Using machine learning to uncover the climatic drivers of US corn yields

David Lafferty, Scott James, Alfonso Ladino

Introduction

The agricultural sector is one of the most vulnerable to changing climate and weather risks. In order to understand how climate change might impact agriculture, we first require models to understand meteorological variables affect crop yields. Machine Learning (ML) provides a powerful set of tools to approach this problem. In this work, we develop a simple ordinary least squares model for end-of-season corn yields across the contiguous United States (CONUS) using various meteorological variables as predictors. We first narrow the set of predictor variables from a wide initial sample, and then perform a cross-validation exercise to show that our selected model performs well out-of-sample. Finally, we analyze whether the influence of extreme heat at different stages of the growing season varies geographically.

Data & Methods

County-level corn yields from 1950-2013 are publically available from the US Department of Agriculture National Agricultural Statistics Service. Climate/weather variables are derived from Livneh et al. (2015) and aggregated to the county level using area-weighted averages. Initially, we choose a wide set of predictor variables by relying on the existing literature: growing and extreme degree days and a quadratic in season-total precipitation from Schlenker & Roberts (2009), and the lowest and highest weekly soil moisture observations as well as a quadratic in season-average soil moisture from Haqiqi et al. (2021). Additionally, Ortiz-Bobea et al. (2019) highlight the importance of growing season timing on predicting yields (for example, that extreme heat closer to harvesting is more damaging than in the early-season), so we calculate each of the above variables at a monthly timescale. The growing season is assumed to be April to September, following previous works.

The first step is to determine a subset of the best predictors. To do this, we utilize SciKit-Learns's KBest function which selects the k-best features by calculating the mutual information between the predictors and target variable. We apply this algorithm to every county and then aggregate over all counties and find that the year, April and September growing degree days (GDDs), August and September extreme degree days (EDDs), season-total precipitation, and irrigation fraction are the best predictor variables. The total number of predictors in the model was determined by minimizing the aggregated out-of-sample error as measured by a 5-fold cross validation: the resulting R^2 for the best model is shown in Figure 1.

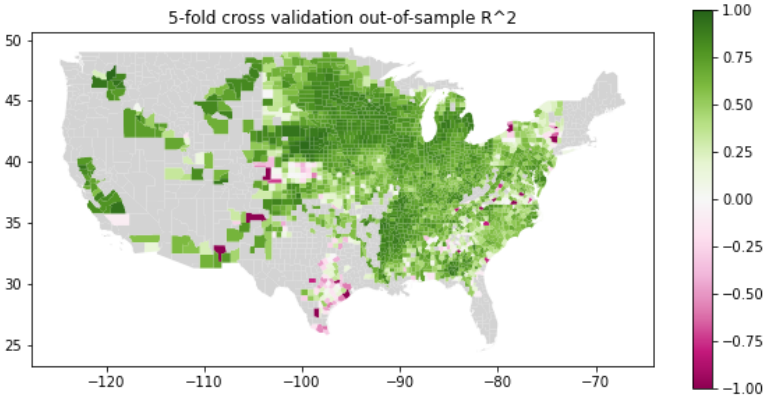

Figure 1: Out-of-sample R^2 as measured by a 5-fold cross-validation

Results & Conclusions

The best predictor variables align with previous works and intuition regarding crop growth. The time trend accounts for historical technological improvements, while the GDD coefficients are generally positive (indicating benefits to yield) and the EDD coefficients are generally negative (indicating harm to yields). Additionally, the feature selection reveals that beneficial temperatures in the form of GDDs are most important at the beginning (April) and end (September) of the season, while extreme temperatures in the form of EDDs are most harmful towards the end of the season.

In Figure 2, we analyze how the coefficient on each of the EDD terms varies across each county. We find that August EDDs are harmful for yields across almost all corn-growing counties in the US and particularly in the Midwest. In contrast, September EDDs show a much more mixed pattern: they are typically harmful for yields in Midwestern counties (particularly in the upper Midwest) but seem less harmful across southern counties. This could be related to different planting and harvesting dates across different latitudes – if southern counties are able to plant and thus harvest earlier than Midwestern counties, they are able to avoid damaging late-season temperatures. Indeed, there seems also to be a similar pattern in August EDDs where southern counties show coefficients of smaller magnitudes and even positive in some cases. An alternative hypothesis is that southern counties are typically warmer across the season and thus farmers have bred their crops to become more heat-tolerant over time.

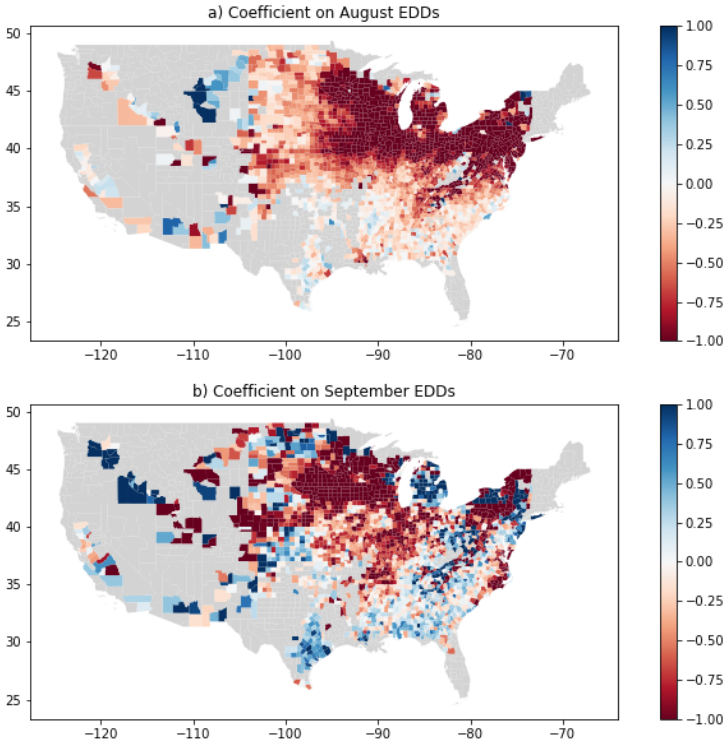

Figure 2: Geographic distribution of the influence of monthly extreme heat on yields

Future work could rely on more sophisticated ML methods to uncover nonlinear relationships among the predictor variables and their affects on yields. Possible candidate approaches include random forest and neural network regressions.

References

Livneh, B., Bohn, T., Pierce, D. et al. A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950–2013. Sci Data **2,** 150042 (2015).

Schlenker, W. & Roberts, M. J. Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proc National Acad Sci* **106**, 15594–15598 (2009).

Haqiqi, I., Grogan, D. S., Hertel, T. W. & Schlenker, W. Quantifying the impacts of compound extremes on agriculture. *Hydrol Earth Syst Sc* **25**, 551--564 (2021).

Ortiz-Bobea, A., Wang, H., Carrillo, C. M. & Ault, T. R. Unpacking the climatic drivers of US agricultural yields. *Environ Res Lett* **14**, 064003 (2019).