Forecasting county-level corn and soybean yield in the Midwest using a multivariate regression model

Cai Chen

Department of Agricultural Economics, Purdue University

AGEC 499: Honors Thesis

Dr. Mindy Mallory

May 5,2023

Abstract

The aim of this paper is to forecast corn and soybean yields across multiple midwestern counties by running various multivariate regression model using environmental factors such as soil texture, pH, organic matter, amount of precipitation, average temperature, and total sunlight hours during the typically growing season for corn/soybean in a county and test the results with 2022 USDA crop production report numbers. Current USDA yield data is limited to a certain counties in the Midwest, and not inclusive of all counties within a state. Using this yield forecasting model, we hope to predict corn/soybean yield for counties not included in the USDA crop production report. Furthermore, a crop yield estimator could be a useful tools for stakeholders who are actively participating in the agricultural commodity markets. Results indicate the best multivariate regressions model for corn yield estimates are models that included a smaller range of dates and models that calculate yield for county using data within the same state that the county is location. The model had more difficulty estimating soybean yields since a country wide regression model made the yield estimates homogenous across counties while yield estimates calculated using only county state data were insufficient to give realistic soybean estimates. Overall, the results of the model that estimated county yield using other counties within that state and limiting yield data from 2010-2021 to 2018-2021 produced on average better yield estimates for corn and soybean, however there was greater variance within the estimates generated due to a smaller dataset. Variance could be decreased in the future by increasing the sample size of the data used to create the model. The yield model which utilized 2018-2021 corn yield data within each state had the highest average R square value of 0.3988 and average adjusted R square value of 0.3274. The yield model which utilized 2018-2021 corn yield data across all states had a mean R square value of 0.19922 and adjusted R square value of

0.18866. From these results we can infer that other variables like fertilizer application and field slope may be necessary to create a better multivariate regression model to calculate corn and soybean yield.

Introduction

The United States is a major producer of corn and soybean. According to the USDA 2022/23 World Markets and Trade report in 2022 the United States accounted for ~30% of global corn production and ~31% of global soybean production producing 13.9 billion bu. of corn and 4.3 billion bu. of soybean. (USDA, 2023) This paper looks to forecast corn and soybean production for counties in Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin using soil and weather data.

The ability to forecast corn and soybean yield is a valuable insight for stakeholders in the agricultural commodity market because corn and soybean basis are heavily influenced by local supply and demand. Any disruptions to U.S. corn and soybean production can cause drastic changes in grain and oilseed future prices, since the U.S. is the worlds largest corn exporter and 2nd largest soybean exporter. The ability to forecast corn and soybean production is important for all stakeholders in the commodity market.

The use of multivariate regression to predict corn and soybean yield using soil and climate data is not a new concept and was researched as far back as 1986 by K.R. Olson and G.W. Olson Use of Multiple Regression Analysis to Estimate Average Corn Yields Using Selected Soils and Climate Data*.(Olson and Olson 1986). Olson research obtained data from research plots across New York State and utilized Climate (Total yearly rainfall), Management (High), Site (Drainage Class), Topography (Slope), Chemical (Sum of bases), Physical

(Porosity) , Minerology (Total silt and very fine sand) , Biology (Organic Carbon), and time variables (One season) to run a multivariate regression model that estimates corn yield. Olson data were obtained from research plots which are highly managed meaning factors such as fertilizer usage, water, and topography could be controlled. Data from our regression model is not controlled for and would be a better representation of real world data. Both regression model seeks to calculate yield across multiple states using similar independent variables like precipitation and soil type. While our regression model can't control or calculate fertilizer usage across different counties, we have included the variable of avg. daily sunlight and avg. daily temperature to our model. Additionally our model allow for the ability of not only estimating corn/soybean yield of plots with available data but also estimating corn/soybean yield for plots that are not currently farmland. This can be advantageous to growers who look to expand farmland acreage and businesses that may want to build grain/oilseed storage/processing plants near high yielding cropland. In recent years, there has been other papers written on estimating corn and soybean yield across counties in the U.S. such as County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model, (Sun, Jie 2019), where they used convolution neural network (CNN) machine learning model to predict soybean yield with great accuracy R^(2) =0.78.

Methodology

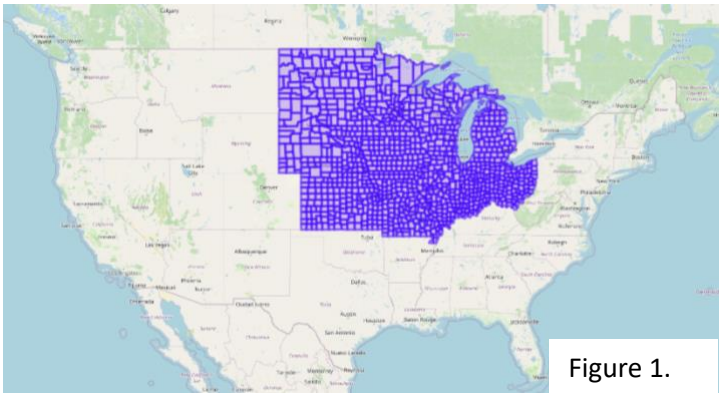


Figure 1.

According to the USDA crop production report, the most productive corn/soybean state are Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. In the 2022 crop year these 12 states accounted for 88.4% of total U.S. corn production and 83% of total U.S. soybean production. Yield data from each county in each state was obtained using the USDA National Agricultural Statistic Service (NASS) Database for all years of available data.

Corn yield in Indiana from 1929-2021 have increased significantly over the past few years with the adoption of double-cross hybrids corn, nitrogen fertilizer, chemical pesticides, agricultural mechanization, an improved soil and crop management practices. (Nielsen R.L 2023) Average corn yield across counties in Indiana was 30.2 bu./ac in 1929, in 2021 average corn yield across counties in Indiana is 192.3 bu./ac. Similarily, soybean yields have also increase drastically from 1937-2021 with average soybean yields in 1937 hovering at 17.5 bu./ac to average soybean yields in 2021 at 58.3 bu./ac.

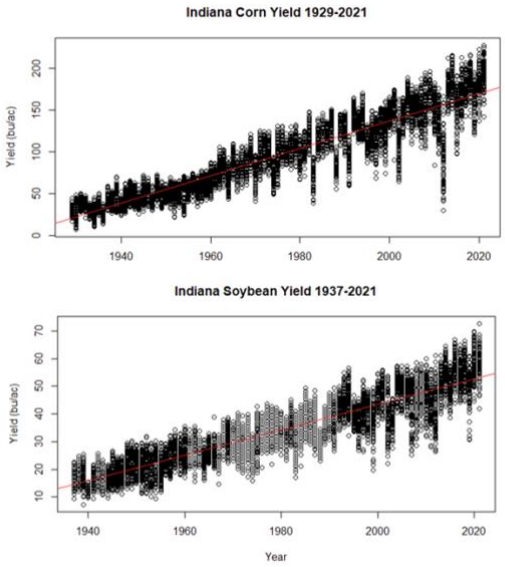


Figure 2.

Gorwing conditions like precipitation, sunlight, and temperature are all important factors that contribute to crop yield.

Corn for Grain Usual Planting and Harvesting Dates – States

State	2009 Harvested acres	Usual planting dates			Usual harvesting dates		
		Begin	Most active	End	Begin	Most active	End
	(1,000 acres)						
Alabama	250	Mar 15	Mar 25 - Apr 25	May 18	Aug 2	Aug 11 - Sep 20	Oct 15
Arizona	20	Mar 10	Apr 1 - May 15	Jun 1	Sep 1	Oct 1 - Nov 1	Dec 1
Arkansas	410	Mar 26	Apr 1 - Apr 26	May 9	Aug 16	Aug 23 - Sep 23	Oct 6
California	160	Mar 15	Apr 1 - Jul 1	Jul 15	Sep 1	Oct 1 - Nov 1	Nov 15
Colorado	990	Apr 19	Apr 28 - May 20	May 29	Sep 28	Oct 8 - Nov 13	Nov 22
Delaware	163	Apr 12	Apr 30 - May 16	May 28	Sep 10	Sep 20 - Oct 15	Nov 5
Florida	37	Mar 1	Mar 15 - Apr 25	May 5	Jul 15	Aug 1 - Sept 10	Oct 1
Georgia	370	Mar 14	Mar 22 - Apr 21	May 4	Aug 6	Aug 16 - Sep 22	Oct 7
Idaho	80	Apr 21	May 5 - May 26	Jun 9	Sep 29	Oct 20 - Nov 10	Nov 24
Illinois	11,800	Apr 14	Apr 21 - May 23	Jun 5	Sep 14	Sep 23 - Nov 5	Nov 20

Figure 3.

Preciptation and sunlight plays a vital role in plant photosythesis, and optimal growing temperatures are essential for daily plant functions. In figure 2, there is a significant drop in corn and soybean yield in 2012 due to a country-wide drought, which prevented corn and soybean plants from reaching their full yield potentials due to lack of water from preciptiation. Using Open-Meteo open-source weather API we were able to extract daily precipitation, sunlight hours, and temperatures for all counties of interest from January 1st of 2010 to March 3rd, 2023. In order to account for variation in growing season of corn/soybean amongst counties due to geography. The USDA Usual Planting and Harvesting Dates for corn/soybean was used to create date ranges for beginning of planting and ending of harvest depending on the state the county was located in. Total precipitation, average temperatures, and average daily sunlight hours for each county and year was then calculated. In figure 3 you can see that different states have different beginning planting and ending harvest dates for Corn.

Different crops have different soil and pH prefences. For example, “The best soybean yields occur on well-drained, but not sandy, soil having a pH of 6.5 or above.” (Cornell CALS). Our model utilizes the XPolaris Package in R to get the pH, organic matter, Clay, Silt, and Sand of every county of interest. XPolaris takes latitude and longitude data and returns pH, organic matter, Clay, Silt, and Sand at 0-5cm, 5-15cm, and 15-30cm. The pH, organic matter, Clay, Silt, and Sand various depth output were averaged to get a single value for each county. Using the Clay, Silt, and Sand data generated from XPolaris, The R soil texture package was then utilized to calculate the soil texture for each county as shown in Figure 4a, where each dot represent a county of within the dataset and the soil texture for that particular county on the soil texture triangle. Across our states of interest there are 1055 counties, however some of these counties are urban areas and others we unable to obtain geospatial data from XPolaris leading to 980 counties represented in model. In figure 4b. we can see the distribution of soil texture across our 980 counties of interest with a majority of them have Silky Loam soil texture. Tables 1-3 shows the summary of our dataset.

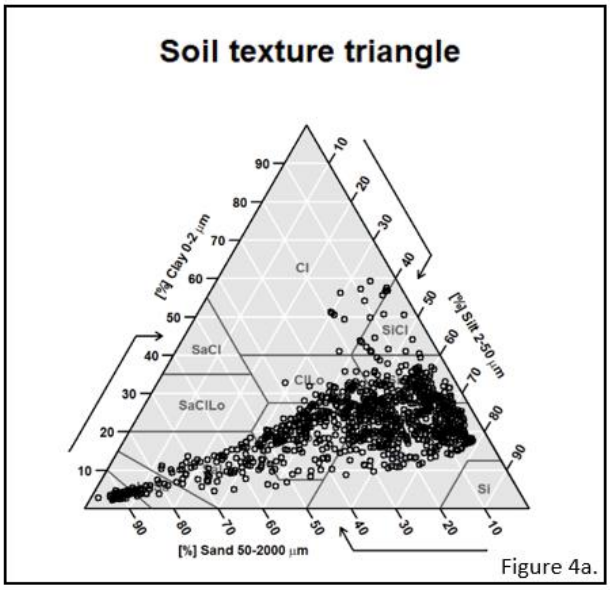


Figure 4a.

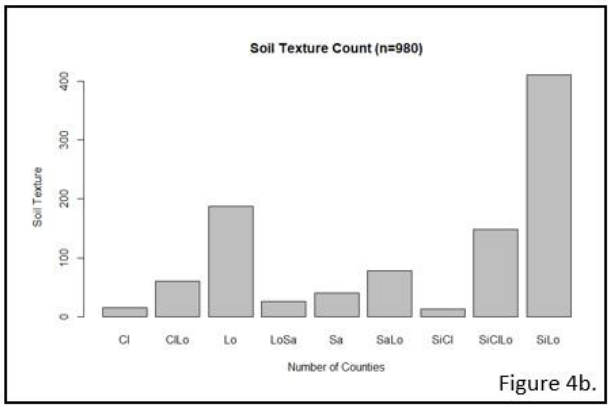


Figure 4b.

Table 1 Descriptive Statistics										
	Statistic	**Yield (bu./acre)**	**total_precip (mm)**	**avg. temp (C*)**	**avg. sunlight (hr)**	**ph**	**om**	**CLAY**	**SILT**	**SAND**
Descriptive statistics soybean (n=9498)	**Min**	6.9	68.3	14.38	13.24	4.073	0.3701	2.229	4.564	2.083
	1st Quartile	41	427.5	18.91	13.56	5.918	1.356	17.744	41.568	9.851
	Median	48.5	521.8	20.17	13.7	6.178	1.914	22.646	53.86	21.071
	Mean	47.3	518.1	20.2	13.73	6.234	2.6479	22.501	51.625	25.874
	3rd Quartile	54.7	605.5	21.54	13.89	6.509	2.9264	27.634	64.345	34
	Max	80.4	973.6	25.92	14.41	8.31	71.9903	59.377	78.475	92.758
Descriptive statistics corn (n=10017)	**Min**	19	73.4	12.53	13.06	4.073	0.3701	2.229	1.637	2.083
	1st Quartile	132.6	392	17.63	13.27	5.915	1.3326	17.394	40.112	10.076
	Median	159.7	485.9	19.24	13.39	6.188	1.8804	22.352	53.218	21.676
	Mean	154.7	484.5	19.38	13.71	6.243	2.7061	22.113	50.427	27.459
	3rd Quartile	181.7	570.4	21.14	13.38	6.565	2.8652	27.488	63.765	35.983
	Max	246.7	947.4	26.47	13.89	8.31	80.3822	59.377	78.475	95.61

In our descriptive statistics table 1, we can see that minimum corn and soybean yield from 2010-2021 are 6.9 bu./acre and 19 bu./acre respectively. Average corn and soybean yield are 154.7 bu./acre and 47.3 bu./acre respectively. Maximum corn and soybean yield are 236.7 bu./acre and 80.4 bu./acre respectively. We can also see that daily sunlight hours and pH is pretty consistent across different counties hovering at 13-14 hours of sunlight and 6-6.5 pH respectively. There is quite some variation in total precipitation and average temperature across different counties.

Results

Estimating yield across state

Our first multivariate regression model seeks to determine whether two counties in two different geographic state would produce the same corn and soybean yield if they were given the same soil texture and weather data. For example, would a county in Minnesota and a county in Indiana that both grew corn on Silky Clay Loam soil with identical pH, organic matter, amount of precipitation, average temperature, and total sunlight hours produce the same yield? To answer this question we used the multivariate regression equation below:

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + B_4X_{4i} + B_5X_{5i}$$

$$i = county$$

$$Y_i = Yield\ (bu./acre)$$

$$B_0 = y - intercept$$

$$X_{1i} = pH\ value\ (0 - 14)\ of\ soil$$

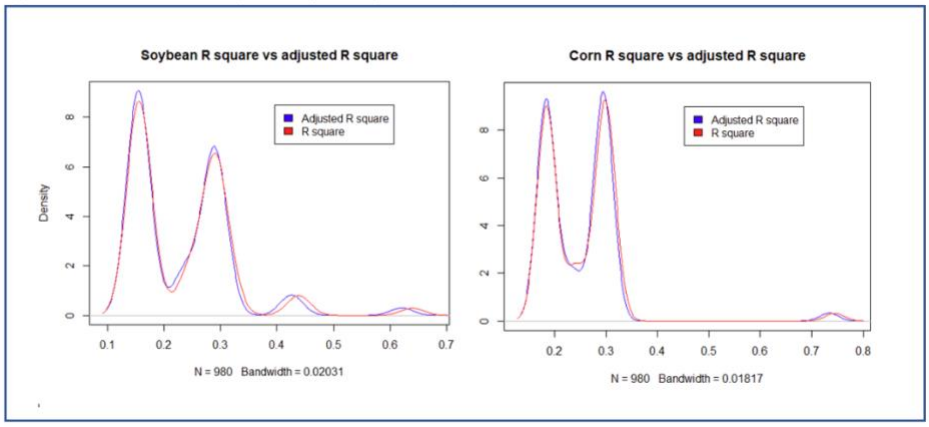
$$X_{2i} = Yearly\ precipitation(mm)\ for\ county$$

$$X_{3i} = Average\ daily\ temperature\ (Celsius)$$

$$X_{4i} = Total\ hours\ of\ sunlight\ in\ growing\ season\ (hours)$$

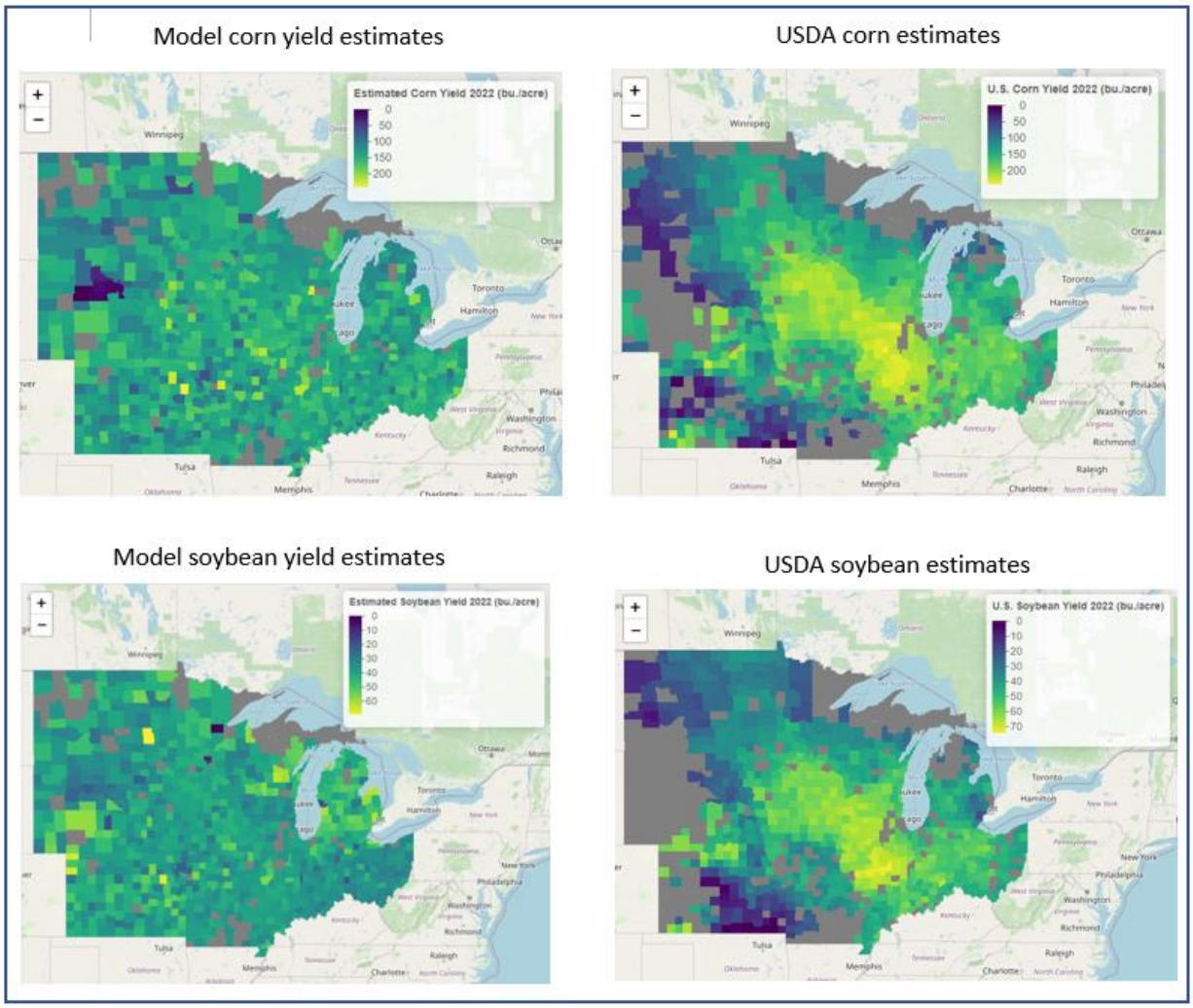
$$X_{5i} = soil\ organic\ matter$$

Figure 5. Across state model R square comparison (2010-2021)



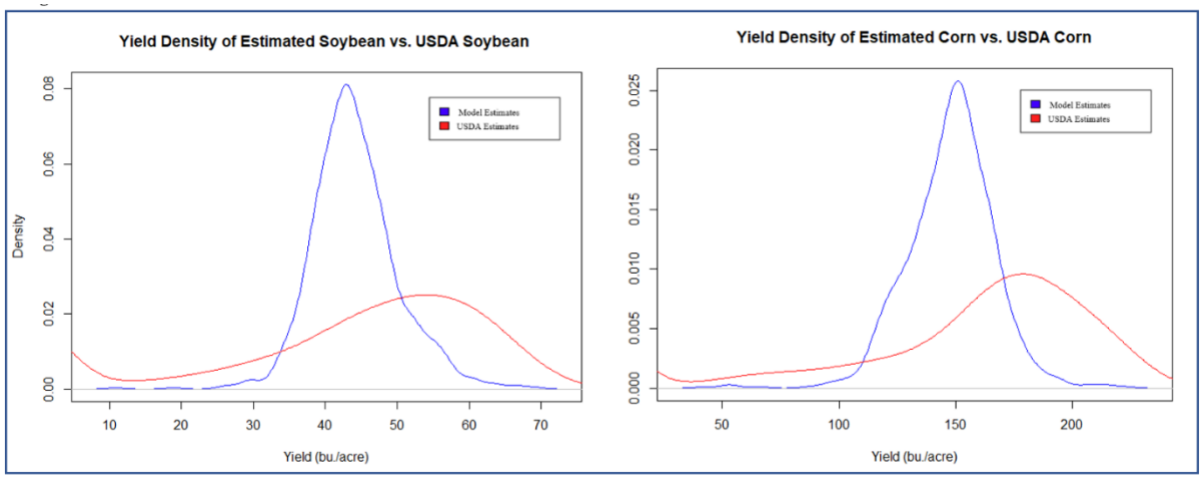
After running a multivariate regression for all counties across all state for corn/soybean. We can see in figure 5 that a high density of the corn and soybean regression model has a R square and adjusted R square value between 0 to 0.40, averaging in the lower 0.20. From the R square and adjusted R square values we can conclude that in this model for most counties 0-40% of the variation in yield can be explained by our soil and weather variables. The density of our R square and adjusted R square are closely aligned in the graph meaning the value gained from the different soil and weather variables are greater than the degrees of freedom loss.

Figure 6. Across U.S. yield model estimates vs. USDA estimates



Using the R we created a choropleth map of yield calculated by our model and USDA yield estimates for visual comparison. In figure 6 we can see that our model corn and soybean estimates have little variation in yield across different states compared to USDA corn and soybean estimates where counties in Illinois have higher corn/soybean yield compared to counties in North Dakota that have lower corn/soybean yield. Figure 7 shows the yield density of the corn/soybean model versus USDA estimates and you can see that the model has very little variance in yield compared to USDA estimates. In figure 7 the model's yield estimates for corn and soybean are not centered at the same point as the USDA estimate, this may infer that our model is biased and with our Beta values not centered at True Beta. Overall the multivariate regression model using data from across the U.S. seems to not be an effective model to measure yield for counties across the U.S. since the yield estimates generated by the model across the U.S. was pretty homogenous throughout the model in figure 6 compared to USDA estimate map.

Figure 7. Yield density of model



Estimating Yield within State

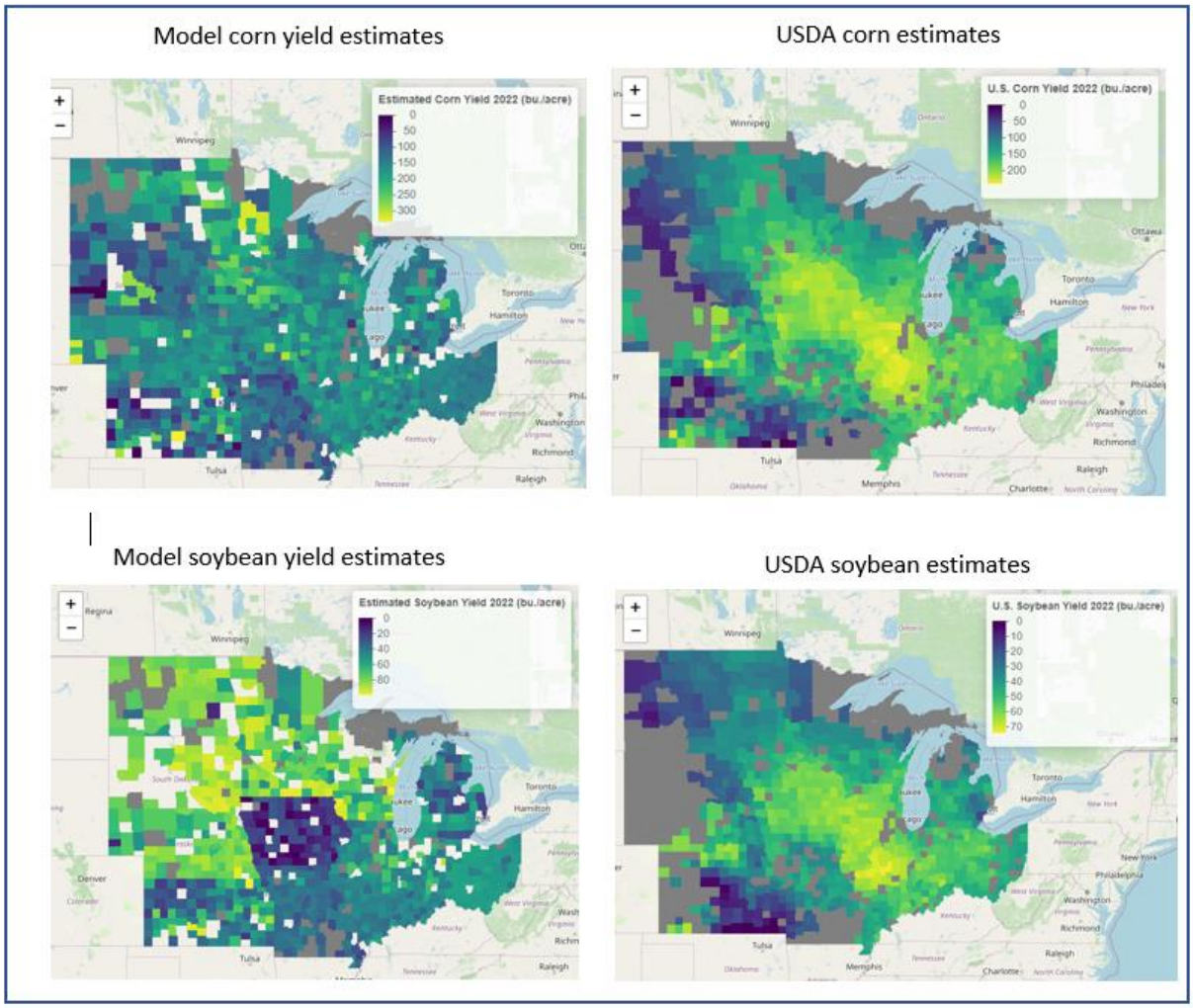
Our previous model tested USDA yield estimates by running a multivariate regression across all counties regardless of the state and produced a homogenous corn and soybean yield estimate across different counties. To create a less homogenous yield estimation model, our 2nd model looks at running multivariate regression models for each counties limiting the regression data to the state the county is located in. This may help create more variation in yield across state and a less homogenous yield estimator. In table 3. We can see that our new model creates yield outliers which heavily skews our corn and soybean yield summary statistics. No corn/soybean hybrid can currently produce a yield of 267,571 bu./acre and 187,637.5 bu./acre under ideal growing conditions. To combat outliers a range for corn and soybean yield were set at 0-350 bu./acre and 0-100 bu./acre respectively.

Table 3. Descriptive Statistics Corn/Bean 2010-2021 Yield

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Model est. corn	-748.27	45.46	61.72	634.06	85.09	187637.5
USDA est. corn	0	28.27	47	39.75	56.73	74.2
Model est. soybean	-159.1	131.6	158.8	3121.5	186.3	267571.1
USDA est. soybean	0	109	166.7	139.7	190.1	240.6

Using the ranges for corn and soybean yield a choropleth map of yield was generated. In figure 8. We can that there is a slight improvement in the model for corn/soybean when a regression was ran on individual counties using only data from the state the county is located in. The model corn estimates seem to be higher in eastern states, then the USDA estimates. For

Figure 8. State by state model estimates vs. USDA estimates



soybean the model seems to underestimate soybean yield in Nebraska and overestimate soybean yield in western states.

Figure 9. County by state model estimates vs. USDA estimates

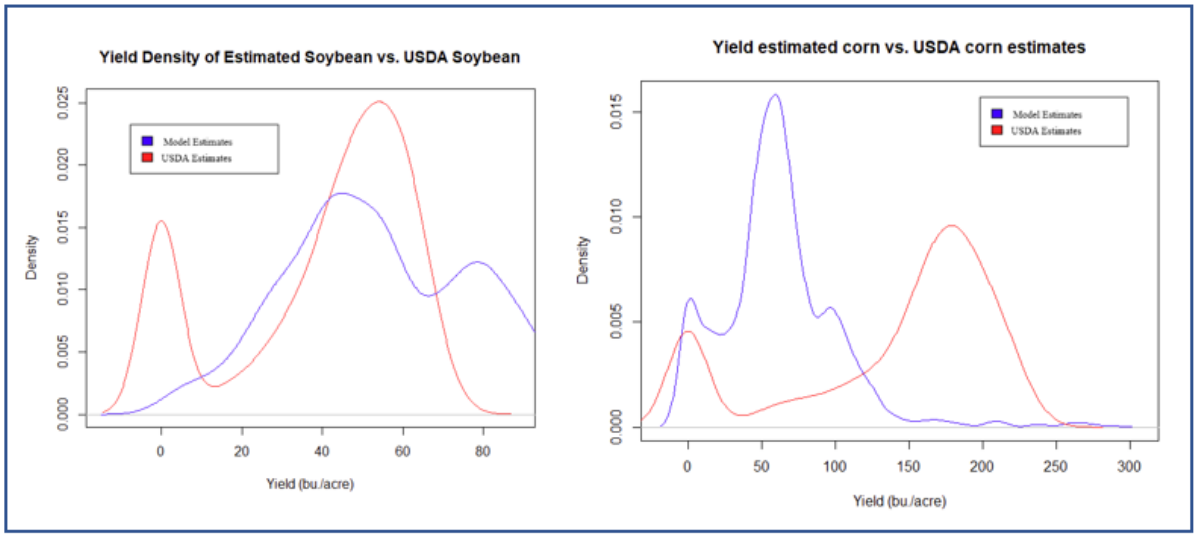
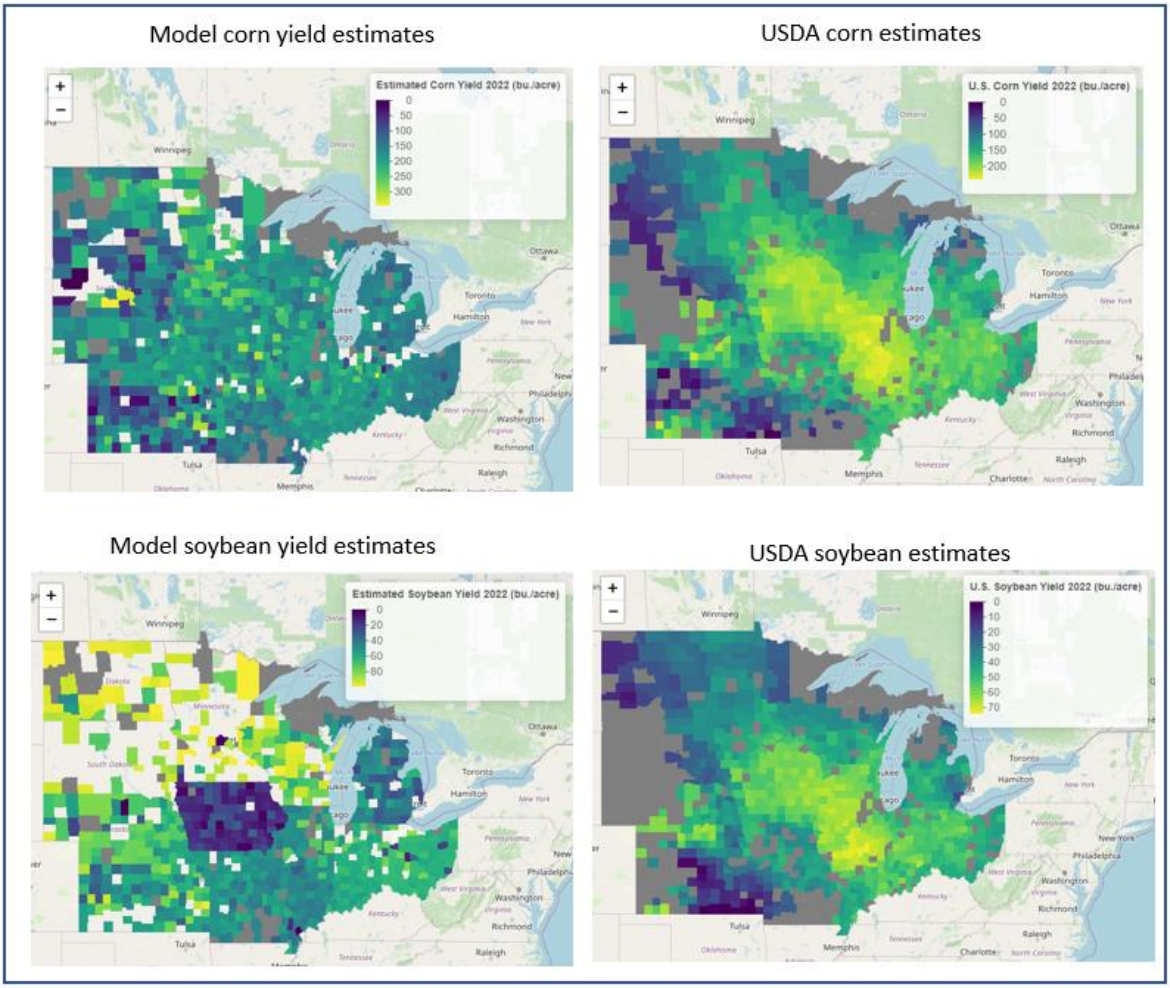


Figure 9. shows the yield density of our county by state model. We can see that while data density is less drastic as our previous model, where our model estimates were heavily biased. The accuracy of our 2nd yield model is not good.

Estimating Yield within State (2018-2021)

Our previous model had bad accuracy in yield. In order to improve the accuracy of the model, we looked at running a multivariate regression model with newer data. This is because corn/soybean yields are constantly improving year after year due to advancements in plant genetics that make plants more tolerant to abiotic stress, more drought resistant, and produce higher yields. By using newer data in our yield estimates our model can account for improve plant genetic seeds. We decided to use the data range of 2018-2021 to run our regression model.

Figure 10. State by State model estimates vs. USDA estimates



By limiting the date ranges for corn and soybean yield and running a regression for each county using only county data from that state figure 10. Was generated. In figure 10. We can see that the accuracy of the model seems to improve slight for corn, but created a lot of outlier soybean yield estimates. Our model soybean yield estimator still over estimates soybean yield in northern regions and underestimates the yield for soybean in Nebraska. Overall, decreasing the date range improved the model yield estimator for corn, but decrease the model yield estimator for soybean.

Conclusion

While using a multivariate regression model to estimate corn and soybean yield based on soil and weather data is feasible, other variables such as fertilizer usage and slope of crop field should be accounted for in the model. Based on our results in table 4. The best regression model were the state specific and latest data range model which had a mean R square value of .3988 and .3933 for corn and soybean respectively. The worst models were the models that ran a regression for county yield across all counties in the U.S. regardless of state and included a large range of date.

Table 4. Descriptive statistics of regression models

	Variables	Corn R^(2)	Corn adj. R^(2)	Soybean R^(2)	Soybean adj. R^(2)
Corn/soybean 2010-2021 state regression model	**Min.**	0.1162	-0.2631	0.0533	-0.3064
	1st Qu.	0.2213	0.2024	0.2337	0.2245
	Median	0.271	0.2578	0.2745	0.2445
	Mean	0.3071	0.2805	0.3093	0.2784
	3rd Qu.	0.3654	0.3349	0.3606	0.3249
	Max.	0.7873	0.7578	0.8911	0.8289
Corn/soybean 2018-2021 state regression model	**Min.**	0.05208	-1.348	0.1034	-2.9509
	1st Qu.	0.18311	0.1609	0.2028	0.1561
	Median	0.38319	0.31	0.3238	0.2882
	Mean	0.3988	0.3274	0.3933	0.2989
	3rd Qu.	0.53443	0.4671	0.5675	0.4427
	Max.	1	0.9691	1	0.954
Corn/soybean 2010-2021 U.S. regression model	**Min.**	0.1845	0.1836	0.1537	0.1527
	1st Qu.	0.1845	0.1836	0.1537	0.1527
	Median	0.2433	0.2383	0.2443	0.2385
	Mean	0.2517	0.2485	0.2345	0.2308
	3rd Qu.	0.2941	0.2909	0.2843	0.2824
	Max.	0.7447	0.7336	0.6396	0.6208
Corn/soybean 2018-2021 U.S. regression model	**Min.**	0.07684	0.07338	0.07902	0.07667
	1st Qu.	0.07684	0.07338	0.07902	0.07667
	Median	0.2178	0.20102	0.21705	0.21169
	Mean	0.19922	0.18866	0.21917	0.21176
	3rd Qu.	0.29035	0.2575	0.36138	0.35777
	Max.	0.77886	0.73937	0.73843	0.71

References

Grain: World markets and trade. USDA Foreign Agricultural Service. (2023, April 12). Retrieved April 29, 2023, from https://www.fas.usda.gov/data/grain-world-markets-and-trade

Oilseeds: World markets and trade. USDA Foreign Agricultural Service. (2023, April 12). Retrieved April 29, 2023, from https://www.fas.usda.gov/data/oilseeds-world-markets-and-trade

Crop Production 2022 Summary. Publication | Crop Production Annual Summary | ID: k3569432s | USDA Economics, Statistics and Market Information System. (n.d.). Retrieved April 29, 2023, from https://usda.library.cornell.edu/concern/publications/k3569432s

Olson, K. R., & Olson, G. W. (1986). Use of multiple regression analysis to estimate average corn yields using selected soils and climatic data. Agricultural Systems, 20(2), 105–120. https://doi.org/10.1016/0308-521X(86)90062-4

Sun, Jie, et al. "County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model." Sensors, vol. 19, no. 20, Oct. 2019, p. 4363. Crossref, https://doi.org/10.3390/s19204363.

Field Crops Usual Planting and Harvesting Dates. Publication | Usual Planting and Harvesting Dates for U.S. Field Crops | ID: vm40xr56k | USDA Economics, Statistics and Market Information System. (n.d.). Retrieved May 3, 2023, from https://usda.library.cornell.edu/concern/publications/vm40xr56k

Nielsen, R. L. (2023, February). *Historical corn grain yields in the U.S.* Historical Corn Grain Yields in the U.S. (Purdue University). Retrieved May 4, 2023, from https://www.agry.purdue.edu/ext/corn/news/timeless/YieldTrends.html

https://cals.cornell.edu/field-crops/soybeans/planting-soybeans#:~:text=The%20best%20soybean%20yields%20occur,August%20will%20have%20disappointing%20yields.

"Moro Rosso, L.H., de Borja Reis, A.F., Correndo, A.A. et al.",

"XPolaris: an R-package to retrieve United States soil data at 30-meter resolution.",

"BMC Res Notes 14, 327 (2021). https://doi.org/10.1186/s13104-021-05729-y"