Predicting Yield in the Corn Belt

Charles L. Hornbaker II
clhornbakerii@gmail.com

J. Benjamin Cook
jbenjamincook@gmail.com

Abstract

We present a spatial model of maize yields in the US Corn Belt from 1970-2012 using a Bayesian prior which induces spatial smoothness among the regression coefficients to mitigate the effects of noisy data across regions and to improve yield forecasting. We outline how this model can be used as one step in an in-season forecasting system, and present one result which indicates why this remains a difficult problem.

1 Background and motivation

The United States is the world's top producer of corn, with over 85 million acres of land reserved for corn production and nearly 14 billion bushels of corn harvested in 2013. Farming is a challenging business. Reduced availability of water and increasing input costs are squeezing farmers' bottom lines during a time of extreme volatility in the commodities markets. New regulations on irrigation, pesticide use, and fertilizer application mean farmers must find a new way to boost yield.

Fortunately, unprecedented amounts of data are available on modern farms, from yield monitors to weather sensors to infrared imaging. Agricultural production, however, lags far behind other industries in using data to make decisions. Soon, using data to optimize decision-making will no longer be a novel luxury – it will be essential for farmers to stay in business.

In this paper, we present a spatial Bayesian model of maize yields over space and time. Because of the abundance of regional data available in the United States, we focus on modeling maize yields at the county level in the Midwest United States with the idea that a similar procedure could be conducted on a more local scale if the right dataset were to become available. We outline how this model can be used as one step in an in-season forecasting system, and present one result which indicates why this remains a difficult problem.

2 Data

Our model focuses on 591 counties in the Corn Belt states of Illinois, Indiana, Iowa, Kansas, Missouri, and Nebraska. Our primary outcome variable is estimated average maize yield (in bushels per acre) for a given county each year from 1970-2012.

This data is derived from sample surveys by the National Agricultural Statistics Service (NASS) of the United States Department of Agriculture (USDA). We aquired this data from the NASS Quick Stats API. [7]

Climate data comes from the United States Historical Climatology Network (USHCN). There are more than 1,200 USHCN weather stations recording daily precipitation (in hundredths of an inch), daily minimum temperature (in degrees fahrenheit) and daily maximum temperature (also in degrees fahrenheit).

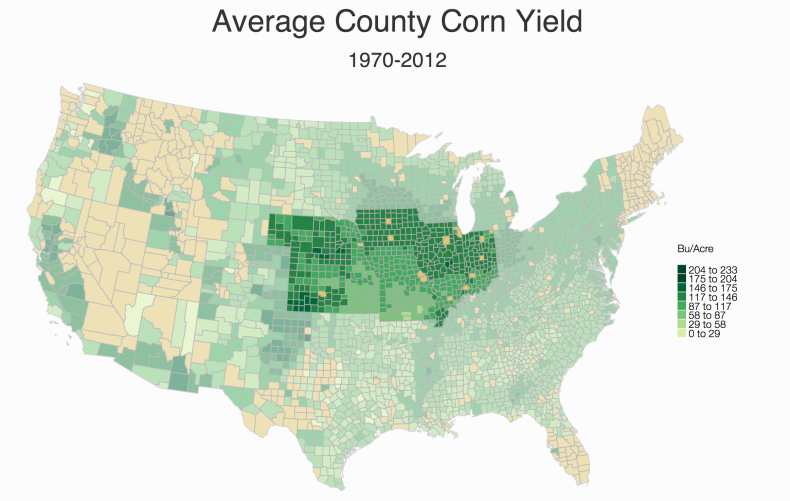


Figure 1: Average yield by county

In our model, temperature is summarized by converting to celsius and computing daily Growing Degree Days (GDD) which is a function of the daily minimum and maximum temperatures,

$$GDD = max\left(0, \frac{Tmax - Tmin}{2} - T_{base}\right)$$

, where $T_{base} = 10$. [2]

Since soil moisture is a major component driving crop yield, we also include available water storage (0 - 100cm) in the soil at the county centroid, which is an approximation of average available water storage throughout the county. This data comes from a website maintained by the California Resource Soil Lab at the University of California, Davis [3].

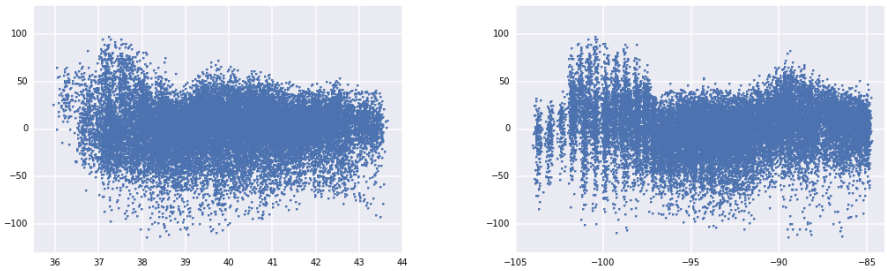
3 Approach

The core of our approach is to model yield at the county level using a spatial Bayesian regression model. That is, we estimate a separate set of regression coefficients for each state and use a Bayesian prior over these coefficients that is spatially smooth. This approach, is straightforward and has been shown to be an effective way to model crop yield [1]. This section outlines modeling decisions, the spatial Bayesian regression model and an in-season forecasting procedure.

3.1 Modeling decisions

In order to estimate GDD and precipitation in the counties of interest, we assign a daily average of the five nearest weather stations to each county centroid from April 1 to October 31, a time range that contains the typical growing season. This procedure allows us to overcome missing weather data in a simple way and is a reasonable approximation of weather in the county.

Because daily climate observations are high-dimensional, it is unwise to include all of the observations directly into the regression model. We reduce the dimensionality of GDD by including average GDD and squared average GDD in our regression models. We choose a simple method of summarizing GDD because it works well in practice and it accords with intuition. More GDD's are helpful up to a certain point and then begin to damage crops. To reduce the dimensionality of precipitation, we perform a non-negative matrix factorization [6], which disallows transformed values to be negative. We include the first two bases in the regression models.



Residuals vs Latitude

Residuals vs Longitude

Figure 2: Simple Linear Model Residuals

3.2 Least squares

We begin with a simple linear model, using features from climate, soil, and historical yield:

$$\begin{aligned} y_{j,t} =& \beta_0 + \beta_1 t + \beta_2 \overline{GDD}_{j,t} + \beta_3 \overline{GDD}^2_{j,t} + \\ & \beta_4 Prcp1_{j,t} + \beta_5 Prcp2_{j,t} + \beta_6 SW_j + \epsilon_{j,t} \end{aligned}$$

The time variable (in years) allows us to represents the technology trend directly in the model. The climate variables GDD, GDD^2, Prcp1 and Prcp2 are summaries of daily climate as described above. Finally, we include an estimated soil water storage level (SW) measured in cm for each county.

The least squares model is too simple to capture the spatial variability that exists in the data. Figure 2 shows residuals by longitude (left panel) and by latitude (right panel). There are a number of ways to account for spatial dependence between counties. One simple approach is to fit a separate linear regression model to each state. Unfortunately, this approach results in overfitting, which hurts the predictive performance of the model. Figure 5 shows the root mean square error (RMSE), a measure of predictive performance for the multiple least squares models compared to the spatial Bayesian regression model we describe in the next section.

3.3 Spatial Bayesian regression

A more robust way to account for spatial variation is to put a Bayesian prior on the regression coefficients. The multivariate conditional autoregressive (MCAR) prior takes the following form:

$$\begin{aligned} p(\beta) &= \mathcal{N}(0, \Lambda^{-1}) \\ \Lambda &= (D - \alpha W) \otimes \mathbb{I} \end{aligned}$$

where W is a (6×6) adjacency matrix that describes the neighborhood structure of states. We assign an edge to states that share a border. $|\alpha| < 1$ is a spatial smoothing parameter, and $D = diag(m_i)$, where m_i is the sum of neighbors for state i. The right hand side of the kronecker product is required to be a positive definite matrix and we use the identity matrix for simplicity [4].

The design matrix is constructed in such a way that each state gets a separate column for each variable. So, for example, the first few rows and columns look like the following:

$$X = \begin{bmatrix} 1 & 0 & \cdots & \text{Year}_1 & 0 & \cdots \\ 1 & 0 & \cdots & \text{Year}_2 & 0 & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots \\ 1 & 0 & 0 & \text{Year}_n & 0 & 0 \\ 0 & 1 & 0 & 0 & \text{Year}_1 & 0 \\ 0 & 1 & 0 & 0 & \text{Year}_2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Because this is a Markov Random field for the regression coefficients, zeros in the precision matrix indicate conditional independence between two coefficients [5]. This means the technology trend coefficient for Nebraska is conditionally independent of the technology trend coefficient for Illinois given the coefficient for Iowa. This allows states to borrow statistical power from one another and mitigates the problem of overfitting. The next section presents how this model can be used for in-season forecasting.

Figure 3 displays the regression coefficients for each of the six states for all variables in our model. In general, states that share a border have more similar coefficient values, but the coefficients are not completely spatially smooth.

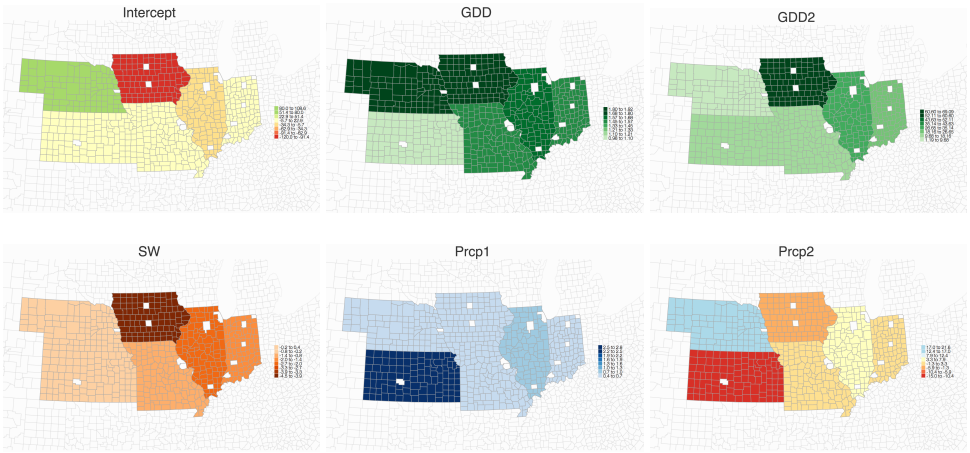


Figure 3: Model Coefficients

3.4 In-season forecasting

The ultimate goal in yield modeling is in-season forecasting, therefore, it is important to consider how this spatial Bayesian regression model helps us move in that direction. According to our model, the only information we are missing before harvest is the daily temperature and precipitation data for the remainder of the season. Any method of forecasting before October 31 requires us to first predict GDD and precipitation for the rest of the days until the end of October.

Ideally, any forecasting method should allow us to propogate uncertainty in climate forecasts through to predictions about yield. One way to accomplish this is to bootstrap weather samples from the past in order to get an idea about typical weather patterns for the remainder of the year. Then we can include those bootstrapped estimates in our predictive model to get in-season forecasts and an estimate of the uncertainty about those forecasts. We perform a basic proof-of-concept for this approach and present the result in the next section. Unfortunately, this procedure exposes a flaw in our model.

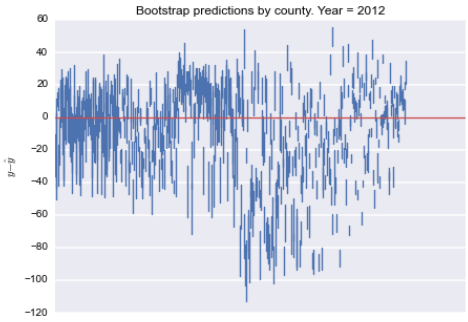


Figure 4: County predictions with bootstrapped weather data

4 Results

To assess performance of the models, we compute R^2 which indicates how much variance in yield is explained by a given model. Since our goal is prediction, a better measure of performance is RMSE. To get RMSE, we fit the model to all but one of the years and then predict yield for the year that we left out. This procedure assumes that we have actually observed the climate time series for the year we left out.

After performing cross-validation we find that the spatial parameter that minimizes the RMSE for out of sample predictions is $\alpha = -0.25$. In terms of model performance, the simple least squares model does not do a good job of capturing the complexity of crop yield. Fitting a separate least squares model for each state improves the fit of the model, but makes generalization to new years problematic because of overfitting. The spatial model with the MCAR prior does almost as well as the multiple least squares approach in terms of variance explained and does drastically better in terms of predictive performance. Figure 5 shows that the MCAR regression model does not completely account for the spatial dependence, but it does dramatically better than fitting multiple least squares models.

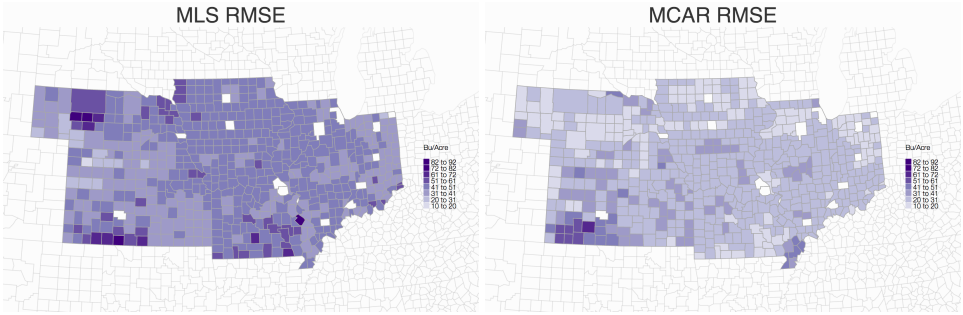


Figure 5: Model RMSE Comparison

Table 1: Model Results

Model	R^2	RMSE
Simple LS	0.44	42.96
LS by State	0.50	65.36
MCAR	0.50	26.11

To assess the in-season forecasting procedure, we leave the most recent year out of the dataset and fit the MCAR model to it. Then, we assign the climate time series from previous years to each of the counties in our test set, reduce the dimensionality as before and predict yield with those variables. This gives us a range of predictions for each county in 2012. Figure 4 shows the range of residuals $(y - \hat{y})$ for each county in the dataset. Zero should fall within this range for a good prediction. Only 29% of the counties contain zero in their prediction ranges. This is not terribly surprising since, in this case we are assuming that we have not seen any of the year's weather data. Unfortunately, after seeing more and more days of the actual weather time series, the predictions improve only marginally. For 2012, we can get as many as 30% of the county residual ranges to contain zero.

5 Conclusions

In order to effectively model crop yield over a large area, we need to model the spatial dependence between locations. The MCAR model reduces the effects of overfitting apparent in the multiple least squares model grouped by states and exhibits significantly lower RMSE than either of the least squares models.

It is possible that our inability to effectively predict yield in-season is due to a weakness in the bootstrap method for estimating weather. But we suspect that it is actually uncovering a weakness in our regression model. Since we cannot move our predictions much with very different climate time series, it is likely the case that we are not properly modeling the relationship between yield and climate.

References

[1] BORNN, L., AND ZIDEK, J. Efficient stabilization of crop yield prediction in the canadian prairies. *Agricultural and Forest Meteorology 152* (2011).

[2] BUTLER, E. E., AND HUYBERS, P. Adaptation of US maize to temperature variations. *Nature Climate Change 3* (2013), 68–72.

[3] CALIFORNIA SOIL RESOURCE LABORATORY. `http://casoilresource.lawr.ucdavis.edu/drupal/`.

[4] JIN, X., CARLIN, B. P., AND BANERJEE, S. Generalized hierarchical multivariate car models for areal data. *Biometrics 61*, 4 (2005), 950–961.

[5] KINDERMANN, R., SNELL, J. L., ET AL. *Markov random fields and their applications*, vol. 1. American Mathematical Society Providence, RI, 1980.

[6] LEE, D. D., AND SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature 401*, 6755 (1999), 788–791.

[7] NATIONAL AGRICULTURAL STATISTICS SERVICE. `http://www.nass.usda.gov/Quick_Stats/index.php`.