State Level Modeling

Julie Wisch, Oliver Causey, Song Yuanhong, Charles Yeh, Gavan Tredoux

10/5/2020

Background

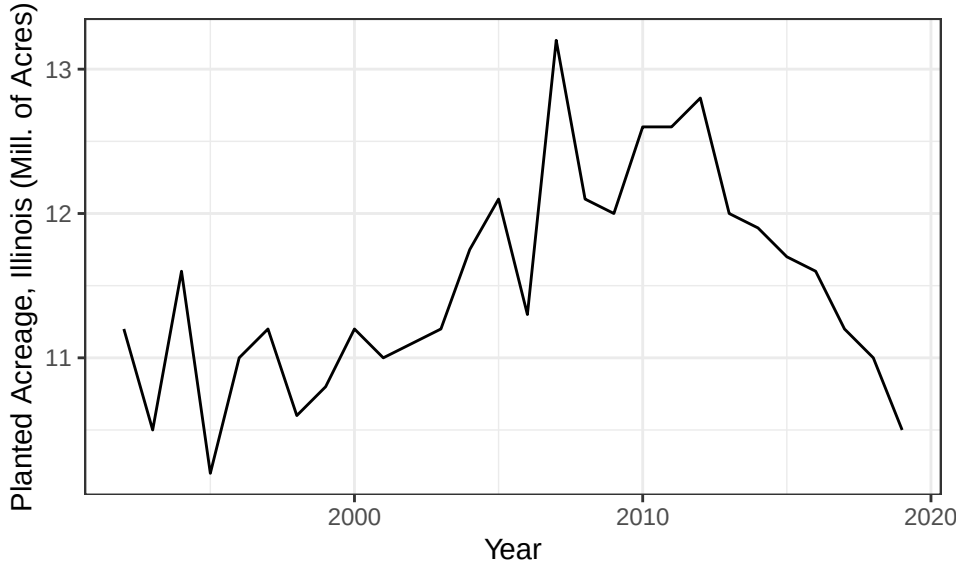
The Bayer Crop Science (BCS) team successfully built a model that could generate national level predictions for acres planted with a 30 - 70% reduction in error as compared to the freely available USDA national level projections. The purpose of this document is to investigate modeling at the state level. We find that we can generate state-level predictions with a 40 - 80% reduction in error as compared to the naive case (assuming what was planted last year will be planted this year). The USDA does not release state-level projections, so we were unable to compare our model performance to a USDA forecast.

USDA Planting Projections and Total Plantings

The USDA publishes many crop planting projections annually, including both corn and soy. These projections are released at a national level only. No state or regional level projections are published.

It is the BCS team's understanding that these models are developed in October, preliminarily released in November, and re-released in final form in February.

The USDA also conducts annual surveys in the first two weeks of March in order to project the total number of acres planted in the US for that spring. The USDA contacts more than 100,000 farmers across the United States. These surveys are re-conducted in June and August of the same year in order to update the planting totals. The USDA uses these probabilistic surveys to calculate the total number of acres planted annually. These totals can be found here: https://quickstats.nass.usda.gov/results/823A5054-B8F9-3341-AC31-8B03C6990BC1 . These surveys are conducted at the FIPS-code level and aggregated to both the state and national level. Below is a sample of the available data, showing the reported planted acreage of corn by year for the state of Illinois.



Modeling Approach

The proposed BCS model is a simple linear regression using a sliding 8 year training window. State Level Planting Acreage is the repsonse variable and State is the only regressor.

A sliding 8 year window means that we train on 8 years of data and test on the 9th. First we train the model on years 2000 - 2007, and then test on year 2008; then, we train the model on years 2001 - 2008 and test on year 2009, etc. We test on years 2008 - 2019, which means that mean average errors are calculated off of 12 datapoints. We would caution the reader against over-interpreting the quality of the model based on this limited number of testing points.

The team considered several other more complex models, including:

- Lasso Regression
- Generalized Additive Models
- XGBoost Modeling, with and without lasso regression for feature regularization

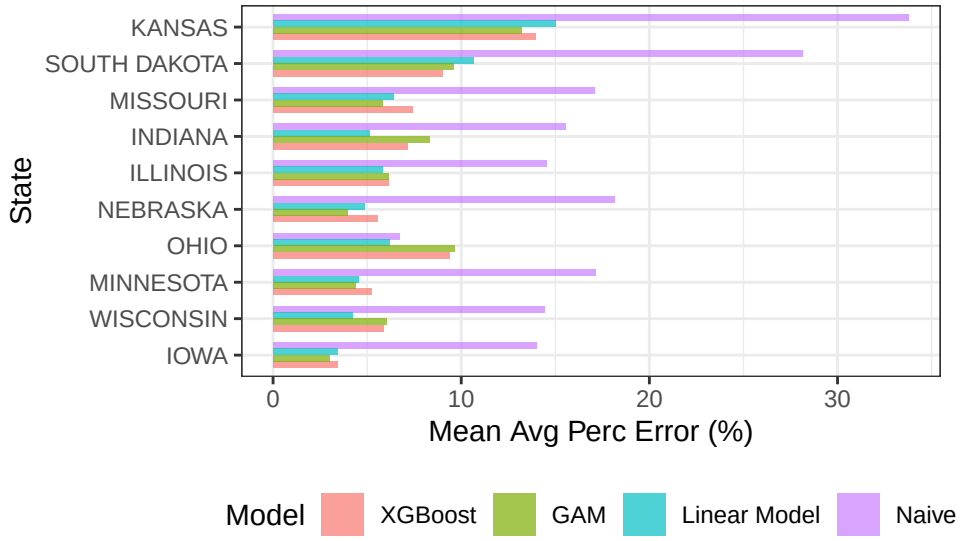
The more complex models considered the following features:

- Historical national planting corn acreage (source: USDA)
- Historical state level planting corn acreage (source: USDA)
- Historical state level planting soy acreage (source: USDA)
- Historical price received per bushel of corn, state level (source: USDA)
- State (this variable was target encoded using last year's state level planting acreage as the response)
- Historical state level sales (source: Internal Data)
- In season state level sales as they become available (source: Internal Data)
- USDA Model Projection for National Planting (source: USDA)
- Estimated USDA State Level Projection for Planting based off of last year's proportion of state level planting and the national planting forecase

Despite including substantially more information, none of the models the team evaluated showed a significant lift beyond the estimate that the very basic linear regression could provide.

Results

The BCS team produced several models that were all better than the naive prediction (that what happened last year will happen this year). Below we show the results of the XGBoost model (performed following regularization), the GAM, and the simple linear model on the 10 largest states for corn planting.

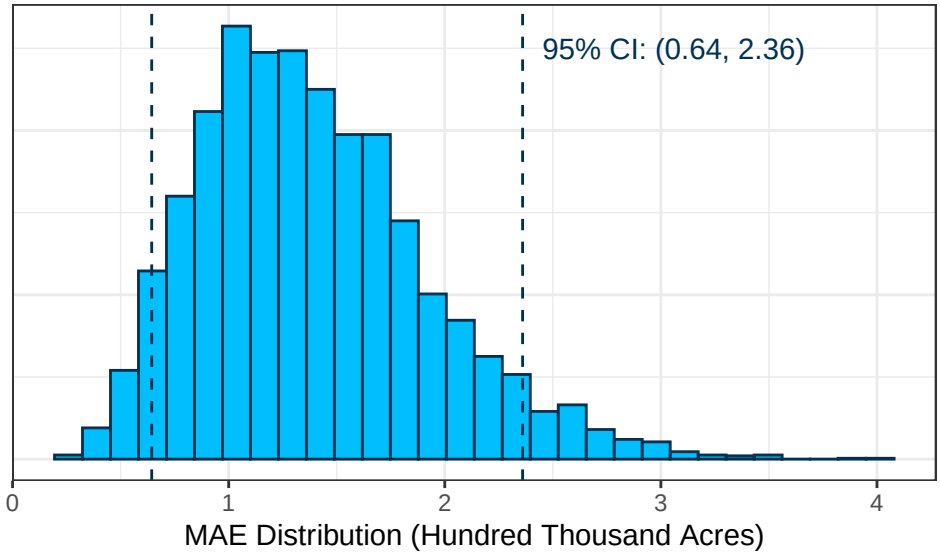


Although the more complex models outperform the linear model for some states, overall there is a negligible difference in performance between the three non-naive models when averaged across all states.

Table 1: Error Metrics Associated with Different Models

Model	MAE (acres/state)
Naive	356000
Linear Model	140000
XGBoost	151000
GAM	144000

In order to confirm our subjective judgment that "there is no difference" between the simple linear model and the two more complex models, we ran a 5000 iteration bootstrap to generate a distribution and estimate a 95% confidence interval. Below we see that although neither the XGBoost nor GAM models fall outside the bounds of the linear models expected error range, the linear model definitively produces an error less than that of the naive prediction.



Recall that for the national case, the team could produce estimates for national corn acreage planting with an MAE of roughly 1.2 million acres. If we aggregate our state level predictions up to the national level, we observe an MAE of 3.8 - 4.6 million acres. Note that the more complex models (XGBoost and GAM) do perform substantially better than the linear model at this level of aggregation. This suggests that the linear model performs better in states that grow less corn, while the XGBoost and GAM perform better in states that grow more corn. While we previously asserted that we recommend pursuing the simple linear model because its simplicity belied the negligible improvements by the GAM and XGBoost model, this particular observation indicates that it may be worth a discussion of the relative importance of estimating planting acres in a state like Illinois as compared to Connecticut, for example.

Table 2: Error Metrics Associated with Nationally Aggregated Models

Model	MAE (acres)
Previously Described National Model	1200000
Linear Model	4620000
GAM	3820000
XGBoost	3810000

It is not surprising that aggregated state level projections are not as good as a single national projection.

As a problem increases in granularity, its difficulty increases. Further, we had relevant data available at the national level (USDA published projections) that does not exist at the state level (except via rough extrapolation). The national model previously developed by the BPI team is still a relevant and important tool, provided the business wants to be able to estimate national planting acreage for corn.

Conclusions

The BCS team can produce a model that is 40 - 80% better than the naive case. On average, the BCS team can predict the state-wide planting acreage to within 140,000 acres (note that the average state evaluated in this document plants 1.8 million acres/year). The state-level model does not substantially improve with the addition of in-season sales data. This means that the BCS team can produce their prediction for the upcoming market year as soon as the acreage planting for the prior season becomes available. Given our teams observations of the USDA's state acreage publication rate, it seems likely that the team could produce a model as early as June 2019 for a 2020 planting season.

Due to the large discrepancy between the national level MAE and the state level MAE, we suggest that if the business is interested in both state and national level projections, two independent models should be implemented.