Agribusiness and Applied Economics Report No. 813 March 2022

DEVELOPMENT OF FARM MODEL FOR ND and NGP

Prediction of Corn and Soybean Yields in the Presence of Random Shocks

Kwame Asiam Addey, Ph.D. Candidate
Saleem Shaik, Ph.D.
William Nganje, Ph.D.

Department of Agribusiness and Applied Economics
Agricultural Experiment Station
North Dakota State University
Fargo, ND 58108-6050

Acknowledgments

The authors would like to express their gratitude to the North Dakota Soybean Council and North Dakota Soybean Growers' Association for their support, suggestions, comments and discussions during the course of project. All views expressed in this paper are those of the authors and do not reflect the opinions and interest of the supporting organizations or NDSU.

Kwame Asiam Addey, Ph.D. Candidate Center for Agric. Policy & Trade Studies, Dept. of Agribusiness and Applied Econ., North Dakota State University, Fargo, ND-58108, USA

Saleem Shaik, Ph.D., Branch Chief, Agric. Policy and Models Branch, U.S. DA, Economic Research Service, MS 9999, Kansas City, MO 64141-6205, USA

William Nganje, Ph.D. Professor and Chair, Department of Agribusiness and Applied Economics, North Dakota State University, Fargo, ND-58108, USA

Table of Contents

List of Tables

List of Figures

Executive Summary

Rationale of the project

Farm policy is dependent on the ability to predict relevant policy variables based on given indicators. Crop yields and prices are important for farm income stability in North Dakota. Prediction of crop yields and prices are also essential for knowing the impacts of farm programs embedded in the U.S. farm bill. The accuracy of prediction of crop yields and prices is often hampered by the occurrence of random events, e.g. weather, input supply or natural disasters such as Covid19. Inaccurate prediction of yields may lead to under (over) payment of producers in insurance programs and ineffective design of farm policies. Making inferences from such data require modelling the data-generating mechanism of the data.

This report builds on Phase I of the CAPTS – NDSC/NDCUC Project. The first phase presented the trends in yields, prices, revenue and production of corn and soybean. Based on the findings, the proposed future research was to:

- Examine the sources of yield variability across North Dakota counties
- Evaluate acreage price elasticities across U.S. and North Dakota counties
- Examine the sources of price volatilities on international markets

Objective of the report

This part of the study is focused on developing accurate models to predict crop yields. We focus on the variability of crop yields across North Dakota counties. This report proposes the Dirichlet Process Mixture Model (DPMM) as a measure of accounting for randomness in the crop yield data in a Hierarchical Linear Model (HLM).

Mixture models have been widely applied in several areas as a tool for modeling population heterogeneity, allowing for posterior unsupervised classification of the observations. They are also beneficial for prediction under conditions of complicated probabilities, multimodality, skewness and heavy tails. The specific choice of a Dirichlet distribution is a result of its ideal statistical properties for Bayesian analysis. For instance, it is one of the few distributions that are conjugate distributions (the generated posterior is always equal to the known prior distribution). In mixture models, the randomness in the data (generating process) are captured through unsupervised cluster assignments of data with similar characteristics.

Data and methodology used in the research study

The states in the northern great plains are the major producers of corn and soybean. Of the top 15 corn producing states in U.S., 9 are within the great plains while 7 out of the 15 soybean producing states are from the great plains. The National Agricultural Statistics Service (NASS) publishes a wide range of data on U.S. agricultural production at different levels of aggregation (county, state, regional and national) over different periods (monthly, quarterly and annually). Corn and soybean yield at county levels were obtained by year from the NASS database. A longitudinal data set of county-average yield data was analyzed using 104 counties for corn and 66 counties for soybean from 1972 through 2018.

The proposed model (DPMM) employs a Gibbs sampler as a clustering algorithm for inference on the random density functions. The impact of the clusters is examined on the prediction accuracy of corn and soybean yields among northern great plain states.

Key findings

- The Bayes classification of the random density functions signify that the yields for both corn and soybean are drawn from two mixture components.
- Of the 6 states used for the corn analysis, the distribution of yields for North Dakota is statistically identical to the distribution of yields in Kansas and South Dakota.
- For the 5 soybean states, the distribution of yields in North Dakota is statistically similar to the distribution of soybean yields in Kansas.
- The best prediction model for corn yield is the HLM structure that incorporates the Bayes classifier, state, county and year.
- The best prediction model for soybean yield is the HLM specification that incorporates the Bayes classifier, state, crop reporting district, county and year.
- Despite the crop reporting district being the difference between the optimal corn and soybean yield prediction model, the Bayes classifier from the proposed DPMM accurately classifies random events.
- The DPMM (Bayes) classifier is therefore integral for the prediction of both crops.
- It is also noteworthy that, neighbouring states with underlying similarities of characteristics are important in the prediction of North Dakota corn and soybean yields.

Future Research

The findings from this study reveal that it is important to consider the impact of random events on the prediction of crop yields using a Bayesian classification of the data as a layer in the HLM. The report also reveals that drawing observations from neighboring states can help improve the prediction of states yields due to similarity of characteristics among different states. Given this finding, the next phase of this research will evaluate the sources of variability of crop yields based on the proposed model. The specific objectives to be pursued as part of this goal will:

- Evaluate the impact of farm and conservation program payments on ND corn and soybean production.
- Examine the impact of farm fertilizer utilization on greenhouse gas emissions and productivity.

1. Introduction

The evaluation of the impact of policies and risk management strategies is dependent on the knowledge of the probability distributions of the expected output and magnitude of risks encountered. Meanwhile, these risks vary across different sectors, geographical space and over time. Given the agricultural sector, the distribution of soybean yields in the Midwest region are expected to be different from that of Northern great plains region of the U.S. For a specific state such as North Dakota, the distribution of corn or soybean yields from 2000 to 2009 is not likely to be the same as its distribution from 2010 and 2019 (Shaik and Addey, 2020a, b). Hence, there is the need for a constant review of the modelling strategies for the evaluation of agricultural risks and prediction of crop yields across geographical locations over time.

The primary focus of this report is to examine the accuracy of crop yield prediction for longitudinal data in the presence of random events and latent characteristics. There are primarily three schools of thought for predicting crop yields. The first group emphasize on the need to account for temporal shocks through the use of trend models (Harri et al, 2011; Brester et al., 2019). A peculiar challenge with the trend models is that, failure to identify the accurate trend may cause misleading results in crop insurance analysis (Atwood et al., 2003 and Finger, 2013).

The second group believe in the use of spatial models (Liu and Ker, 2020 and Ramsey, 2020). Given that the quality of agricultural land and weather conditions vary across space, spatial prediction models present the best way of accounting for spatial heterogeneity. However, Brady and Irwin (2011) emphasized that data generated from proximal spatial points are spatially dependent and as such econometric models that do not include all the relevant spatial variables will suffer from omitted spatially correlated variables that can bias results. Of course, the high probability of encountering missing yield data (Liu and Ker, 2020) implies that there is an increased likelihood of omitting relevant spatial variables. On the other hand, since spatial dependence will lead to multiple interactions, this may result in multiple predictors which could also lead to overfitting.

Finally, the third group emphasize on accounting for randomness through distributional assumptions (Duarte et al., 2018) or natural clusters in the model (Shaik and Bhattacharjee, 2016). In general, researchers consider cluster models based on the natural clusters when using longitudinal data. This empirical strategy is useful because it accounts for the shared variances among the natural clusters. Despite this advantage, following the natural clusters may lead to an incorrect partitioning of variances and dependencies in the data which may lead to a type I error. In crop yield predictions, it is essential to consider models that rightly classify these variables due to the presence of random unobservable factors.

The present study proposes a hierarchical linear model (HLM) that incorporates a Bayesian classified stratum generated from the Dirichlet process mixture model (DPMM). The DPMM is a Bayesian clustering method that links a response vector to a set of fixed or random covariates through nonparametric regression methods (Malsiner-Walli et al., 2017). The application of Bayesian nonparametric methods has been extensively employed to account for randomness in different fields. For instance, Kleinman and Ibrahim (1998) modelled the random effects in a health data using a Dirichlet process prior. The DPMM not only accounts for randomness and latency associated with temporal, spatial and natural clusters but also allows for posterior unsupervised classification of the observations despite the heterogeneity across counties. Another advantage of employing the DPMM is its robustness for prediction under conditions of complicated probabilities, multimodality, skewness and heavy tails. The model proposed in this report is a simple tractable model which accounts for random events using three processes sequentially. These are;

1. Temporal smoothing using the penalized B-spline.
2. Bayesian clustering of crop yields using the DPMM.
3. Prediction of crop yields using the HLM with the Bayes classification from the DPMM considered as a stratum.

The concept of mixture models is not new in crop yield prediction. The model being employed in this article is similar to that employed by Tolhurst and Ker (2015). The basic concept of their model was to employ a two-component mixture model to account for technological changes and randomness in subpopulations. However, the present study introduces a class variable (level) in the HLM structure using a DPMM. In statistical literature, the proposed model was employed by Bush and MacEachern (1996) in a hierarchical model of randomized block design which allows for the Dirichlet process to be inserted in the middle stage for the distribution of the block effects.

The contribution of this article is three-fold. First, to the best of our knowledge, this is the first study to introduce a DPMM as a hierarchical level to evaluate systematic risk of U.S. crop yield distributions. The results indicate that this model has a potential to improve the prediction of crop yields and the impact of farm policies in the presence of random shocks.

Secondly, rare events such as early blizzards, late planting or hail in certain locations or years may lead to partial/total loss of yields when they occur. However, the frequency of occurrence of these events are often too low to develop a trend or spatial pattern for analysis. On the other hand, certain areas and periods may experience very good cropping conditions leading to high yields. Due to the possibility of extremely low or high yields, policymakers and researchers, particularly in crop insurance studies are often concerned about the effects of outliers and overdispersion of the data. The presence of overdispersion, if unaccounted for may lead to biased estimates of the variance-covariance matrix. Implementation of a DPMM allows clustering of the response variable based on latent functions.

The final contribution of this study emanates from its focus on the northern great plains. Several studies on crop yield predictions have focused on midwestern states such as Ohio, Oklahoma, Illinois, Iowa, Kansas, Minnesota and Nebraska. One state which has not been given much attention in the crop yield prediction and crop insurance literature is North Dakota. Meanwhile, the state is a major contributor to U.S. agricultural production. In addition, the climate within the state implies that crop production is often faced with a lot of risks.

The rest of the paper is organized as follows. Section 2 discusses the empirical application strategy using real U.S. crop yield data. The structure of the HLM with the DPMM Bayes strata is presented in this section. In addition, the sources of data used are also presented in section 2. The results of the estimation are presented in Section 3. The results comprise of the descriptive statistics, Kruskal-Wallis test, Wilcoxon rank sum test for differences in state yield distributions and prediction of corn and soybean yields using alternative HLM structures. This section also discusses the empirical results for the prediction of corn and soybean yields. Section 4 presents the conclusions while section 5 presents the proposed future research.

2. Empirical Application

2.1 Type and sources of data for empirical analysis

The states in the northern great plains are major producers of corn and soybean. Of the top 15 corn producing states in U.S., 9 are within the great plains (Shaik and Addey, 2020c) while 7 out of the top 15 soybean producing states are from the great plains (Shaik and Addey, 2020d). The National Agricultural Statistics Service (NASS) publishes a wide range of data on U.S. agricultural production at different levels of aggregation (county, state, regional and national) over different periods (monthly, quarterly and annually).

The empirical application of this study employs annual corn county yields from Colorado, Kansas, Nebraska, North Dakota, South Dakota and Texas while soybean county yields are drawn from Kansas, Nebraska, North Dakota, Oklahoma and South Dakota. The analysis comprises of purposively selected counties with complete yield series from 1972 to 2018. For corn, the number of counties with complete series were 2, 9, 57, 4, 14 and 18 for Colorado, Kansas, Nebraska, North Dakota, South Dakota and Texas respectively. The number of complete counties series for soybean were 19, 32, 4, 5 and 6 for Kansas, Nebraska, North Dakota, Oklahoma and South Dakota respectively. This gives 104 counties for corn and 66 counties for soybean across all the 47 years. County yields are measured as the total production divided by the planted acreage to capture the tails of the yield distribution.

2.2 Penalized B-spline for temporal smoothing

The idea of smoothing the dataset creates an approximating function that captures important patterns over time in the data while leaving out noise or other fine-scale structures. The penalized B-spline (Eilers and Marx, 1996) is a detrending procedure that fits a smooth curve through a scatter plot with an automatic selection of the smoothing parameter. To account for temporal variations in the yields, we employ the penalized B-spline.

2.3 Hierarchical linear model with Dirichlet process mixture model Bayes strata

The HLM is the primary tool of multilevel analysis. This method allows for examining data with nested sources of variability. As a first step, we derive the cluster of discrete random densities using the DPMM. Following this step, the generated Bayes classification of random densities are incorporated into the HLM as a level. The proposed HLM model uses data comprising of two levels, i.e. state and county. This implies that the addition of an extra level using the Bayes classifier to account for randomness leads to a three-level HLM. Data structured in such format are often estimated based on the general linear model (Addey, 2021).

For the northern great plains crop yield data, consider a set of $t \times 1$ vector of observed responses $\boldsymbol{y}_{ijkt} = (y_{ijk1}, \dots, y_{ijkt})'$ for the ith county for $i = 1, \dots, a_i$ within the jth state for $j = 1, \dots, b_j$ in the kth Bayes class for k $= 1, \dots,$ c in time t for $t = t_{ijk1}, \dots, t_{ijkt}$, with $t_{ijk1} < \dots < t_{ijkt} < \cdots < T$. This yields a standard linear mixed model given as;

$$\boldsymbol{y}_{ijkt} = \boldsymbol{\mu} + \boldsymbol{\beta}_{\text{ijk}}\boldsymbol{X'}_{ijk} + \boldsymbol{\gamma}_{ijk}\boldsymbol{Z'}_{ijk} + \boldsymbol{\varepsilon}_{ijkt} \quad \begin{cases} \text{i} = 1, \dots, \text{a} \;\; \text{County} \\ \text{j} = 1, .., \text{b} \;\; \text{State} \\ \text{k} = 1, \dots, \text{c} \;\; \text{Bayes Class} \\ \text{t}_{\text{ijk1}} = 1, .. \text{t}_{\text{ijkt}} \;\; \text{Year} \end{cases} \tag{3.1}$$

where $\boldsymbol{\mu} = (\mu_1 \dots \mu_t)'$ is a vector of intercept parameters, $\boldsymbol{\beta}_{\text{ijk}}$ are the p-dimensional vectors of time-varying functional coefficients which are heterogeneous over i, j and k. $\boldsymbol{Z'}_{ijk}$ represents the latent individual specific random effects for county i in state j within Bayes class k characterized by random coefficients $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_t)'$ and $\boldsymbol{\varepsilon}_{ijkt} = (\boldsymbol{\varepsilon}_{ijk1}, \dots, \boldsymbol{\varepsilon}_{ijkT})'$ is a vector of independently distributed measurement errors measuring idiosyncratic term with mean zero. The primary goal of this model is to examine the implication of latent response distribution of the groups on prediction of the dependent variable (yield). It is typically common to see that most studies address this issue through a mean regression model (Dunson, 2006), in which $E(\boldsymbol{Z}_{ijk}) = \boldsymbol{\beta}_{\text{ijkt}}\boldsymbol{X'}_{ijkt}$ and $V(\boldsymbol{Z}_{ijk}) = 1$. The variance of the latent variable density is fixed at 1 for identifiability. In this report, the variance of the outcome variable conditional on the random coefficients is given as $V(\boldsymbol{y}_{ijkt} | \boldsymbol{\gamma}_{ijk}) = g(\boldsymbol{\mu}) * \boldsymbol{\phi}$, where g is a monotonic differentiable link function which describes how the expected value of $\boldsymbol{y}_{ijkt}$ is related to $\boldsymbol{\mu}$**;** and $\boldsymbol{\phi}$ is the dispersion parameter (either known or estimated). Hence, equation (3.1), is re-written to exclude the fixed covariates to give;

$$\boldsymbol{y}_{ijkt} = \boldsymbol{\mu} + \boldsymbol{\gamma}_{ijk}\boldsymbol{Z'}_{ijk} + \boldsymbol{\varepsilon}_{ijkt} \tag{3.2}$$

Considering the three different levels in this HLM, the empirical model specification is;

$$\boldsymbol{y}_{ijkt} = \boldsymbol{\gamma}_1 County_{ijk} + \boldsymbol{\gamma}_2 State_{ij} + \boldsymbol{\gamma}_3 Bayes_k + \boldsymbol{\gamma}_4 Year_{ijkt} + \boldsymbol{\varepsilon}_{ijkt} \tag{3.3}$$

3. Results and Discussions

3.1 Descriptive statistics

The results of the empirical application are presented in this section. To validate the robustness of the proposed model, a range of models are estimated and compared based on their goodness of fit diagnostics. The descriptive statistics of corn and soybean yields for the states are presented in Table 1. From the table, Colorado has the highest average corn yield of 152.67bu/acre followed by Nebraska with 132.08bu/acre. Following these are Kansas and North Dakota with 103.21bu/acre and 100.84bu/acre respectively. The next is South Dakota with average corn yield of 97.21bu/acre and the highest standard deviation of 44.17bu/acre. The state with the least average corn yield was Texas with 88.90bu/acre. The least standard deviation was observed for Colorado with 29.16bu/acre.

From Table 1, the highest average soybean yield was observed for Nebraska with 41.62bu/acre, followed by South Dakota with 34.15bu/acre. The next highest average soybean yield was observed for Kansas and North Dakota, with 30.77bu/acre and 29.19bu/acre respectively. The least average soybean yield was 23.40bu/acre for Oklahoma. The standard deviations about the average yields were fairly distributed with Nebraska having the highest of 11.35bu/acre while the least was observed for Oklahoma with 7.34bu/acre.

Table 1: Descriptive statistics of corn and soybean yield by state

State name	N	Mean	Std Dev	Min	Max
		Corn			
Colorado	96	152.67	29.16	92.00	218.50
Kansas	431	103.21	34.43	23.00	196.00
Nebraska	2736	132.08	37.38	18.40	228.50
North Dakota	192	100.84	37.70	36.00	207.60
South Dakota	672	97.21	44.17	1.500	194.50
Texas	864	88.90	42.02	15.00	233.20
		Soybean			
Kansas	912	30.77	10.36	8.00	62.50
Nebraska	1536	41.62	11.35	15.40	72.20
North Dakota	192	29.19	8.14	9.20	48.30
Oklahoma	240	23.40	7.34	5.60	44.50
South Dakota	288	34.15	10.13	4.30	60.40

3.2 Kruskal-Wallis and Wilcoxon rank-sum test for differences in state yield distributions

The box plot for the distributional characteristics of the corn and soybean yields are presented in Figure 1(a and b). For corn, a similar median is observed for Kansas, North Dakota, South Dakota and Texas while it is similar for Kansas and North Dakota for soybean. Despite the similarities, we see varying quartile ranges for the various states. The use of the DPMM is beneficial over the gaussian counterparts when the natural clusters of the data set are not normally distributed.

Figure 1: Box plot of corn and soybean yield by state

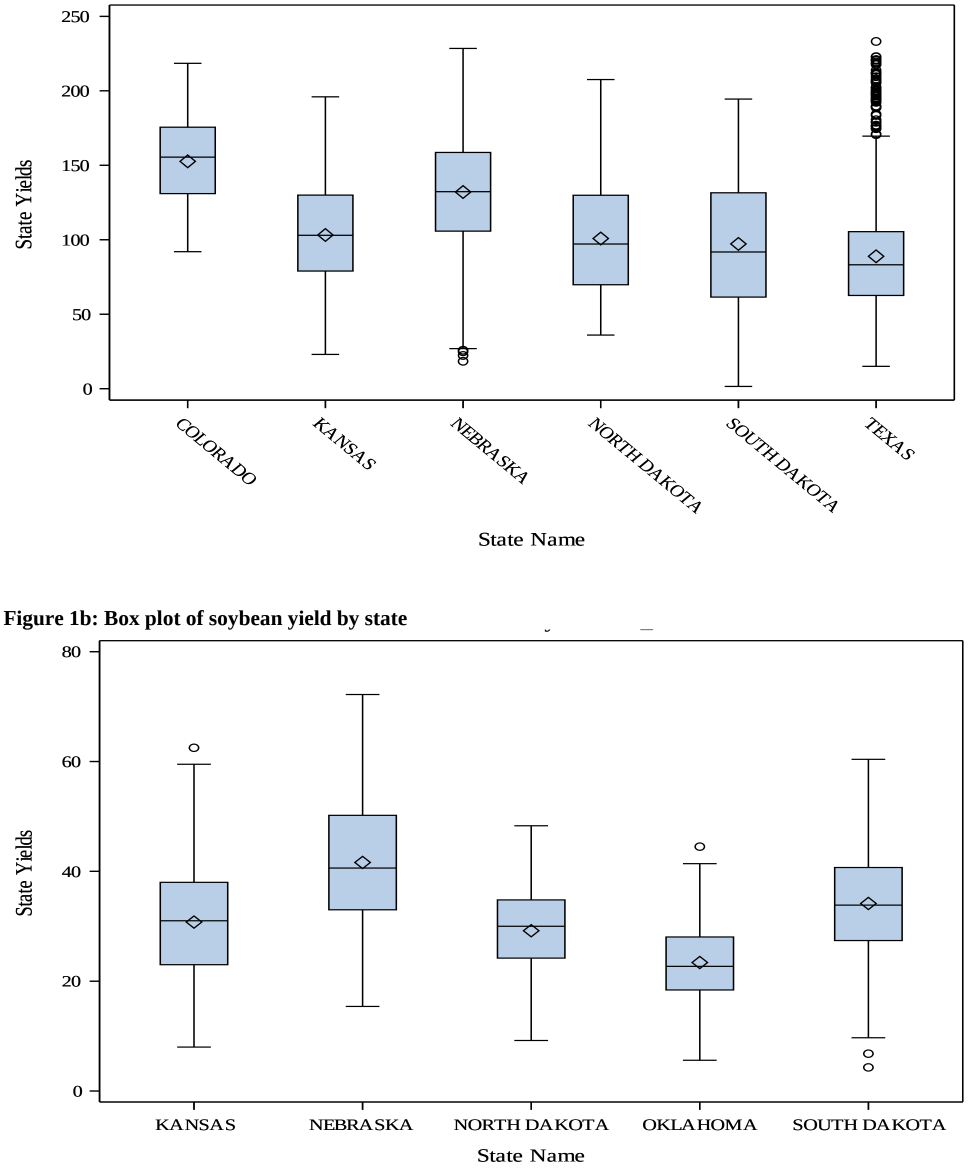


Figure 1a: Box plot of corn yield by state

Figure 1b: Box plot of soybean yield by state

Visual observations using kernel distributions for corn yields show Colorado, Kansas, North Dakota, South Dakota and Texas to have non-normal yields. The kernel distributions of corn yield for the 6 states are presented in Figure 2 (a to f). For soybean yield, the kernel distribution for Kansas and Oklahoma yield reveal a near normal distribution. However, the distributions of soybean yield for Nebraska, North Dakota and South Dakota reveal non-normality. The kernel distributions of soybean yield for these states are presented in Figure 3 (a to e).

A basic requirement for the need for mixture models is the existence of differences among the different levels (states) of the dataset. To proof that the distributions are from different sources and hence affected by varying random factors, a Kruskal-Wallis nonparametric test was conducted. The chi square value is significant at 1% for both crops, implying that the distributions of the crop yields from the states are statistically different from each other for both crops. For corn and soybean yield, the chi square probability values in Table 2 confirms the differences in yield distributions among the states.

It is typical to observe the use of a t-test for pairwise comparison of subgroups in a longitudinal dataset. However, the standard t-test is useful under conditions when the distributions of the subpopulations are assumed to be known. Under conditions of latency where the distribution of the subgroups is assumed to be unknown, it is useful to employ the Wilcoxon rank sum test. For pairwise comparison of the differences in yield distribution among the states, a two-sample Wilcoxon rank sum test was performed. Table 3 presents the results for corn and soybean. From the table, the test of equality between the distributions of corn yield for Kansas and North Dakota reveals a probability value of 0.8151 which implies that there is no statistical difference between the yield distributions from these two states. In addition, corn yield distribution for North Dakota is revealed to be statistically similar to the yield distribution of corn from South Dakota, having a probability value of 0.8441. From the table, the probability value for the equality of distribution for Kansas and North Dakota soybean yield is 0.3547, indicating that there is no statistical difference between the distribution of the soybean yield from these two states.

Figure 2: Distribution of corn yield by state

Figure 2a: Distribution of Colorado corn yield

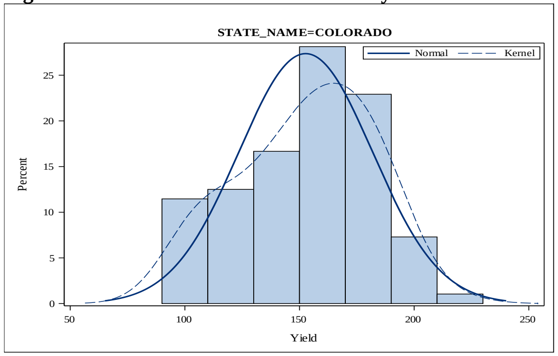


Figure 2b: Distribution of Kansas corn yield

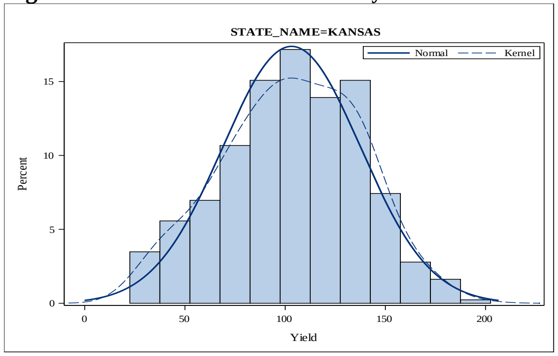


Figure 2c: Distribution of Nebraska corn yield

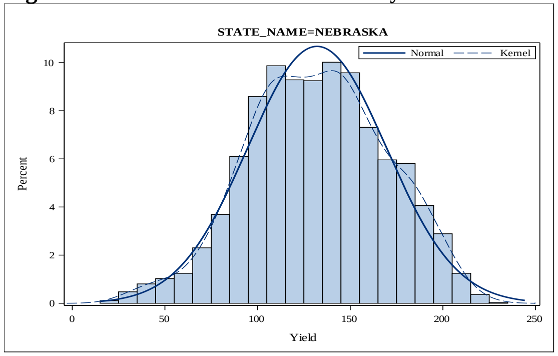


Figure 2d: Distribution of North Dakota corn yield

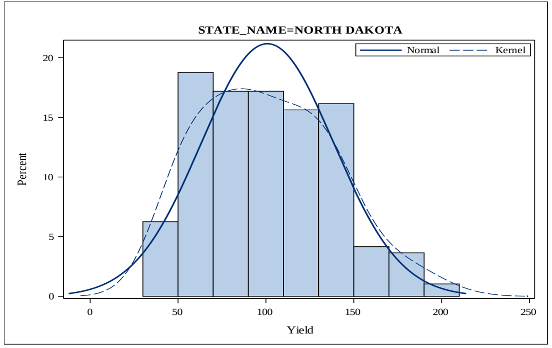


Figure 2e: Distribution of South Dakota corn yield

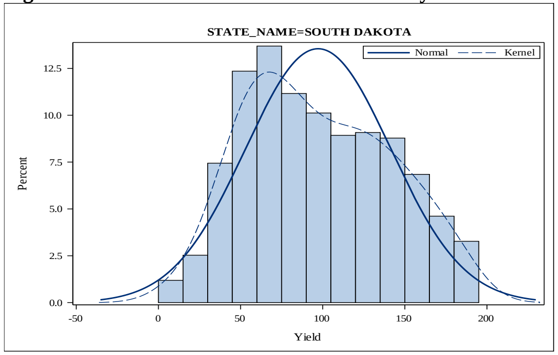


Figure 2f: Distribution of Texas corn yield

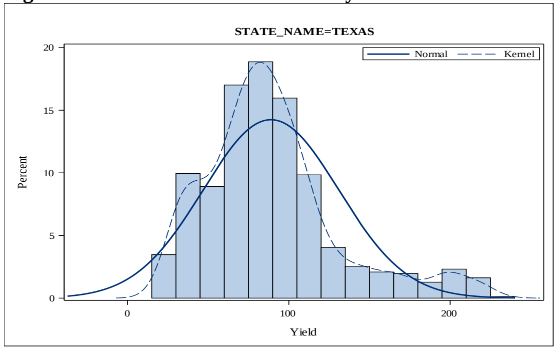


Figure 3: Distribution of soybean yield by state

Figure 3a: Distribution of Kansas soybean yield

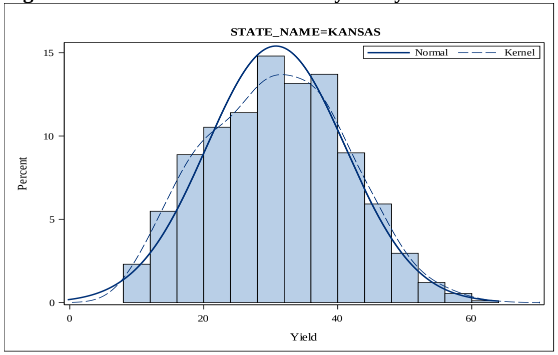


Figure 3b: Distribution of Nebraska soybean yield

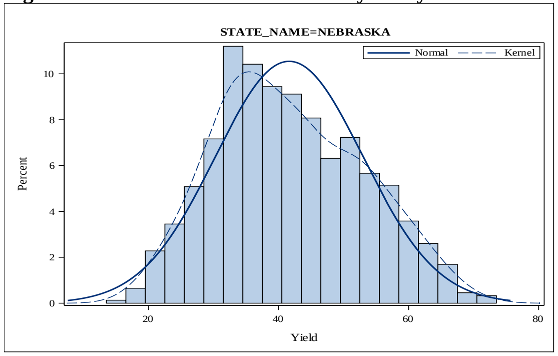


Figure 3c: Distribution of North Dakota soybean yield

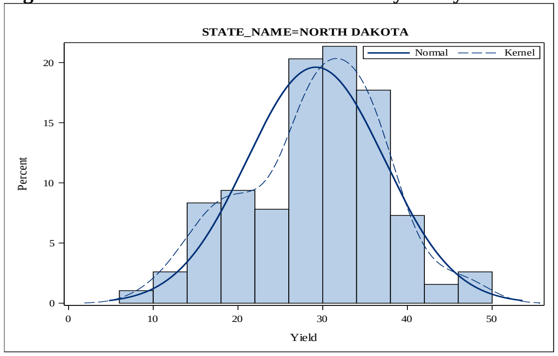


Figure 3d: Distribution of Oklahoma soybean yield

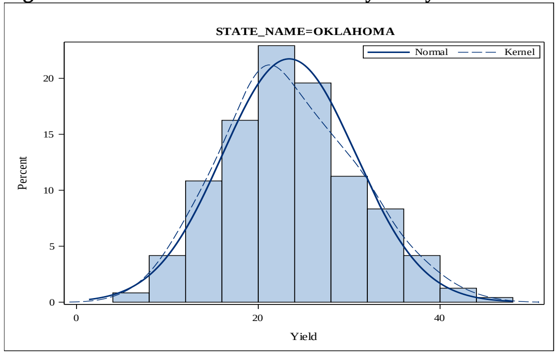


Figure 3e: Distribution of South Dakota soybean yield

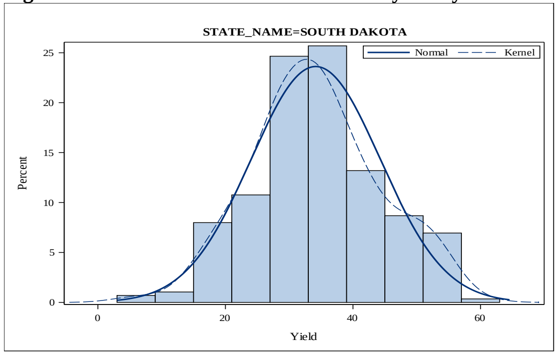


Table 2: Test of equality of corn and soybean yield distributions across states

State name	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Corn					
Colorado	96	358534.0	239616.0	13981.65	3734.73
Kansas	431	884522.0	1075776.0	28593.51	2052.26
Nebraska	2736	8264941.5	6829056.0	50661.47	3020.81
North Dakota	192	373969.5	479232.0	19578.18	1947.76
South Dakota	672	1263325.5	1677312.0	34747.43	1879.95
Texas	864	1312243.5	2156544.0	38514.17	1518.80
Kruskal-Wallis Test					
Chi-Square		1022.77	**DF** 5	**Pr > ChiSq**	<.0001
State name	**N**	**Sum of Scores**	**Expected Under H0**	**Std Dev Under H0**	**Mean Score**
Soybean					
Kansas	912	1125736.0	1445064.0	23309.38	1234.36
Nebraska	1536	3106209.0	2433792.0	25728.79	2022.27
North Dakota	192	209130.5	304224.0	12283.75	1089.22
Oklahoma	240	150714.5	380280.0	13622.44	627.98
South Dakota	288	427906.0	456336.0	14799.81	1485.79
Kruskal-Wallis Test					
Chi-Square		807.6401	**DF** 4	**Pr > ChiSq**	<.0001

Table 3: Pairwise comparison of corn and soybean yield among states

State name	Wilcoxon Z	DSCF Value	Pr > DSCF
Corn			
Colorado vs. Kansas	10.9958	15.5504	<.0001
Colorado vs. Nebraska	5.4972	7.7742	<.0001
Colorado vs. North Dakota	9.8145	13.8798	<.0001
Colorado vs. South Dakota	10.7448	15.1955	<.0001
Colorado vs. Texas	12.5745	17.7830	<.0001
Kansas vs. Nebraska	-13.9913	19.7866	<.0001
Kansas vs. North Dakota	1.2440	1.7593	0.8151
Kansas vs. South Dakota	2.9175	4.1260	0.0412
Kansas vs. Texas	8.3265	11.7754	<.0001
Nebraska vs. North Dakota	10.3221	14.5977	<.0001
Nebraska vs. South Dakota	17.8336	25.2205	<.0001
Nebraska vs. Texas	26.5773	37.5859	<.0001
North Dakota vs. South Dakota	1.1852	1.6761	0.8441
North Dakota vs. Texas	4.4797	6.3353	0.0001
South Dakota vs. Texas	3.9551	5.5933	0.0011
Soybean			
Kansas vs. Nebraska	-20.7009	29.2755	<.0001
Kansas vs. North Dakota	1.8324	2.5915	0.3547
Kansas vs. Oklahoma	10.0801	14.2554	<.0001
Kansas vs. South Dakota	-4.4668	6.3170	<.0001
Nebraska vs. North Dakota	13.7342	19.4231	<.0001
Nebraska vs. Oklahoma	20.6888	29.2583	<.0001
Nebraska vs. South Dakota	9.7564	13.7976	<.0001
North Dakota vs. Oklahoma	7.2912	10.3113	<.0001
North Dakota vs. South Dakota	-5.1121	7.2296	<.0001
Oklahoma vs. South Dakota	-12.0786	17.0817	<.0001

The four counties with full observations of corn yield from North Dakota over the study period are Cass, Grand Forks, Richland and Sargent. In Figure 4(a and b), we present visual distributions of their actual yield compared to the trend of yield smoothened based on the penalized B-spline. We observe an upward trend for all four counties. In addition, most of the yields are near the penalized B-spline trend line with few further away for Richland and Sargent counties. For soybean, the counties with complete yield data over the study period are Cass, Richland, Sargent and Traill. A comparison of the actual yield distribution to the penalized B-spline trend line is presented in Figure 5. From the figure, there are upward trends for the four counties. The observed dispersion about the penalized B-spline trend is fairly even despite Cass, Sargent and Traill having a few observations further away from the trend line.

Figure 4: Comparison of smoothened corn yield to actual yield by ND counties

Figure 4a: Smoothened versus actual corn yield for ND counties

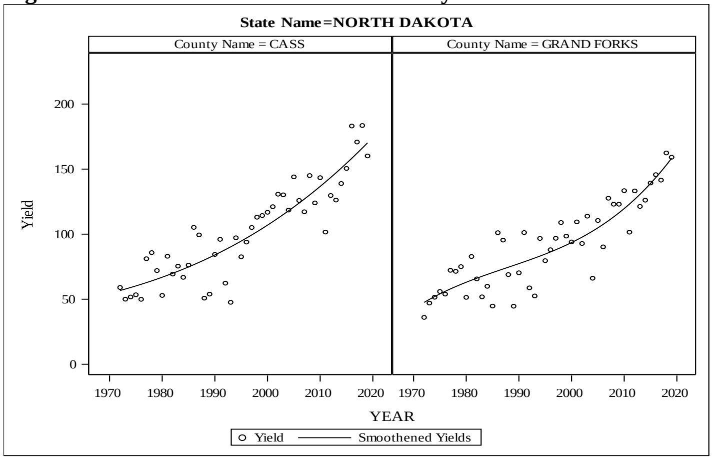


Figure 4b: Smoothened versus actual corn yield for ND counties

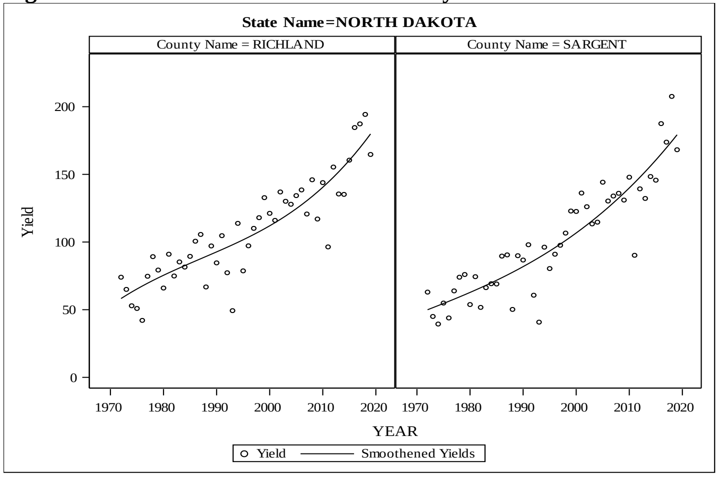
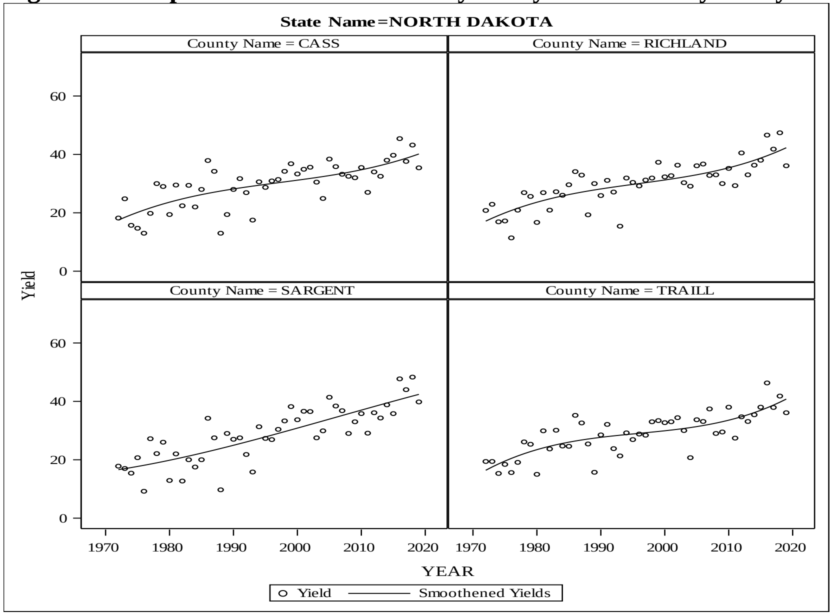


Figure 5: Comparison of smoothened soybean yield to actual yield by ND counties



3.3 Robustness check using alternative hierarchical linear model structures

A set of alternative models were estimated in addition to the proposed model to measure comparative robustness. These models are presented in Table 4. To evaluate the performance of the proposed model, 12 HLM structures were implemented and compared. The differences in the structures are the levels considered in the HLM and whether the data is smoothened or not.

Model one uses the unadjusted/actual yield as a function of the Bayes classifier and year. Model two employs the actual yield as a function of the Bayes classifier and the individual years. Model three is the unadjusted yield as a function of state, county nested in state and individual years. Model four evaluates the actual yield as a function of the state, crop reporting district nested in state, the county nested in the crop reporting district and individual years. Model five considers the actual yield as a function of the Bayes class, state, crop reporting district nested in state, county nested in the Bayes classifier and individual years. Model six is the actual yield as a function of the Bayes class, state nested in the Bayes class, county nested in the Bayes class and individual years.

Model seven uses the adjusted/smoothened yield as a function of the Bayes classifier and year. Model eight employs the adjusted yield as a function of the Bayes classifier and the individual years. Model nine is the adjusted yield as a function of state, county nested in state and individual years. Model ten evaluates the smoothened yield as a function of the state, crop reporting district nested in state, the county nested in the crop reporting district and individual years. Model eleven considers the adjusted yield as a function of the Bayes class, state, crop reporting district nested in state, county nested in the Bayes classifier and individual years. Model twelve is the smoothened yield as a function of the Bayes class, state nested in the Bayes class, county nested in the Bayes class and individual years.

Table 4: Structure of comparative models

Model	Type of Data	Bayes Class	State	CRD	County	Year
Model one	Actual yield	Yes	No	No	No	Yes
Model two	Actual yield	Yes	No	No	No	Yes[1]
Model three	Actual yield	No	Yes	No	Yes	Yes
Model four	Actual yield	No	Yes	Yes	Yes	Yes
Model five	Actual yield	Yes	Yes	Yes	Yes	Yes
Model six	Actual yield	Yes	Yes	No	Yes	Yes
Model seven	Smoothened yield	Yes	No	No	No	Yes
Model eight	Smoothened yield	Yes	No	No	No	Yes[2]
Model nine	Smoothened yield	No	Yes	No	Yes	Yes
Model ten	Smoothened yield	No	Yes	Yes	Yes	Yes
Model eleven	Smoothened yield	Yes	Yes	Yes	Yes	Yes
Model twelve	Smoothened yield	Yes	Yes	No	Yes	Yes

Table 5: Comparison and selection of HLM structures

Model	-2log Likelihood	AIC	AICC	BIC	Pearson Statistic	Unscaled Pearson Chi sq.
Corn						
Model 1	48678.9	48686.9	48686.9	48713	4990.9	5029622
Model 2	48089.8	48189.8	48190.9	48515.6	4990.4	4469680
Model 3	43224	43528	43537.6	44518.3	4992.2	1686047
Model 4	43224	43528	43537.6	44518.3	4990.9	1686047
Model 5	42458.5	42894.5	42914.5	44314.9	4996	1446311
Model 6	42454.2	42896.2	42916.8	44336.1	4986.3	1445072
Model 7	46855.5	46863.5	46863.5	46889.6	4992	3484515
Model 8	46767.3	46867.3	46868.4	47193.1	4991.5	3423496
Model 9	38520.3	38824.3	38833.9	39814.7	4992	656136
Model 10	38520.3	38824.3	38833.9	39814.7	4992	656136
Model 11	36032.3	36468.3	36488.3	37888.7	4991.3	398602
Model 12	**35999.4**	**36441.4**	**36461.9**	**37881.3**	**4993.1**	**395982**
Soybean						
Model 1	23324.7	23332.7	23332.7	23356.9	3168	292306
Model 2	22600.3	22700.3	22701.9	23003.3	3167.8	232561
Model 3	19387.8	19615.8	19624.4	20306.7	3168	84360.7
Model 4	19387.8	19615.8	19624.4	20306.7	3167.9	84360.7
Model 5	19069.7	19343.7	19356.2	20174	3168.2	76301.7
Model 6	19082.9	19354.9	19367.2	20179.2	3167.8	76621.7
Model 7	21710.8	21718.8	21718.8	21743	3168	175629
Model 8	21675.4	21775.4	21777	22078.4	3162.2	173678
Model 9	15538.4	15766.4	15775	16457.4	3168	25028.8
Model 10	15538.4	15766.4	15775	16457.4	3168.1	25028.8
Model 11	**14476.5**	**14750.5**	**14763**	**15580.9**	**3167.9**	**17900.6**
Model 12	14531.6	14803.6	14815.9	15627.9	3168	18214.6

[1] In this model, the impact of the individual years is considered rather than all years together, as in model one.

[2] In this model, the impact of the individual years is considered rather than all years together, as in model one.

3.4 Empirical results for the prediction of corn yield

Based on the unsupervised Bayesian clustering using the DPMM, a two-component mixture was found as the optimal mixture component for county corn yields. We specified 1000 burn-in samples and 10000 samples after burn-in for all models used. A separate trace panel is produced for each sampled parameter. The panels for North Dakota specific slopes and intercepts are presented in Figure 6. There is a good mixing in the chains. The modest autocorrelation diminishes after about 10 successive samples for component 1 of the intercept. For North Dakota, the autocorrelation diminishes after 5 successive samples for both component 1 and 2. The trace and density plots signify successful convergence in the models. The density plots reveal a smooth and unimodal density function which implies that the samples provide a good representation of the posterior distribution. The specific slopes for other states can be found in the supplementary files.

The model diagnostics for the 12 corn models are presented in Table 5. The diagnostics used are the AIC and BIC (results revealed are consistent for both diagnostics). The least AIC found was for model 12 (36,441.4) while its BIC was also 37,881.3. This implies that corn yield is most accurately predicted using model 12. This model consists of the smoothened yield as a function of the Bayes classifier to account for random events, state nested in the Bayes classifier, county yields nested in the Bayes classifier and the individual years. The structure of this model suggests the importance of considering random events at each hierarchical structure level.

Based on the prediction of corn yields from model 12, the predicted average yield for Colorado was 152.67bu/acre. The predicted average for Kansas and Nebraska were 103.12bu/acre and 132.07bu/acre. Using the same model, the predicted corn yields for North Dakota and South Dakota were 100.84bu/acre and 97.27bu/acre while that of Texas was 88.89bu/acre. Even though the predicted average yields for the optimal HLM structure were equivalent to predicted mean yields for models 3, 4, 5, 6, the smoothened and actual yields, the dispersion about the means were found to be different. The standard deviation (minimum, maximum) of model 12 for Colorado, Kansas and Nebraska were 22.54 (103.67, 195.91), 29.31 (35.35, 163.68) and 31.30 (51.95, 212.43) respectively. For North Dakota, South Dakota and Texas, their respective values were 31.15 (46.27, 168.15), 36.54 (19.01, 173.43) and 40.44 (18.95, 211.75). Table 6 presents the descriptive statistics of the predicted models.

Figure 6: Markov chain Monte Carlo diagnostics for ND corn yield

Figure 6a: Trace panels for intercept of component one for North Dakota corn yield

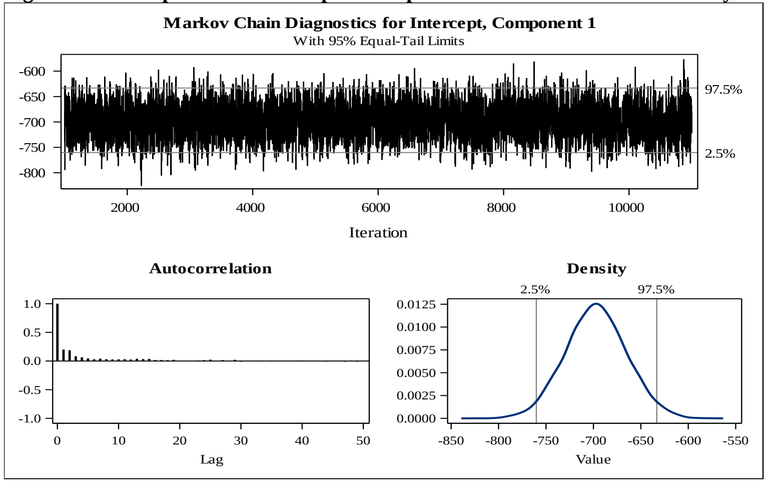


Figure 6b: Trace panels for slope of component one for North Dakota corn yield

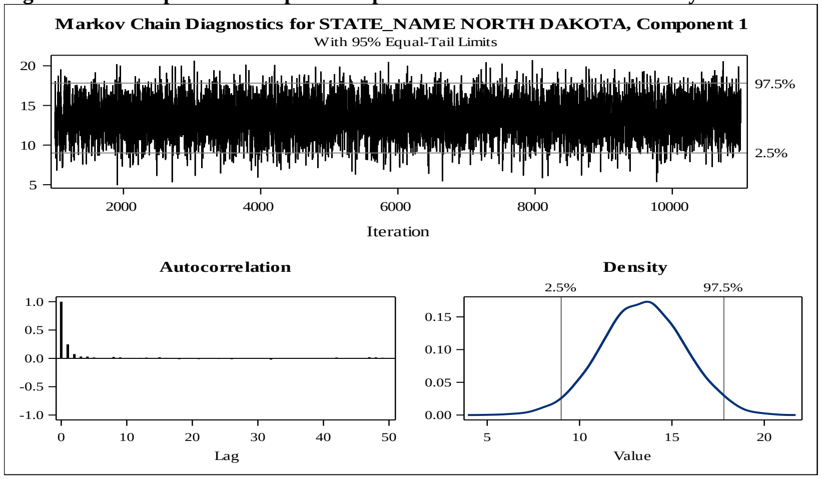


Figure 6c: Trace panels for intercept of component two for North Dakota corn yield

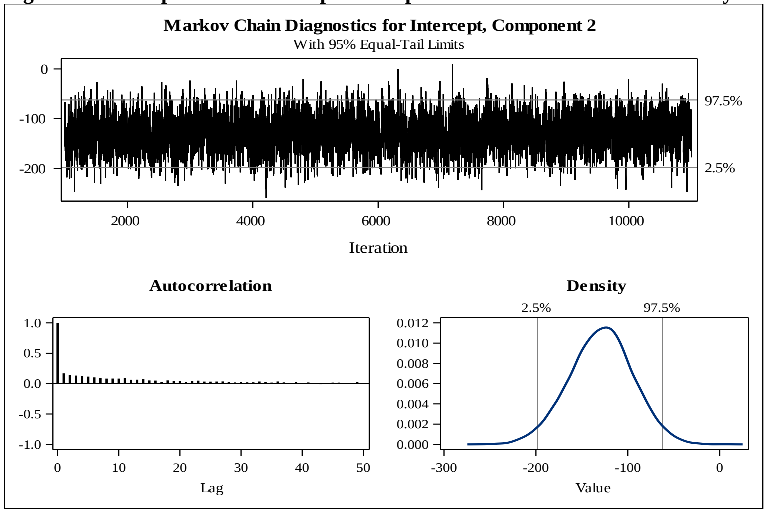


Figure 6d: Trace panels for slope of component two for North Dakota corn yield

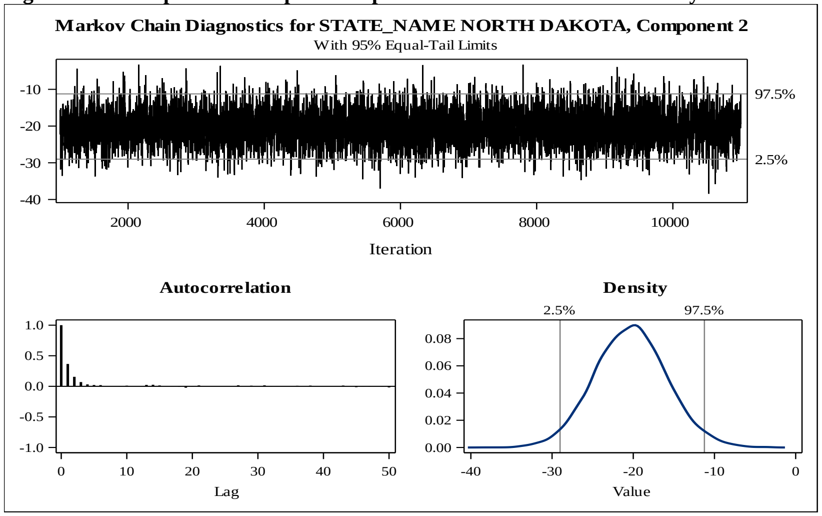


Table 6: State-wise prediction for all corn yield models

PREDICTED MODEL	**COLORADO**					**KANSAS**					**NEBRASKA**				
	N	Mean	Std Dev	Min.	Max.	N	Mean	Std Dev	Min.	Max.	N	Mean	Std Dev	Min.	Max.
Model 1	96	109.36	18.14	78.75	139.97	432	114.50	24.42	78.75	182.69	2736	116.68	30.35	78.75	182.69
Model 2	96	108.83	20.58	59.36	139.33	432	114.35	27.08	59.36	185.16	2736	116.69	31.94	59.36	185.16
Model 3	96	152.67	28.71	88.77	206.49	432	103.04	33.73	19.67	178.59	2736	132.08	31.76	37.05	213.08
Model 4	96	152.67	28.71	88.76	206.49	432	103.04	33.73	19.67	178.59	2736	132.07	31.76	37.04	213.06
Model 5	96	152.67	24.82	92.44	194.51	432	103.05	31.77	23.43	168.52	2736	132.07	32.90	40.72	210.85
Model 6	96	152.67	24.80	92.48	194.53	432	103.05	31.31	23.47	165.82	2736	132.08	32.86	40.76	210.78
Model 7	96	109.30	18.08	78.79	139.81	432	114.48	24.45	78.79	182.88	2736	116.68	30.41	78.79	182.88
Model 8	96	108.77	17.89	70.51	135.82	432	114.33	25.35	70.51	181.98	2736	116.69	30.40	70.51	181.98
Model 9	96	152.67	26.88	99.91	208.81	432	103.12	32.13	31.59	180.91	2736	132.07	30.14	48.18	215.40
Model 10	96	152.67	26.88	99.91	208.81	432	103.12	32.13	31.59	180.91	2736	132.07	30.13	48.18	215.39
Model 11	96	152.67	22.57	103.62	195.93	432	103.12	30.01	35.30	170.21	2736	132.07	31.34	51.90	212.44
Model 12	**96**	**152.67**	**22.54**	**103.67**	**195.91**	**432**	**103.12**	**29.31**	**35.35**	**163.68**	**2736**	**132.07**	**31.30**	**51.95**	**212.43**
Smoothened Yield	**96**	**152.67**	**26.24**	**93.43**	**190.80**	**432**	**103.12**	**25.81**	**51.67**	**160.79**	**2736**	**132.07**	**33.78**	**51.16**	**224.41**
Actual Yield	**96**	**152.67**	**29.16**	**92.00**	**218.50**	**431**	**103.21**	**34.43**	**23.00**	**196.00**	**2736**	**132.07**	**37.38**	**18.40**	**228.50**
PREDICTED MODEL	**NORTH DAKOTA**					**SOUTH DAKOTA**					**TEXAS**				
	N	Mean	Std Dev	Min.	Max.	N	Mean	Std Dev	Min.	Max.	N	Mean	Std Dev	Min.	Max.
Model 1	192	117.15	31.54	78.75	182.69	672	120.42	33.80	78.75	182.69	864	115.10	24.65	78.75	182.69
Model 2	192	117.18	32.64	59.36	185.16	672	120.70	35.62	59.36	185.16	864	114.98	27.37	59.36	185.16
Model 3	192	100.85	28.48	34.32	155.47	672	97.21	32.31	4.11	167.72	864	88.89	43.75	4.05	217.63
Model 4	192	100.84	28.48	34.32	155.46	672	97.21	32.31	4.11	167.72	864	88.89	43.75	4.02	217.62
Model 5	192	100.85	31.77	35.16	163.22	672	97.21	37.11	7.78	171.78	864	88.89	41.73	7.74	211.16
Model 6	192	100.84	31.42	35.21	163.15	672	97.21	37.47	7.82	171.77	864	88.89	41.82	7.77	211.18
Model 7	192	117.15	31.61	78.79	182.88	672	120.45	33.89	78.79	182.88	864	115.08	24.69	78.79	182.88
Model 8	192	117.19	31.36	70.51	181.98	672	120.73	34.13	70.51	181.98	864	114.97	25.48	70.51	181.98
Model 9	192	100.84	26.65	45.46	157.78	672	97.21	30.71	15.24	170.04	864	88.89	42.58	15.19	219.94
Model 10	192	100.84	26.65	45.46	157.78	672	97.21	30.71	15.24	170.04	864	88.89	42.58	15.17	219.94
Model 11	192	100.84	31.33	46.22	168.16	672	97.21	36.05	18.96	173.41	864	88.89	40.35	18.90	211.75
Model 12	**192**	**100.84**	**31.15**	**46.27**	**168.15**	**672**	**97.21**	**36.54**	**19.01**	**173.43**	**864**	**88.89**	**40.44**	**18.95**	**211.75**
Smoothened Yield	**192**	**100.84**	**34.03**	**47.55**	**179.87**	**672**	**97.21**	**39.86**	**21.72**	**183.11**	**864**	**88.89**	**37.69**	**14.98**	**207.10**
Actual Yield	**192**	**100.84**	**37.69**	**36.00**	**207.60**	**672**	**97.21**	**44.17**	**1.50**	**194.50**	**864**	**88.89**	**42.02**	**15.00**	**233.20**

3.5 Empirical results for the prediction of soybean yield

The panels for soybean yield of North Dakota specific slopes and intercepts are presented in Figure 7. There is a good mixing in the Markov chain Monte Carlo. The modest autocorrelation diminishes after about 10 successive samples for component 1 of the intercept. For North Dakota, the autocorrelation diminishes after 5 successive samples for both component 1 and 2. The trace plot and the density plots signify successful convergence in the models. The density plots reveal a smooth and unimodal density function which implies that the samples provide a good representation of the posterior distribution. The supplementary files contain the diagnostics for the other states for soybean yields.

The model diagnostics for the 12 soybean models are presented in Table 5. The diagnostics used are the AIC and BIC. We find the results to be consistent for both goodness of fit indicators. The least AIC is found for model 11 with 14750.5 while its BIC is also the smallest (15580.9). This implies that soybean yields are most accurately predicted using model 11. This model consists of the smoothened yield as a function of the Bayes classifier to account for random events, state nested in the Bayes classifier, crop reporting districts nested in the states, county yields nested in the crop reporting district and the individual years. The structure of this model suggests that it is important to consider the impact of random events at each hierarchical structure level.

Using the prediction from model 11, the predicted average soybean yield for Kansas is 30.76bu/acre. The predicted averages for Nebraska and North Dakota are 41.62bu/acre and 29.19bu/acre for model 11 while Oklahoma and South Dakota had predicted soybean yields of 23.40bu/acre and 34.15bu/acre. This predicted average for the optimum model is equal to the predicted average yields from models 3,4,5,6, 9,10, 12, the smoothened and actual yields for all the states. However, discussions on crop insurance are more concerned about dispersion of risk (standard deviation, minima, maxima and tails). Hence, the importance of the optimum model is evident in the predicted dispersion parameters. It can be seen that the standard deviation for model 11 is 7.65bu/acre, 9.08bu/acre, 6.97bu/acre, 6.92bu/acre and 7.23bu/acre for Kansas, Nebraska, North Dakota, Oklahoma and South Dakota respectively. Among the predicted standard deviations, that of model 12 for Nebraska, North Dakota, Oklahoma and South Dakota are equal to that of model 11. However, their minima and maxima are different. The predictions based on the comparative models for soybean yields are presented in Table 7.

Figure 7: Markov chain Monte Carlo diagnostics for ND soybean yield

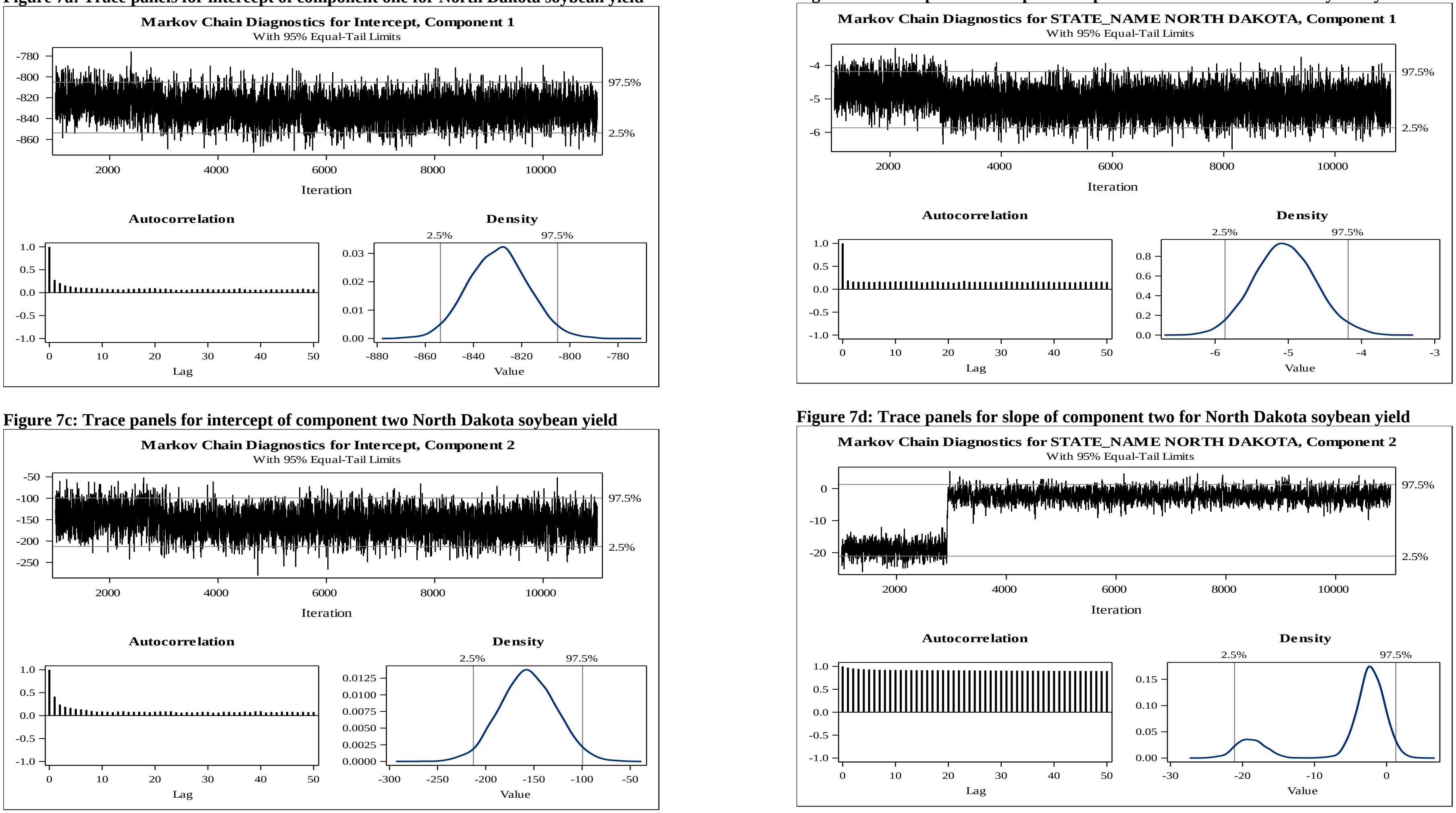


Figure 7a: Trace panels for intercept of component one for North Dakota soybean yield

Figure 7b: Trace panels for slope of component one for North Dakota soybean yield

Figure 7c: Trace panels for intercept of component two North Dakota soybean yield

Figure 7d: Trace panels for slope of component two for North Dakota soybean yield

Table 7: State-wise prediction of all soybean yield models

PREDICTED MODEL	**KANSAS**					**NEBRASKA**					**NORTH DAKOTA**				
	N	Mean	Std Dev	Min	Max	N	Mean	Std Dev	Min	Max	N	Mean	Std Dev	Min	Max
Model 1	912	35.52	7.18	24.08	55.37	1536	36.01	8.10	24.08	55.37	192	35.03	6.47	24.08	45.98
Model 2	912	35.53	8.33	18.46	61.06	1536	36.00	9.22	18.46	61.06	192	35.05	7.82	18.46	52.03
Model 3	912	30.76	9.15	8.09	55.90	1536	41.62	9.42	16.68	68.30	192	29.19	8.45	11.52	48.21
Model 4	912	30.76	9.15	8.09	55.90	1536	41.62	9.42	16.68	68.30	192	29.19	8.45	11.52	48.21
Model 5	912	30.76	8.78	9.70	55.14	1536	41.62	10.07	16.78	73.40	192	29.18	8.17	11.61	47.45
Model 6	912	30.76	8.76	9.70	55.14	1536	41.62	10.07	16.78	73.40	192	29.19	8.17	11.61	47.45
Model 7	912	35.52	7.17	24.07	55.27	1536	36.01	8.09	24.07	55.27	192	35.04	6.48	24.07	46.00
Model 8	912	35.53	7.17	24.52	57.26	1536	35.99	8.13	24.52	57.26	192	35.06	6.56	24.52	48.37
Model 9	912	30.76	8.09	14.14	52.89	1536	41.62	8.40	22.73	65.30	192	29.19	7.29	17.57	45.21
Model 10	912	30.76	8.09	14.14	52.90	1536	41.62	8.40	22.73	65.30	192	29.19	7.29	17.57	45.21
Model 11	**912**	**30.76**	**7.65**	**15.58**	**51.69**	**1536**	**41.62**	**9.08**	**22.87**	**69.46**	**192**	**29.19**	**6.97**	**17.71**	**44.01**
Model 12	912	30.76	7.63	15.58	51.67	1536	41.62	9.08	22.88	69.44	192	29.19	6.97	17.71	43.99
Smoothened Yield	**912**	**30.76**	**6.88**	**14.96**	**55.45**	**1536**	**41.62**	**10.14**	**23.82**	**71.77**	**192**	**29.19**	**6.47**	**16.39**	**42.39**
Actual Yield	**912**	**30.76**	**10.36**	**8.00**	**62.50**	**1536**	**41.62**	**11.35**	**15.40**	**72.20**	**192**	**29.19**	**8.13**	**9.20**	**48.30**

PREDICTED MODEL	OKLAHOMA					SOUTH DAKOTA				
	N	Mean	Std Dev	Min	Max	N	Mean	Std Dev	Min	Max
Model 1	240	35.50	6.88	24.08	55.37	288	35.03	6.47	24.08	45.98
Model 2	240	35.50	8.19	18.46	61.06	288	35.05	7.82	18.46	52.03
Model 3	240	23.40	8.78	4.01	45.57	288	34.15	8.67	13.03	54.71
Model 4	240	23.40	8.78	4.01	45.57	288	34.15	8.67	13.03	54.71
Model 5	240	23.40	8.05	4.10	43.68	288	34.15	8.40	13.12	53.95
Model 6	240	23.40	8.05	4.10	43.68	288	34.15	8.40	13.12	53.95
Model 7	240	35.50	6.88	24.07	55.27	288	35.04	6.48	24.07	46.00
Model 8	240	35.51	6.99	24.52	57.26	288	35.06	6.55	24.52	48.37
Model 9	240	23.40	7.67	10.06	42.57	288	34.15	7.54	19.08	51.71
Model 10	240	23.40	7.67	10.06	42.57	288	34.15	7.54	19.08	51.71
Model 11	**240**	**23.40**	**6.92**	**10.20**	**40.66**	**288**	**34.15**	**7.23**	**19.22**	**50.51**
Model 12	240	23.40	6.92	10.20	40.64	288	34.15	7.23	19.22	50.49
Smoothened Yield	**240**	**23.40**	**4.62**	**17.07**	**37.97**	**288**	**34.15**	**8.22**	**14.74**	**54.79**
Actual Yield	**240**	**23.40**	**7.34**	**5.60**	**44.50**	**288**	**34.15**	**10.13**	**4.30**	**60.40**

4. Conclusions

We extend the literature on crop yield predictions in the presence of random shocks. Previous studies have examined crop yield distributions based on factors that affect yields, e.g. technological change, climate change and weather-related factors and soil and topographical characteristics. However, agricultural yield data is often affected by random and latent characteristics which lead to dissimilarity of data groups. Meanwhile, mixing data from dissimilar distributions often lead to farm policy implementation problems when not properly accounted for (Liu and Ker, 2020).

To overcome the challenges in the accurate prediction of crop yields, there is the need to properly account for the random data generating processes in crop yield data. Traditional HLM models assume a gaussian distribution. This may lead to bias inference if the subgroups within the data structure are not normally distributed. In this light, the present article proposes a novel statistical method to account for randomness in crop yield data. To accomplish the objective of this article, 12 HLM structures were estimated and compared based on the AIC and BIC values using U.S. northern great plains corn and soybean yield data. For corn, it was found that the HLM structure that accounts for randomness effectively is the smoothened data with a level for the Bayes classifier, state, county and year. For soybean, the structure that accurately accounts for randomness is the smoothened data with a layer considering the Bayes cluster assignment, state, crop reporting district, county and year. Overall, the best model for both commodities imply that accounting for randomness with the unsupervised Bayesian cluster assignment improves the prediction accuracy of the yields. Our results indicate that, assigning Bayes class as a level in an HLM improves prediction of the crop yield data in the presence of random and latent characteristics. Finally, the results from this study also indicate that drawing observations from neighboring states can help improve the prediction of states yields due to similarity of characteristics among different states.

We employed a balanced data set to test the proposed model due to the incorporation of a penalized B-spline for temporal smoothing. Researchers who do not wish to use this method for temporal smoothing can choose to use either a balanced or unbalanced data. However, it must be noted from our results that, incorporating the temporal smoothing technique improved the prediction of the model.

5. Future Research

Primarily, farm policy is dependent on the ability to predict relevant policy variables based on a given set of indicators. In the U.S., most of the policies introduced to improve the agricultural sector are embedded in the Farm Bill. Some of these policies include the conservation programs, farm program payments, trade & taxes, and crop insurance.

Despite the policies enacted to help boost agricultural production and producer welfare, the risks and uncertainties involved in agricultural production often lead to loss of income for producers. Hence, to evaluate their impacts, it is necessary to have an appropriate statistical method. The improvements in prediction accuracy from the proposed model implies that policy variables of interest such as conservation program payments, climate change, change in consumer preference for processed agricultural products and other farm bill policies can be evaluated using the hierarchical linear model with a Bayesian layer obtained from a Dirichlet process mixture model classification. Given this finding, the next phase of this research will evaluate the sources of variability of crop yields based on the proposed model. The specific objectives to be pursued as part of this goal will;

- Evaluate the impact of farm and conservation program payments on ND corn and soybean production.
- Examine the impact of farm fertilizer utilization on greenhouse gas emissions and productivity.

References

Addey K.A. (2021). The cost of partners' genetically modified organisms regulatory index on U.S. corn and soybean exports. Food and energy security 10 (1).

Brady M. and Irwin E. (2011). Accounting for spatial effects in economic models of land use: Recent developments and challenges. Environmental and Resource Economics 48:487-509.

Brester G.W., Atwood J., Watt M.J. and Kawalski A. (2019). The influence of genetic modification technologies on U.S. and EU crop yields. Journal of Agric and Resource Economics. 44(1):16-31.

Bush C.A. and MacEachern S.N. (1996). A semiparametric Bayesian model for randomized block designs. Biometrika 83,275 -285.

Duarte G.V., Braga A., Miquelluti D.L. and Ozaki V.A. (2018). Modelling of soybean yield using symmetric, asymmetric and bimodal distributions: Implications for crop insurance. Journal of Applied Statistics, 45(11):1920-1937.

Dunson D.B. (2006). Bayesian dynamic model of latent trait distributions. Biostatistics. 7(4): 551 -568.

Eilers P.H.C. and Marx B.D. (1996). Flexible smoothing with B-splines and penalties. Statistical science 11:89-121.

Finger R. (2013). Investigating the performance of different estimation techniques for crop yield data analysis in crop insurance applications. Agr. Econ. 44:217-230.

Harri A., Coble K.H., Ker A.P. and Goodwin B.J. (2011). Relaxing heteroscedasticity assumptions in area-yield crop insurance rating. Amer. J. of Agr. Econ. 93 (3): 707 -717.

Kleinman K.P. and Ibrahim J.G. (1998). A semiparametric Bayesian approach to the random effects model. Biometrics 54: 921-938.

Liu Y. and Ker A.P. (2020). Rating crop insurance contracts with nonparametric Bayesian Model Averaging. Journal of Agricultural and Resource Economics. 45(2): 244 -264.

Malsiner-Walli G., Fruhwirth-Schnatter S. and Grun B. (2017). Identifying mixture of mixtures using Bayesian Estimation. Journal of Computation and Graphical Statistics 26:2 (285-295).

Ramsey A.F. (2020). Probability Distributions of Crop Yields: A Bayesian Quantile Regression Approach. Amer. J. Agr. Econ. 102(1): 220 -239.

Shaik S. and Addey K.A. (2020a). Corn Production Indicators Report: Appendix III. Agribusiness and Applied Economics Report No. 802 – Appendix III. Doi: http://dx.doi.org/10.22004/ag.econ.308261

__(2020b). Soybean Production Indicators Report: Appendix III. Agribusiness and Applied Economics Report No. 803 – Appendix III. Doi: http://dx.doi.org/10.22004/ag.econ.308299

__(2020c). Corn Production Indicators Report: Appendix II. Agribusiness and Applied Economics Report No. 802 – Appendix II. Doi: http://dx.doi.org/10.22004/ag.econ.308259

__(2020d). Soybean Production Indicators Report: Appendix II. Agribusiness and Applied Economics Report No. 803 – Appendix II. Doi: http://dx.doi.org/10.22004/ag.econ.308298

Shaik S. and Bhattacharjee S. (2016). Hierarchical crop yield linear model. Lett. Spat. Res. Sci. 9, 219-231.

Tolhurst T.N. and Ker A.P. (2015). On Technological Change in crop yields. Amer. J. of Agr. Econ. 97 (1): 137 -158.