of two stochastic processes,

$$\begin{aligned} (7)\quad & p_2(\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}, \boldsymbol{\varepsilon}_t |\, \boldsymbol{\Theta}) \\ & = \frac{1}{\sqrt{(2\pi)^N |\Sigma_\epsilon|}} \exp\left[-\frac{1}{2}\boldsymbol{\epsilon}\, \acute{\Sigma}_\epsilon{}^{-1}\boldsymbol{\epsilon}\right] \times \frac{1}{\sqrt{(2\pi)^N |\Sigma_P|}} \exp\left[-\frac{1}{2}(\boldsymbol{P} - \overline{\boldsymbol{P}})\, \acute{\Sigma}_P{}^{-1}(\boldsymbol{P} - \overline{\boldsymbol{P}})\right]. \end{aligned}$$

Prior layer

The prior payer imposes priors for the hyper-parameters $\boldsymbol{\Theta}$. The hyper-parameters include the plateau ($\boldsymbol{P}$), spatially correlated plateau error ($\boldsymbol{\epsilon}$), Kriging parameters for each spatially correlated parameter (sill: ρ_P and ρ_ϵ, range: θ_P and θ_ϵ), and the i.i.d. error variance σ^2 in the process layer. The model assumes that the priors for the hyper-parameters are independent each other. Therefore, multiplication of all priors for the hyper parameters forms the prior layer. First, we impose spatial Gaussian priors (multivariate normal priors) for the location specific plateau $\boldsymbol{P}$ and spatially correlated error $\boldsymbol{\epsilon}$. For the i.i.d. error variance σ^2 and two Kriging sill parameters (ρ_P and ρ_ϵ), we impose general inverse gamma priors. For the two Kriging range parameters (θ_P and θ_ϵ), which describe the length of spatial correlation among the location specific parameters, however, more carefully considered priors should be imposed. Bayesian statistics literature (Banerjee, Carlin, and Gelfand 2004; Cooley, Nychka, and Naveau 2007) argue that improper priors induce bad convergence and improper posteriors. Following to the suggestions of the literature (Banerjee, Carlin, and Gelfand 2004; Sahu, Gelfand, and Holland 2006; Cooley, Nychka, and Naveau 2007) and agricultural economics literature (Park, Brorsen, and Harri 2016, 2017), we impose informative priors on both the Kriging parameters (ρ_P, ρ_ϵ, θ_P, θ_ϵ) and the coefficient parameters ($\boldsymbol{\mu}, \boldsymbol{P}, \boldsymbol{\epsilon}$) using the empirical information. We use the information (maximum distance among the all locations in the dataset) because the range parameter could neither fall below zero nor exceed the maximum distance in the dataset.

To impose proper priors for coefficient parameters, we run a stochastic linear plateau model under the classical econometric manner[6] and use the estimates to impose priors. We impose multivariate normal priors for the plateau P with $P \sim N(264.76, 10^3)$ and the plateaus means are allowed to vary spatially with $GP(\overline{\boldsymbol{P}}, \Sigma_P)$ where $\mathrm{E}(\boldsymbol{P}) = \overline{\boldsymbol{P}}$ and $\Sigma_P = \rho_P e^{-D_{ij}/\theta_P}$, and impose general inverse gamma priors on the Kriging sill parameter $\rho_P \sim IG(0.1, 0.1)$. For Kriging range parameter for the plateau mean, we impose uniform prior $\theta_P \sim U\big(0, \max\big(D_{ij}\big)\big)$, where $\max\big(D_{ij}\big)$ is maximum distance among the all locations in the dataset. For all the spatially correlated plateau error $\boldsymbol{\epsilon}$, we impose the multivariate normal priors $\boldsymbol{\epsilon} \sim MVN(\boldsymbol{0}, 10^3\boldsymbol{I})$. They are allowed to vary spatially with $GP(\boldsymbol{0}, \Sigma_\epsilon)$, where $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\Sigma_\epsilon = \rho_\epsilon e^{-D_{ij}/\theta_\epsilon}$. We impose inverse gamma prior and uniform prior for sill and range parameter of the $\boldsymbol{\epsilon}$, where $\rho_\epsilon \sim IG(0.1, 0.1)$ and $\theta_\epsilon \sim U\big(0, \max\big(D_{ij}\big)\big)$. Finally, we impose general inverse gamma prior for the i.i.d. error variance σ^2, where $\sigma^2 \sim IG(0.1, 0.1)$.

The third layer in equation (3) is then structured as

(8) $$p_3(\boldsymbol{\Theta}) = p(\boldsymbol{P})p(\boldsymbol{\epsilon})p(\rho_P)p(\rho_\epsilon)p(\theta_P)p(\theta_\epsilon)p(\sigma^2).$$

Monte-Carlo Simulation of the model

To evaluate accuracy of the model, we conduct Monte-Carlo simulation. We generate the spatially correlated plateau parameters and the correlated error in the hypothetical farming field, and re-estimate those known parameters using the Bayesian Kriging regression model. The Monte-Carlo datasets are generated under the three different scenarios for the spatial structure: (a) strong spatial correlation on the plateau, (b) randomly distributed plateau, and (c) uniform plateau over all locations. To impose reliable level of inputs and plateau level, we estimate a Bayesian stochastic linear plateau model by using the dataset we have, and get the average and standard deviation of the parameter values. Then we generate each location's (say location i) different plateau levels, say P_i, that are spatially correlated (by distance). We first create 1,000 bootstrapped samples for each observation / location combination based on the hypothetical true parameter values, and re-estimate the parameters by using both the Kriging regression model and a general maximum likelihood estimation (MLE) model (regress by locations) to get each location's plateau parameters. Then we calculate the RMSE for the estimated parameters from each model ($\tilde{P}_i = Kriging,\ \hat{P}_i = MLE$) compared to the true known plateau parameter values β_{1i} and β_{2i},

(9) $$RMSE_{Kriging} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \tilde{P}_i\right)^2}\ ,\quad RMSE_{MLE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(P_i - \hat{P}_i\right)^2}\ ,$$

and calculate the ratio of RMSE, which can be defined as,

(10) $$D = \frac{RMSE_{Kriging}}{RMSE_{MLE}}.$$

We calculate the index D for each bootstrapped sample, and count the number of greater than 1 out of 1,000 samples. Table 2 presents estimated p-values from the bootstrapped samples. We find that the Kriging parameter smoothing model performs very well for 5 observations but need to have enough number of locations to obtain good estimator. For 50 number of observations, the Kriging model less outperforms to the MLE but still works well when we have enough locations (50 or 100 locations).

Expected conclusions

Table 1. Descriptive Statistics of Gridded Data

		Observations	Mean	St. Dev.	Min	Max
Yield (Bu/Acre)		640	216.2	18.3	163.5	274.0
	2012	160	229.9	20.9	174.8	264.5
	2014	160	213.8	18.0	163.5	274.0
	2015	160	207.8	10.8	177.9	226.9
	2016	160	213.4	13.8	166.9	242.8
Seed (1,000/Acre)		640	32.1	2.4	25.2	36.5
	2012	160	35.0	0.6	34.0	36.5
	2014	160	31.4	1.1	25.2	34.0
	2015	160	28.9	1.1	27.4	31.9
	2016	160	33.1	0.9	31.6	34.9
Nitrogen (Lb/Acre)		640	145.3	58.8	60.7	887.5
	2012	160	93.4	6.3	81.6	157.5
	2014	160	113.5	71.8	82.0	887.5
	2015	160	162.0	5.8	139.7	183.0
	2016	160	212.2	13.1	60.7	229.5

Table 2. Average of 90 Percent Coverage Premium Rates across Counties

State / Smoothing Space		Physical Space		Climate Space	
Model Structure		GP	AR	GP	AR
Iowa	Premium Rate (%)	1.73	1.79	1.57	1.59
Illinois	Premium Rate (%)	1.52	1.54	1.54	1.55
Nebraska	Premium Rate (%)	1.38	1.27	1.28	1.31
Minnesota	Premium Rate (%)	1.59	1.57	1.63	1.59
Indiana	Premium Rate (%)	1.65	1.69	1.64	1.75
Colorado	Premium Rate (%)	1.48	1.62	2.41	2.53

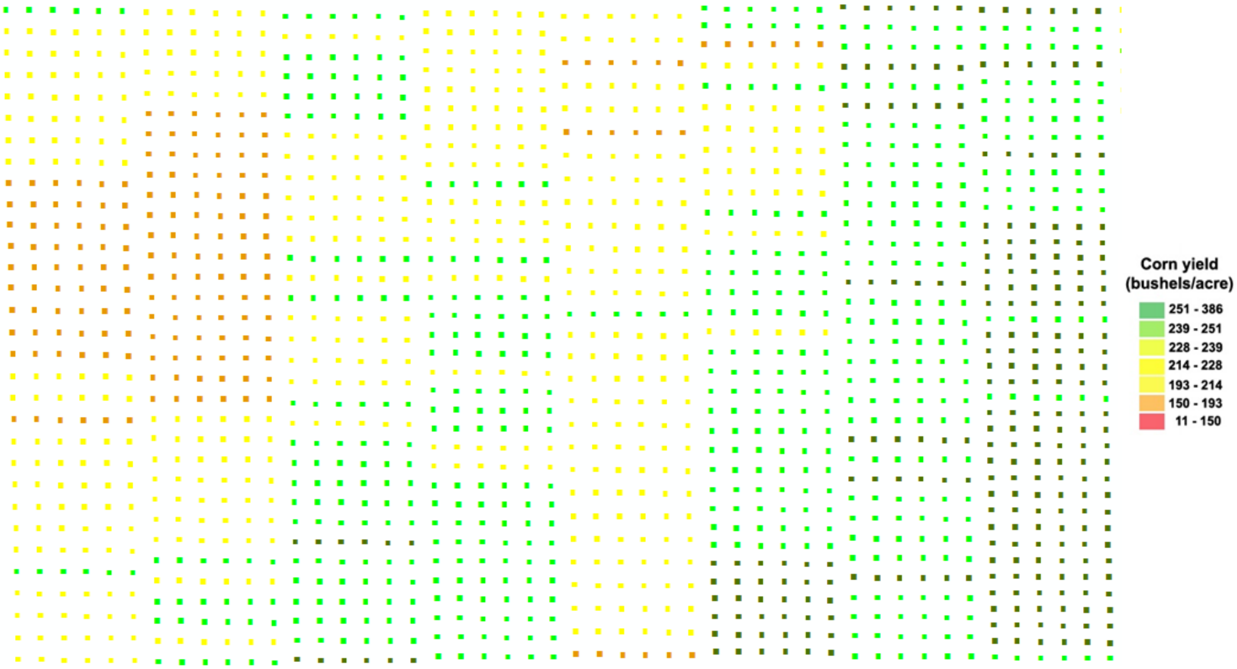

Figure 1. Illustration of the original yield monitor data.

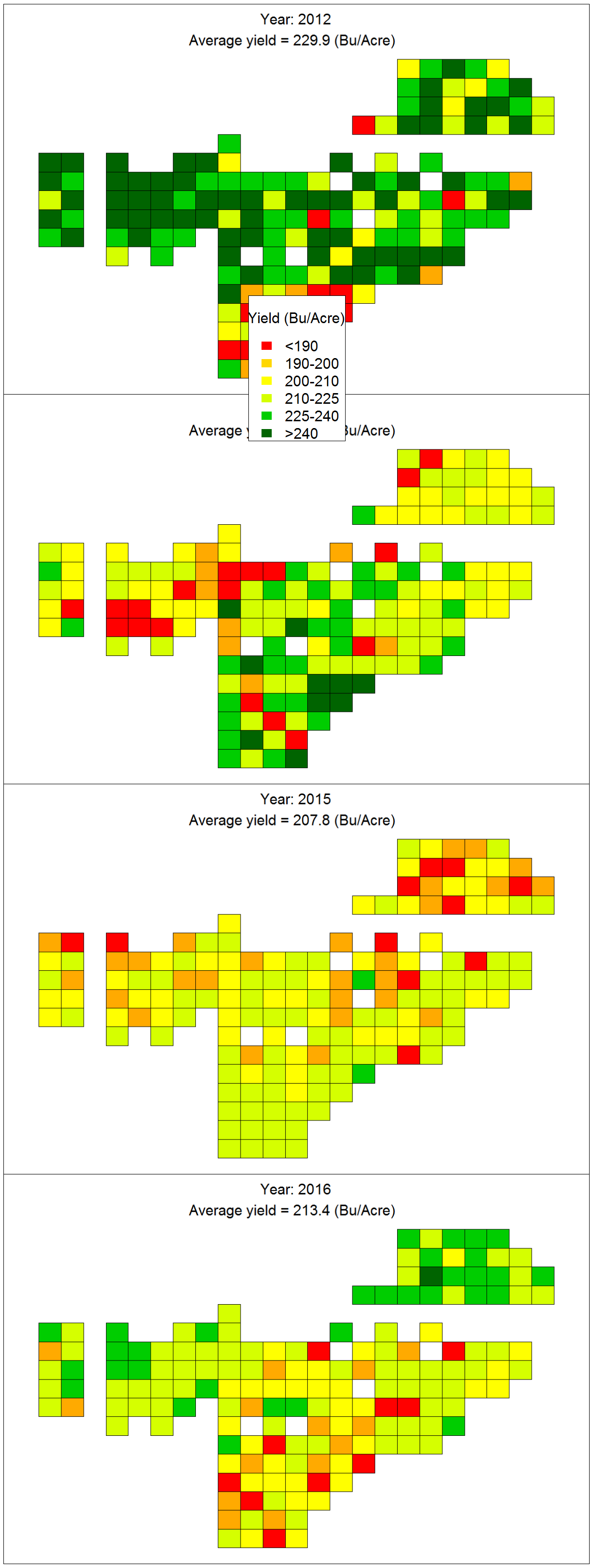

Figure 2. Mapping of grid-level (100 meters) corn yield by year.

[1] In the Bayesian statistics, parameters are random. Therefore, spatially varying coefficients (location specific parameters) assume to follow the Gaussian spatial process. Every location in the process is normally distributed, and the finite collection of those locations follows the multivariate normal distribution. Therefore, the process can be regarded as the stochastic joint distribution of all those locations. The Gaussian spatial process is stationary and isotropic, which means that the covariance of the random variable (i.e., location specific parameters) between any two locations is determined only by Euclidean distance, not by position itself or direction of the distance. We assume that spatially correlated stochastic terms in the model follow the Gaussian spatial process and non-spatially correlated terms follow a general multivariate normal distribution.

[2] The Euclidean distances for the model can be used both from the basic 2-dimensional (longitude-latitude) or 3-dimensional physical spaces (longitude-latitude-altitude). It is also noteworthy that the distances can not only be the distance from the physical space but also be from the soil type space, weather space, etc.

[3] Note that the spatial covariance matrix Σ is $N \times N$ matrix the follows exponential type spatial function, then

$$\Sigma = \rho e^{-D_{ij}/\theta} = \rho \begin{bmatrix} 1 & & e^{-D_{1N}/\theta} \\ \vdots & \ddots & \vdots \\ e^{-D_{N1}/\theta} & & 1 \end{bmatrix}.$$

[4] The location specific coefficients are generated in our MCMC estimation procedure. The model generates K MCMC draws and each random draw creates an $N \times 1$ (N is the number of locations) vector $\boldsymbol{z}$= $[z_1, \ldots, z_N]'$. Note that for any kth MCMC draw, $\boldsymbol{z}_k \sim N(0, 1)$, and therefore $\sum_{i=1}^{N} z_{ik} \approx 0$. Next, the model conducts a Cholesky decomposition for kth covariance matrix based on the kth updated Kriging parameter values, $\Sigma_{Pk}(\rho_{Pk}, \theta_{Pk}) = \boldsymbol{L}_k \boldsymbol{L}_k{}'$, where $\boldsymbol{L}_k$ is a lower triangular matrix. Then using the equation $\boldsymbol{P}_k = \overline{\boldsymbol{P}} + \boldsymbol{L}_k \boldsymbol{z}_k$, the model draws the kth posterior value $\boldsymbol{P}$, say $\boldsymbol{P}_k$, from the kth decomposed covariance matrix $\boldsymbol{L}_k$ and the random draw vector $\boldsymbol{z}_k$. The model then accepts or rejects the draw under the Metropolis-Hastings algorithm. The spatially correlated error term $\boldsymbol{\epsilon}_t$ is updated under the same gaussian spatial process procedure with the plateau parameter.

[5] It is noteworthy that the proposed Kriging regression model can allow the spatially varying intercept (α) and input response parameter (β) as well by using the Gaussian spatial process. We assume the deterministic response here just due to the empirical application.

[6] We use SAS PROC NLMIXED procedure to get estimates for the priors.