Predicting Soybean Yield with NDVI using a Flexible Fourier Transform Model

Chang Xu and Ani Katchova
Department of Agricultural, Environmental, and Development Economics
The Ohio State University

January 17, 2018

Selected Paper prepared for presentation at the Southern Agricultural and Applied Economics Association's 2018 Annual Meeting, Jacksonville, FL, February 3-6, 2018.

Predicting Soybean Yield with NDVI using a Flexible Fourier Transform Model

Abstract

We study how to incorporate the Normalized Difference Vegetation Index (NDVI) derived from remote sensing satellites to improve soybean yield predictions in ten major producing states in the United States. Unlike traditional methods which assume that a global OLS model applies to all observations, we account for geographical heterogeneity by using the Flexible Fourier Transform (FFT) model. Results show that there is considerable heterogeneity in how responsive soybean yield is to NDVI over the growing season. Out-of-sample cross-validation indicates that accounting for geographical heterogeneity improves the forecasts in terms of smaller prediction error compared to models assuming away geographical heterogeneity.

JEL Codes: C14, C53, Q16

Keywords: Crop Yield, Flexible Fourier Transform Model, Forecasting, NDVI, Remote Sensing.

1. Introduction

Many agencies, both public and private, exert significant effort to make crop yield forecasts (Irwin et al., 2014). Accurate and timely crop yield forecasts are valuable in many ways for a number of market participants. At the aggregate level, crop yield forecasts help the price discovery process and improve market efficiency and they aid decision-makers in making rapid decisions to accommodate humanitarian actions and combat food insecurity. At the individual level, crop yield forecasts are used to set insurance premiums for insurance companies and they provide critical information for market participants, especially on the production side, to make adjustments to improve profitability.

In recent years, there has been an increasing interest in using remote sensing data to help improve crop yield forecasting. Remote sensing collects, archives, processes and distributes satellite-derived data (Senay, 2016). For example, the Normalized Difference Vegetation Index (NDVI) contains helpful information generated by remote sensing procedures that can be used to predict crop yield. NDVI is a measure of biomass density on the surface of the earth, usually produced by a space platform. NDVI is defined as:

$$NDVI = (NIR-RED)/ (NIR+RED)$$

where NIR stands for the reflectance of the near-infrared bands of the electromagnetic spectrum and RED stands for reflectance of the visible bands of the electromagnetic spectrum. According to electromagnetic theory, live vegetation absorbs the blue and red bands of sunlight and reflects most of the green band of sunlight. Dying vegetation, to the contrary, absorbs mostly the green band of sunlight and reflects mostly the blue and red bands of sunlight. Barren soil reflects moderately both the visible and near-infrared bands of the electromagnetic spectrum. Generally,

the higher the NDVI, the more NIR light is reflected and the less RED/visible light is reflected, and therefore the target area includes more vegetation.

Because remote sensing provides information at a homogeneous level of accuracy and accessibility, regardless of the location and economic development of the country, using remote sensing data to predict crop yield has the potential to be applied in less-developed countries in a cost-effective manner. In comparison, traditional, survey-based forecasts are relatively expensive and labor-intensive.

Previous NDVI-based forecasting studies (Lv, 2014) utilized Ordinary Least Square (OLS) regression, which assumes that a global coefficient applies to each location invariantly. However, a global coefficient may hide location variation. Due to the complexity of local climate, soil conditions, and farm practices, the effect of NDVI on crop yield may be highly localized. Using a global coefficient to forecast site-specific crop yield may be biased and thus may cause less informed decisions by market participants.

We use Flexible Fourier transform (FFT) model to account for spatial heterogeneity in crop yield forecasts. This is the first study to our knowledge to examine how the correlation between NDVI and soybean yield varies by location and to use this spatial heterogeneity to improve forecast performance for soybean yields. We then compare FFT with OLS in terms of out-of-sample forecast performance. Two hypotheses are tested: 1) the coefficients of NDVI on crop yield are heterogeneous across sites; and 2) models that account for spatial heterogeneity outperform OLS in terms of ex-ante forecasting accuracy.

This paper is organized as follows: Section 2 presents some background information about the current practices on crop yield forecasting and remote sensing for crop yield forecasting; Section 3 introduces the data sources and the FFT method we use; Section 4 presents

descriptive analysis, regression results and forecasting results, making a comparison between the FFT method used in this paper and traditional OLS method; and Section 5 concludes the paper.

2. Background and Related Literature

2.1 Overview of Current Crop Yield Forecast Methods

There are two types of crop forecasts: survey-based forecasts and regression-based forecasts. Survey-based forecasts tend to be more accurate, especially when the harvest date is approaching, usually available shortly before or around harvest time, but they are also more expensive and labor-intensive; regression based forecasts are more cost-effective, and can be available largely ahead of harvest, while their accuracy may be compromised.

Survey-based forecasts that are used by USDA-NASS are made by conducting annually an Agricultural Yield Survey (AYS) and an Objective Yield Survey (OYS) to collect information on crop growing conditions. In AYS, farmers are asked to self-report the growing condition of their crop. In OYS, NASS sends technical personnel to the field to take objective measurements and counts of the plants. Both AYS and OYS are conducted at the beginning of each month from June to September, and forecasts are generated and updated in September, October, and November. The final forecast is released in January of each year. The typical cycle of soybean production in the major producing states in the U.S. is as follows: planting is in May and June, flowering is in July (this is its moisture/temperature sensitive stage), filling is in August, maturation is in September, and harvesting is between October and November.

The second type of crop forecasts is regression-based forecasts. This type of forecasts is used mostly by private agencies, and occasionally as supplementary forecasts by public agencies. For example, the World Agricultural Outlook Board (WAOB) releases World Agricultural Supply and Demand Estimation (WASDE) regression-based forecasts which use trend analysis

and crop weather regression models. Unlike the forecasts released by NASS, the WAOB releases forecasts throughout the growing season, from May to August (Irwin et al., 2014). Unfortunately, the detailed process and exact information used by WAOB to determine the crop yield is not available to the public. Irwin et al. (2014) recommends that "this document be available on the WAOB website." The crop weather model (a.k.a. the modified model) utilizes the year trend variable, monthly weather variables, and an indicator if the crop is planted late. The crop condition model utilizes year trend variable, portion of crop planted after a certain date, e.g. May 30th for soybeans (Irwin, Good, and Tannura, 2009) and the portion of crop rated good or excellent by the USDA (Crop Progress Report). The model we propose in this paper adds NDVI variables to the crop weather model. According to the literature, the modified Thompson model produces a good fit but performs poorly when events that cannot be captured by the weather variable such as insects and diseases negatively impact crop yield. We hypothesize that using NDVI can also monitor for insects and diseases because NDVI is a direct indicator of the greenness/health of the vegetation, with the additional benefit that NDVI data are immediately available at a low cost compared to the methods that rate crop conditions.

2.2 Crop yield forecasting by Remote Sensing

There have been numerous studies documenting the correlation between NDVI and crop yield forecasts, at the national (Maselli & Rembold, 2002), regional, county level (Bolton & Friedl, 2013) and field level (Ferencz et al., 2004). Tucker (1979) determined that the time-integrated NDVI is largely correlated with crop yields when the vegetation is at the maximum level of greenness. Some studies focus on intra-annual variability, that is, how the correlation between the vegetation index and crop yields varies by the crop cycle/planting date (Basnyat et al., 2004). They suggest choosing NDVI data over a specific period for each type of crop to

produce better forecasts. The weekly availability of NDVI data makes this crop-specific specification achievable. Lv (2013) suggests using earlier May NDVI and the change of NDVI over the crop's planting and harvesting for the most accurate yield forecasting. D. M. Johnson (2014) finds that crop yield is highly associated with NDVI and daytime Land Surface Temperature. The author conducts a regression of crop yield on NDVI for every week of the growing season, and finds that the week where the association is at its peak is at the beginning of August.

In addition to NDVI derived from Earth Observing System-Moderate Resolution Imaging Spectoradiometer, called eMODIS, other indexes and images have been used. For example, Doraiswamy and Cook (1995) is one of the earliest studies that used Advanced Very High Resolution Radiometer (AVHRR) imagery. AVHRR data are coarser, eMODIS data are finer; AVHRR data are available for an extended period while eMODIS data are only available after 2000. Following works using AVHRR include Ferencz et al. (2004), in which they used a vegetation index called General Yield Reference Index. Bolton and Friedl (2013) suggest incorporating crop phenology and using a combination of EVI2 (Two-end Enhanced Vegetation Index), NDVI, and Normalized Differenced Water Index (NDWI) for crop yield forecasting. They distinguish between semi-arid and non-semi-arid areas. They find that vegetation indexes are the best type of indexes for predicting in non-semi-arid areas, whereas the NDWI is the best index for prediction in semi-arid areas, because the water index is sensitive to irrigation in these semi-arid areas.

Instead of using traditional statistical models, Bose et al. (2016) utilize spiking neural networks (SNNs) from machine learning to analyze a remote sensing spatiotemporal

relationship. Their work focuses on finding the optimum number of variables (or "features" in machine learning) to be included in regression analysis using machine learning techniques. They find that this type of prediction can be made six weeks before harvest with an average accuracy of 95.64%. They find year 2002 had the largest forecast error due to the 2002 drought. Adrian (2012) applies the Bayesian hierarchical model. This model is suitable for modeling data with clusters. It produces unique estimates for each state while requiring the estimates from each state to also follow a prior distribution. M. D. Johnson et al. (2016) focus on comparing forecast performance using linear versus non-linear machine learning techniques and find that non-linear models are not necessarily advantageous compared to linear models. (Li et al., 2007) find that Neural Network techniques improve corn predictions compared to multivariate analysis. Kaul et al. (2005) find that a non-linear model only outperforms the linear model for barley. Mkhabela et al. (2011) categorize the Census Agricultural Regions (CARs) into three distinct agro-climatic zones, however, even within CARs, there might be multiple soil types. Bolton and Friedl (2013) emphasize the importance to delineate the boundary between farmland and non-farmland such as grassland and forests, as non-farmland may contaminate the NDVI-crop yield relationship. Delineation can be done by using a land cover map such as the Landsat Thematic Mapper data (Bolton & Friedl, 2013). Another method of delineation is to identify single pixels as agricultural or non-agricultural vegetation using statistical correction analysis (Maselli & Rembold, 2002). Among those studies, there are soybean forecasts in the United States using remote sensing (Lobell & Asner, 2003; Prasad et al., 2006). J. Chang et al. (2007) focus on using NDVI to map corn and soybean farmland.

Fieuzal et al. (2017) make corn yield forecasts using both a real-time approach and a diagnostic approach. The real-time approach updates the estimates dynamically after the newest

image is acquired whereas the diagnostic approach utilizes all the image data throughout the season. The authors find the two best estimates perform comparably. Burke and Lobell (2017) regresses the agreement between satellite-based yields and field-reported yields as a function of farm size and find the vegetation index can most accurately predict crop yield when the field size is large.

The above-mentioned literature all employ a global model to produce the regression result that fits all observations, with the major difference among the studies being the specific model used. To our best knowledge, our paper is the first one to employ models that produce site-specific regression results, allowing heterogeneous response of soybean yield across counties. This is also the first paper to our knowledge that applies the Flexible Fourier Transform Model to examine the Yield-NDVI relationship.

3. Data and Methods

3.1 Data

We use data from 797 counties from ten major soybean producing states in the U.S. from 2000 to 2016. According to NASS, the soybean production from these ten states accounted for 77-83% of total soybean production in the U.S. from 2000 to 2016 (see Table 1 for soybean production and yield by state). Mkhabella et al. (2011) state that if a crop is not the dominant crop in the region, NDVI usually gives a poor prediction of crop yield, because it cannot distinguish between different crops. The soybean yield data are obtained from the USDA-NASS QuickStats. This database provides official published aggregate statistics on U.S. soybean yields and value of production of soybeans. Soybean yield is measured in bushels per acre. County-level NDVI data are obtained from the United States Geological Survey (USGS) and Ag-

Analytics. Ag-Analytics is an open-source, open-access database that provides data on agricultural finance, environmental finance, insurance, and risks (Woodard, 2016). The USGS uses eMODIS and the satellite platform Terra to obtain images at the resolution of 250 meters from 2000 onwards. Ag-Analytics converts the 250m-resolution raw images to county-level NDVI mean values every 7 days. We calculate county-level monthly NDVI values by taking an average of the weekly NDVI values. Climatological data obtained from PRISM Climate Data from Oregon State University and Ag-Analytics. We include two weather variables: maximum temperature over a month and average monthly precipitation. County boundary shapefiles are obtained from the United States Census Bureau. We obtain a sample of 12,027 county-year observations for the FFT analysis.

[Place Table 1 approximately here]

3.2 Flexible Fourier Transform Model

When estimating crop yield response to input variables (for fertilizer input see Li et al.,2016), traditional models use regional and temporal dummies to capture spatial and inter-temporal heterogeneity. Adding dummy variables can only capture the difference in the value of the dependent variable across locations and time; however, it does not take into account how the relationship varies according to site-specific and time-specific characteristics. Another type of model uses a quadratic functional form to estimate the relationship between crop yield and weather variables, assuming that crop yield is non-linearly related to the weather variable. However, these models may suffer from model misspecification, especially if there is a threshold effect, driven by environmental risks such as drought and flooding (Cooper et al., 2017).

Gallant (1984) first proposed flexible Fourier functional transform to generate unbiased production function approximation and proved its mathematical validity. Cooper et al. (2017) applied a flexible Fourier transform function to estimate the relationship between crop yield and temperature. The flexible Fourier function we use can be presented as follows:

$$y = u_0 + x'b + x'Dx + 2\sum_{\alpha=1}^{A}(\sum_{j=1}^{J}(v_{j\alpha}\cos[jk_\alpha's(x_{FFT})] - w_{j\alpha}\sin[jk_\alpha's(x_{FFT})])) + \varepsilon \quad (1)$$

In this model, y is the dependent variable soybean yield; u_0 is the constant term; x denote the independent variables; $s(x)$ is the scaled version of x, such that $s(x)$ is in the range of $[0, 2\pi]$; and x_{FFT} denote transformed variables, in our case, NDVI. The first two terms represent the linear regression part. D is a parameter matrix to be estimated. The third term represents the quadratic term. The last term models the functional flexibility using FFT. Similar to the Taylor expansion which uses a series of polynomial terms to approximate the true function, the Fourier function uses a series of trigonometric terms to approximate the true function. The Fourier functional form is believed to be the only known functional form that satisfies the Sobolev condition, meaning that the difference between the approximated function and the true function approaches zero as the sample size becomes arbitrarily large. For a proof that the Fourier function satisfies the Sobolev condition, refer to Gallant (1994). In the model, k_α (α=1, 2, …, A) is the elementary multi-index vector, whose dimension equals the dimension of x_{FFT}, whereas A is the total number of elementary multi-indexes. The vector k_α can be obtained in the following way: first, exhaust the list of k_α, such that k_α has only integer elements and the sum of the absolute value of each element in k_α is no greater than K, where K is predetermined; second, delete any k_α whose first non-zero element is negative; third, delete any k_α whose elements have

a common integer divisor. Monahan (1981) introduced a Fortran code to produce the set of elementary multi-index vectors. Also in the model, *J* is the order of the Fourier transformation whereas $v_{j\alpha}$ and $w_{j\alpha}$ are parameters to be estimated. We use the following parametrization: $K = 2, J = 2$, which are chosen such that the rule of thumb — the number of variables after transformation is roughly the square root of the number of observations (Fenton & Gallant, 1996) — is satisfied. Since there are 12,027 observations in the data we use, we include a total of 120 variables after the adding the transformed NDVI variables.

The regression equation we use in our study can be simply written as follows:

$$\text{soybean yield} = \beta_0 + \sum_{m=April}^{August}(\beta_{1m}MaxTemp_m + \beta_{2m}MaxTempSquare_m) + \sum_{m=April}^{August}(\beta_{3m}Precipitation_m + \beta_{4m}PrecipitationSquare_m) + \sum_{m=April}^{September}(\beta_{5m}NDVI_m) + \delta_0 TimeTrend + \sum_{s=1}^{9}\delta_s StateDummy_s + 2\sum_{\alpha=1}^{A}\sum_{j=1}^{J}(v_{j\alpha}\cos[jk'_\alpha s(NDVI)] - w_{j\alpha}\sin[jk'_\alpha s(NDVI)]) + error \quad (2)$$

where $MaxTemp_m$, $Precipitation_m$, $NDVI_m$ are the maximum temperature, the average precipitation, the average NDVI in month *m*, respectively. $PrecipitationSquare_m$ and $MaxTempSquare_m$ are the squared terms of $MaxTemp_m$ and $Precipitation_m$. *TimeTrend* equals the year minus 1999. $StateDummy_s$ is the state dummy variable. *NDVI* is a vector with each element being $NDVI_m$. We include the weather variables from April to August, following the standard specification in the literature (Cooper et al., 2017). We include NDVI variables through September, following the remote sensing literature (Li et al., 2007). The advantage of the flexible Fourier transform function is that not only does it allow for model flexibility, but also it incorporates multivariate estimation which is difficult to achieve through other non-parametric models such as kernel regression. The model degenerates to the traditional OLS model with

quadratic terms when $v_{j\alpha} = 0$ and $w_{j\alpha} = 0$. In the following discussion, the OLS model is the above model with $v_{j\alpha} = 0$ and $w_{j\alpha} = 0$ imposed. By testing the statistical significance of variable $v_{j\alpha}$ and $w_{j\alpha}$, we can conclude whether the traditional quadratic model is rejected in favor of the more flexible FFT model.

A review of the relevant literature reveals that the FFT model has been used/tested by scholars in different studies, fields, and situations. Chang (2016) used the FFT to model the non-linear effect of temperature on electricity demand. Becker et al. (2006) proposed a unit root test with a Fourier functional transform. Enders and Li (2015) approximated structural breaks in US GDP trends using Fourier forms. Jones and Enders (2014) provided a summary on using Fourier forms to model structural breaks.

3.3 Prediction and Forecast

We compare the prediction performance of the FFT model versus the OLS model. We conduct out-of-sample predictions and evaluate the prediction performance by comparing prediction errors measured by the root mean square error (RMSE) and the mean absolute error (MAE), between FFT and OLS, for four schemes: time-series prediction, cross-sectional prediction, panel prediction, and dynamic prediction. RMSE and MAE are defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \quad (4)$$

Both RMSE and MAE are commonly used measures to evaluate prediction performance. They measure the difference between true and fitted values. The unit for both RMSE and MAE is bushels per acre. In a time-series prediction, we first select a year for prediction, then we use observations from all other years to generate the model, then we predict the soybean yield for the selected year using the fitted model, weather data, and NDVI data from the selected year. In cross-sectional prediction, similarly, we select a state for prediction, then we use observations from all other states to generate the model, then we predict the soybean yield for the selected state using the fitted model, weather data, and NDVI data from the selected state; in panel prediction, similarly, we make the prediction for a selected year-and-state. Though commonly used, a shortcoming of using RMSE or MAE to measure prediction performance is that we do not know whether the predicted yield overestimates or underestimates the final actual yield.

In dynamic prediction or forecast, which is more realistic, we make predictions in each month throughout the growing season (May to September). For example, we produce the May prediction using only the information that is available until May. In our case, we use the precipitation, temperature and vegetation index from April and May. For each following month, the prediction is updated in a dynamic fashion by adding the most up-to-date weather and NDVI data into the model.

We make predictions and forecasts using the regression results from above-mentioned models. In this paper, prediction refers to cases where we may use data afterwards to predict for a specific time; forecast refers to case where we only use data up to a certain year to make predictions for that year.

4. Results

4.1 Descriptive analysis

The descriptive statistics for the main variables are reported in Table 2. The average soybean yield across all states and years is 43.11 bushels per acre. From April to July, the average maximum temperature and average NDVI increase steadily, and reach their peak levels in August. The average precipitation is highest in the months of May and June. These variables are included as suggested by the modified Thompson model (Thompson, 1963) to account for weather effects.

[Place Table 2 approximately here]

4.2 Flexible Fourier Transform Regression Results

All FFT models are developed using Matlab R2017a (The MathWorks, Inc.), following the methodology in Cooper et al. (2017). Figures showing FFT results are made using ArcMap 10.3 software. The estimation results from the model incorporating FFT terms are reported in Table 3. Due to the substantial number of variables (including 84 transformed NDVI variables), we only report the results for main variables including the untransformed weather variables and NDVI variables. However, the rest of the transformed variables are also included in the model fitting process. We calculate elasticities by applying the mean value theorem to get the numerical approximation of the derivatives and fixing the values of independent variables at the median level of each variable for each county. Thus, we obtain an elasticity estimate for each county. We present the minimum, median and maximum of FFT elasticity estimates across counties in column 2 through column 4 in Table 3. For comparison purposes, we also use the OLS regression results to calculate elasticity estimate for each county and report elasticity summary from OLS regression in column 5 through column 7 in Table 3. The OLS model refers to

equation (1) with $v_{j\alpha} = 0$ and $w_{j\alpha} = 0$ imposed. For weather variables, except for the July maximum temperature and the April average precipitation, the median of elasticity estimates from OLS and the median of elasticity estimates from FFT have the same sign. On average, higher temperatures from April to June and higher precipitation levels from June to August lead to higher soybean yields. On the other hand, higher temperatures in August and higher precipitation levels in May are associated with lower soybean yields.

While the median of elasticity estimates for weather variables across counties are very similar between FFT results and OLS results, median elasticity estimates of NDVI variables differ dramatically between OLS results and FFT results, in terms of both sign (September NDVI) and magnitude (April-August NDVI). NDVI elasticities estimated by FFT have a wider range than those generated by OLS, due to the inclusion of the transformed NDVI variables. OLS results suggest that August NDVI has a greater impact on soybean yields than the July NDVI, whereas FFT results suggest the opposite. According to Table 3, when July (August) NDVI increases by 10 percent, the median soybean yield rises by 4.5% (3.4%) or by 1.94 (1.47) bushels per acre.

By testing the significance of the coefficient estimates for the Fourier terms, we can test whether the FFT specification is overfitting the data. In Table 3, we present an F test of the FFT regression versus the OLS regression and we find that the coefficients on the transformed Fourier terms are jointly significantly different from zero and thus the OLS is rejected in favor of the FFT regression.

[Place Table 3 approximately here]

[Place Table 4 approximately here]

The geographical distribution of coefficient estimates from FFT is presented in Figure 1. In each subfigure, we present the geographical distribution of the median of the elasticity estimates of NDVI for each month (April, May, June, July, August and September, respectively) across different counties. For some counties in the north of North Dakota, central Minnesota, central Indiana, west Arkansas, and southwest Missouri, soybean yields are highly responsive to July NDVI, but less responsive to August NDVI. For most counties in Ohio and in east Arkansas, in contrast, the soybean yield is responsive to August NDVI whereas it is less responsive to July NDVI. For some counties in the west parts of North Dakota and South Dakota, soybean yields are responsive to April NDVI whereas they are less responsive to August NDVI. These geographical differences in soybean yield responsiveness to NDVI shows that there is considerable spatial heterogeneity that needs to be considered when making yield predictions.

4.3 Prediction and Forecast Results

The results of time-series prediction and cross-sectional prediction performance for FFT versus OLS are shown in Table 4. The bold numbers show cases where the FFT error is lower than the OLS error. On average, FFT performs better than OLS in time series predictions since both MAE and RMSE for FFT is lower than those for the OLS model. For cross-sectional predictions, FFT has a higher RMSE on average, but a lower MAE than OLS does.

[Place Table 4 approximately here]

Our results show that time-series predictions on average are more accurate than cross-sectional predictions in terms of smaller predicting error. RMSE and MAE from time-series predictions are consistently lower than cross-sectional predictions.

We then conduct out-of-sample panel prediction. We randomly select 1,000 observations from all years and states, and predict the soybean yields for these 1,000 observations by OLS and

FFT, using all other observations excluding these 1,000 observations. We then compare the predicted soybean yields with the actual yields and calculate the RMSE and MAE. We then repeat this sampling process 200 times. The histogram shown in Figure 2 is of the distribution of RMSE and MAE. Two findings are interesting. First, panel prediction has much lower prediction error than both time-series and cross-sectional predictions in Table 4. This suggests that when predicting soybean yield for a certain location, it is useful to include the already publicized yield data from other locations into the training sample. Second, FFT has a consistently lower prediction error than the OLS model. FFT can improve the prediction performance by a modest 0.3% according to MAE, or 0.4% according to RMSE. This percentage is obtained by dividing the prediction error by the mean of crop yield (average MAE is 0.138, average RMSE is 0.1684 and mean soybean yield is 43.11).

[Place Figure 2 approximately here]

The predictions so far may have used data from future periods to predict current soybean yields. For robustness, we also include forecasts where soybean yield predictions are only based on data from previous periods (Table 5). For RMSE, there are 10 years out of 16 years where FFT outperforms OLS. For MAE, there are 12 years out of 16 years in which FFT outperforms OLS. In terms of average error, FFT has smaller RMSE and MAE than OLS does. While the forecasts are more realistic in terms of being based only on data from previous periods, the average prediction errors are unsurprisingly higher than those for the predictions using all data including from future periods in Table 4.

5. Conclusions

In this paper, we used FFT to account for spatial and temporal heterogeneity in the relationship between NDVI throughout the growing season and soybean yield. We produced county-specific coefficients and elasticities of NDVI on soybean yield. We found that the response of soybean yield to NDVI is different across locations, showing spatial heterogeneity of responses. For some counties located in the northern states, soybean yield is highly positively related with the July NDVI, whereas for other counties located in the south, the August NDVI is a better indicator of the soybean yield. Traditional OLS models seem to underestimate the effect of July and August NDVI on soybean yields.

Furthermore, we conducted out-of-sample prediction/forecast and compared performances for the OLS and FFT models. We found that models that account for the spatial heterogeneity generally result in better out-of-sample predictions and forecasts.

A limitation of this work is that it does not distinguish pixels of soybean crop from those of other crops or vegetation types, still incorporating NDVI in the model results in significant coefficients and improved fit. Future work can use filters to select pixels that are highly likely to be soybean crop. However, the use of spatial heterogeneous models may capture the heterogeneous soybean/total land ratios across counties, compared to OLS, thus alleviating the contamination caused by other crops. Future work that applies land cover filters may improve the results even further.

This study uses data from the ten major soybean producing states in the U.S. where data are readily available. Our results show that using the FFT model helps improve the prediction accuracy (lowers the prediction error) especially in panel predictions. The goal is to improve on the forecast accuracy of soybean yield to allow market participants to make more informed decisions with respect to anticipated crop yield and possible resulting prices. The FFT model

also has potential to forecast crop yields in less developed countries where ground field work is too expensive to conduct or where the meteorological network is sparse – making this an alternative feasible solution in making crop yield predictions.

References

Adrian, D. (2012). *A model-based approach to forecasting corn and soybean yields.* Paper presented at the Fourth International Conference on Establishment Surveys.

Basnyat, P., McConkey, B., Meinert, B., Gatkze, C., & Noble, G. (2004). Agriculture field characterization using aerial photograph and satellite imagery. *IEEE Geoscience and Remote Sensing Letters, 1*(1), 7-10.

Becker, R., Enders, W., & Lee, J. (2006). A stationarity test in the presence of an unknown number of smooth breaks. *Journal of Time Series Analysis, 27*(3), 381-409.

Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology, 173*, 74-84.

Bose, P., Kasabov, N. K., Bruzzone, L., & Hartono, R. N. (2016). Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series. *IEEE Transactions on Geoscience and Remote Sensing, 54*(11), 6563-6573.

Burke, M., & Lobell, D. B. (2017). Satellite-based assessment of yield variation and its determinants in smallholder African systems. *Proceedings of the National Academy of Sciences, 114*(9), 2189-2194.

Chang, J., Hansen, M. C., Pittman, K., Carroll, M., & DiMiceli, C. (2007). Corn and Soybean Mapping in the United States Using MODIS Time-Series Data Sets. *Agronomy Journal, 99*(6), 1654-1664.

Chang, Y., Kim, C. S., Miller, J. I., Park, J. Y., & Park, S. (2016). A new approach to modeling the effects of temperature fluctuations on monthly electricity demand. *Energy Economics, 60*(Supplement C), 206-216.

Cooper, J., Nam Tran, A., & Wallander, S. (2017). Testing for Specification Bias with a Flexible Fourier Transform Model for Crop Yields. *American Journal of Agricultural Economics, 99*(3), 800-817.

Doraiswamy, P. C., & Cook, P. W. (1995). Spring wheat yield assessment using NOAA AVHRR data. *Canadian Journal of Remote Sensing, 21*(1), 43-51.

Enders, W., & Li, J. (2015). Trend-cycle decomposition allowing for multiple smooth structural changes in the trend of US real GDP. *Journal of Macroeconomics, 44*(Supplement C), 71-81.

Fenton, V. M., & Gallant, A. R. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics, 74*(1), 77-118.

Ferencz, C., Bognar, P., Lichtenberger, J., Hamar, D., Tarcsai, G., Timar, G., . . . Szekely, B. (2004). Crop yield estimation by satellite remote sensing. *International journal of remote sensing, 25*(20), 4113-4149.

Fieuzal, R., Sicre, C. M., & Baup, F. (2017). Estimation of corn yield using multi-temporal optical and radar satellite data and artificial neural networks. *International Journal of Applied Earth Observation and Geoinformation, 57*, 14-23.

Gallant, A. R. (1984). The Fourier flexible form. *American Journal of Agricultural Economics, 66*(2), 204-208.

Gallant, A. R. (1994). Identification and consistency in semi-nonparametric regression. *Advances in Econometrics, 1*, 145-169.

Irwin, S. H., Sanders, D. R., & Good, D. L. (2014). Evaluation of Selected USDA WAOB and NASS Forecasts and Estimates in Corn and Soybeans. *Marketing and Outlook Research Report, 1*.

Johnson, D. M. (2014). An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. *Remote sensing of Environment, 141*, 116-128.

Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. (2016). Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology, 218*, 74-84.

Jones, P. M., & Enders, W. (2014). On the use of the flexible Fourier form in unit root tests, endogenous breaks, and parameter instability *Recent Advances in Estimating Nonlinear Models* (pp. 59-83): Springer.

Kaul, M., Hill, R. L., & Walthall, C. (2005). Artificial neural networks for corn and soybean yield prediction. *Agricultural Systems, 85*(1), 1-18.

Li, A., Liang, S., Wang, A., & Qin, J. (2007). Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques. *Photogrammetric Engineering & Remote Sensing, 73*(10), 1149-1157.

Lobell, D. B., & Asner, G. P. (2003). Climate and Management Contributions to Recent Trends in U.S. Agricultural Yields. *Science, 299*(5609), 1032-1032.

Lv, X. (2014). Remote sensing, normalized difference vegetation index (NDVI), and crop yield forecasting (Master's Thesis).

Maselli, F., & Rembold, F. (2002). Integration of LAC and GAC NDVI data to improve vegetation monitoring in semi-arid environments. *International journal of remote sensing, 23*(12), 2475-2488.

Mkhabela, M., Bullock, P., Raj, S., Wang, S., & Yang, Y. (2011). Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agricultural and Forest Meteorology, 151*(3), 385-393.

Monahan, J. F. (1981). Enumeration of elementary multi-indices for multivariate Fourier series. *Institute of Statistics, North Carolina State University, Raleigh, Mimeograph Series*(1338).

Prasad, A. K., Chai, L., Singh, R. P., & Kafatos, M. (2006). Crop yield estimation model for Iowa using remote sensing and surface parameters. *International Journal of Applied Earth Observation and Geoinformation, 8*(1), 26-33.

Senay, G. (2016). The Power of Remote Sensing: Global monitoring of weather, water, and crops with satellites and data integration.

Thompson, L. M. (1963). Weather and Technology in the Production of Corn and Soybeans.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment, 8*(2), 127-150.

Woodard, J. (2016). Big data and Ag-Analytics: An open source, open data platform for agricultural & environmental finance, insurance, and risk. *Agricultural Finance Review, 76*(1), 15-26.

Table 1. Soybean Production and Yield in 10 Major Producing States for 2014-2016

State	Soybean Production	Soybean Yield	Soybean Production	Soybean Yield	Soybean Production	Soybean Yield
	2014		2015		2016	
Illinois	547,120	56	544,320	56	592,950	59
Iowa	498,270	51	553,700	56.5	571,725	60.5
Minnesota	301,705	41.5	377,500	50	393,750	52.5
Indiana	301,920	55.5	275,000	50	324,300	57.5
Nebraska	287,820	54	305,660	58	314,150	61
Missouri	259,935	46.5	181,035	40.5	271,460	49
Ohio	246,225	52.5	237,000	50	263,780	54.5
South Dakota	229,950	45	235,520	46	255,915	49.5
North Dakota	202,515	34.5	185,900	32.5	249,000	41.5
Arkansas	158,400	49.5	155,330	49	145,700	47
Ten states	3,033,860	48.6	3,050,965	49	3,382,730	53
U.S. Total	3,927,090	47.5	3,926,339	48	4,306,671	52.1

Note: Soybean production is measured in 1,000 bushels, Soybean yield is measured in bushels/acre.

Table 2. Descriptive Statistics

Variable	Number of Obs.	Min.	Median	Max.	Mean	Standard Deviation
Soybean Yield	12,027	2.9	44	73.1	43.11	10.05
Max Temp. April	12,027	0.37	17.32	27.34	17.06	3.44
Max Temp. May	12,027	13.65	22.33	30.67	22.39	2.56
Max Temp. June	12,027	19.7	27.37	36.06	27.37	2.28
Max Temp. July	12,027	22.82	29.41	38.91	29.58	2.4
Max Temp. August	12,027	20.13	28.76	39.47	28.92	2.35
Precipitation April	12,027	4.17	85.14	424.08	91.26	48.98
Precipitation May	12,027	5.27	108.28	355.26	113.08	52.23
Precipitation June	12,027	7.64	105.01	376.5	115.8	58.07
Precipitation July	12,027	0.89	87.5	354.27	94.43	49.66
Precipitation August	12,027	0	82.04	438	90.01	51.54
NDVI April	12,027	-0.01	0.33	0.79	0.35	0.12
NDVI May	12,027	0.13	0.42	0.85	0.45	0.13
NDVI June	12,027	0.24	0.59	0.87	0.58	0.1
NDVI July	12,027	0.27	0.74	0.89	0.73	0.08
NDVI August	12,027	0.27	0.75	0.88	0.72	0.1
NDVI September	12,027	0.24	0.6	0.87	0.6	0.1

Note: Temperatures are measured in degrees Celsius, precipitation is measured in inches. Negative NDVI denotes snow cover.

Table 3. Elasticity Estimates from FFT and Quadratic OLS Models

	FFT			Quadratic OLS		
	Min.	Median	Max.	Min.	Median	Max.
Max. Temp. April	0.02	0.08	0.29	0.04	0.07	0.21
Max. Temp. May	-0.75	0.22***	3.74	-0.82	0.27***	1.38
Max. Temp. June	-0.11	0.42***	5.99	-0.29	0.39***	1.62
Max. Temp. July	-0.94	-0.04***	1.14	-1.1	0.04***	0.86
Max. Temp. August	-3.82	-0.48**	-0.3	-1.46	-0.53	-0.37
Precipitation April	-0.03	0.0013***	0.04	-0.04	-0.0047**	0.0048
Precipitation May	-0.15	-0.01*	0.01	-0.18	-0.01*	0.004
Precipitation June	-0.05	0.03***	0.3	-0.07	0.04***	0.08
Precipitation July	-0.29	0.04***	0.35	-0.44	0.05***	0.11
Precipitation August	0.01	0.09***	0.53	0.03	0.09***	0.16
NDVI April	-3.27	-0.03***	2.08	-0.22	-0.07***	-0.04
NDVI May	-1.04	-0.06*	1.09	-0.22	-0.09***	-0.04
NDVI June	-8.62	-0.15***	1.14	-0.01	-0.01	-0.0032
NDVI July	-2.27	0.45***	8.36	0.08	0.14***	0.26
NDVI August	-3.74	0.34	7.46	0.13	0.22***	0.32
NDVI September	-5.11	0.09	2.48	-0.11	-0.06***	-0.04
No. of Obs.		12,027			12,027	
State Fixed Effects		Yes			Yes	
Year Trend Effects		Yes			Yes	
Adjusted R-sq		0.721			0.701	
Rank test between Fourier and OLS		F (84,11906) = 11.132				

Notes: Due to the non-linearity of the FFT regression, we report the elasticity estimates rather than the coefficient estimates of the main variables.
Significance here indicated by asterisks corresponds to the significance of untransformed variables.
In addition to these variables, additional 84 Fourier transformed variables of NDVI are included in the analysis - their coefficient estimates are not reported here but they are included in the elasticity calculations.

Table 4. Out-of-Sample Prediction Performance: Time-Series and Cross-Sectional Prediction

	MAE		RMSE			MAE		RMSE	
Year	OLS	FFT	OLS	FFT	State	OLS	FFT	OLS	FFT
2000	4.81137	4.98073	6.07073	6.21356	North Dakota	**5.42118**	**4.92597**	**6.50988**	**6.00264**
2001	3.87899	3.9363	4.92053	5.02227	South Dakota	5.35795	6.27265	6.88285	8.72918
2002	**4.82763**	**4.80833**	**6.23882**	**6.1961**	Iowa	**4.80132**	**4.07755**	**5.85146**	**5.11804**
2003	**5.32093**	**5.02482**	**6.69649**	**6.38478**	Ohio	4.31768	4.71736	5.3346	5.77125
2004	4.24005	4.35191	5.42675	6.02137	Illinois	**6.19968**	**5.57796**	**7.58194**	**6.94729**
2005	**4.26876**	**4.13705**	**5.42658**	**5.23434**	Indiana	**4.61708**	**4.46813**	**5.54583**	**5.45675**
2006	**4.37803**	**4.14817**	**5.55463**	**5.34228**	Nebraska	10.0521	10.3535	12.769	13.22
2007	4.61954	4.74141	6.20303	6.33831	Minnesota	4.97922	4.98107	6.34346	6.53194
2008	4.07258	4.25072	5.22737	5.44965	Missouri	4.68498	4.72269	5.91926	5.92473
2009	**4.3971**	**4.19245**	**5.75316**	**5.41486**	Arkansas	**7.61327**	**7.26225**	**9.56978**	**9.22424**
2010	**3.85706**	**3.75677**	**4.92539**	**4.8753**	Average	**5.80444**	**5.73592**	7.23081	7.29261
2011	4.37394	4.53074	5.57265	5.67107					
2012	**5.92868**	**5.73895**	**7.46748**	**7.32792**					
2013	**4.9454**	**4.89378**	**6.18927**	**6.1736**					
2014	**4.25676**	**4.2375**	**5.40185**	**5.37212**					
2015	**4.492**	**4.46359**	**5.8391**	**5.79755**					
2016	**5.35846**	**5.33611**	6.5839	6.60164					
Average	**4.58984**	**4.56055**	**5.85281**	**5.84922**					

Note: Bold numbers indicate that FFT has lower prediction errors and therefore outperforms OLS.

Table 5. Out-of-Sample Forecast Performance

	MAE		RMSE	
Year	OLS	FFT	OLS	FFT
2001	**11.136**	**10.636**	**12.882**	**12.678**
2002	6.303	6.502	7.956	8.235
2003	**4.696**	**4.327**	**6.188**	**5.639**
2004	**5.956**	**5.395**	7.117	7.394
2005	**7.201**	**6.979**	**8.448**	**8.239**
2006	**4.758**	**4.546**	**6.319**	**6.051**
2007	5.297	5.559	6.865	7.157
2008	4.242	4.758	5.475	6.143
2009	**4.240**	**4.074**	**5.495**	**5.203**
2010	**4.353**	**4.277**	**5.545**	**5.489**
2011	**5.132**	**4.911**	**6.765**	**6.410**
2012	**6.239**	**6.124**	**7.835**	**7.730**
2013	4.845	4.994	6.080	6.247
2014	**4.240**	**4.218**	**5.388**	**5.335**
2015	**4.488**	**4.383**	**5.854**	**5.748**
2016	**5.358**	**5.336**	6.584	6.602
Average	**5.530**	**5.439**	**6.925**	**6.894**

Note: Bold numbers indicate that FFT has lower forecast errors and therefore outperforms OLS.

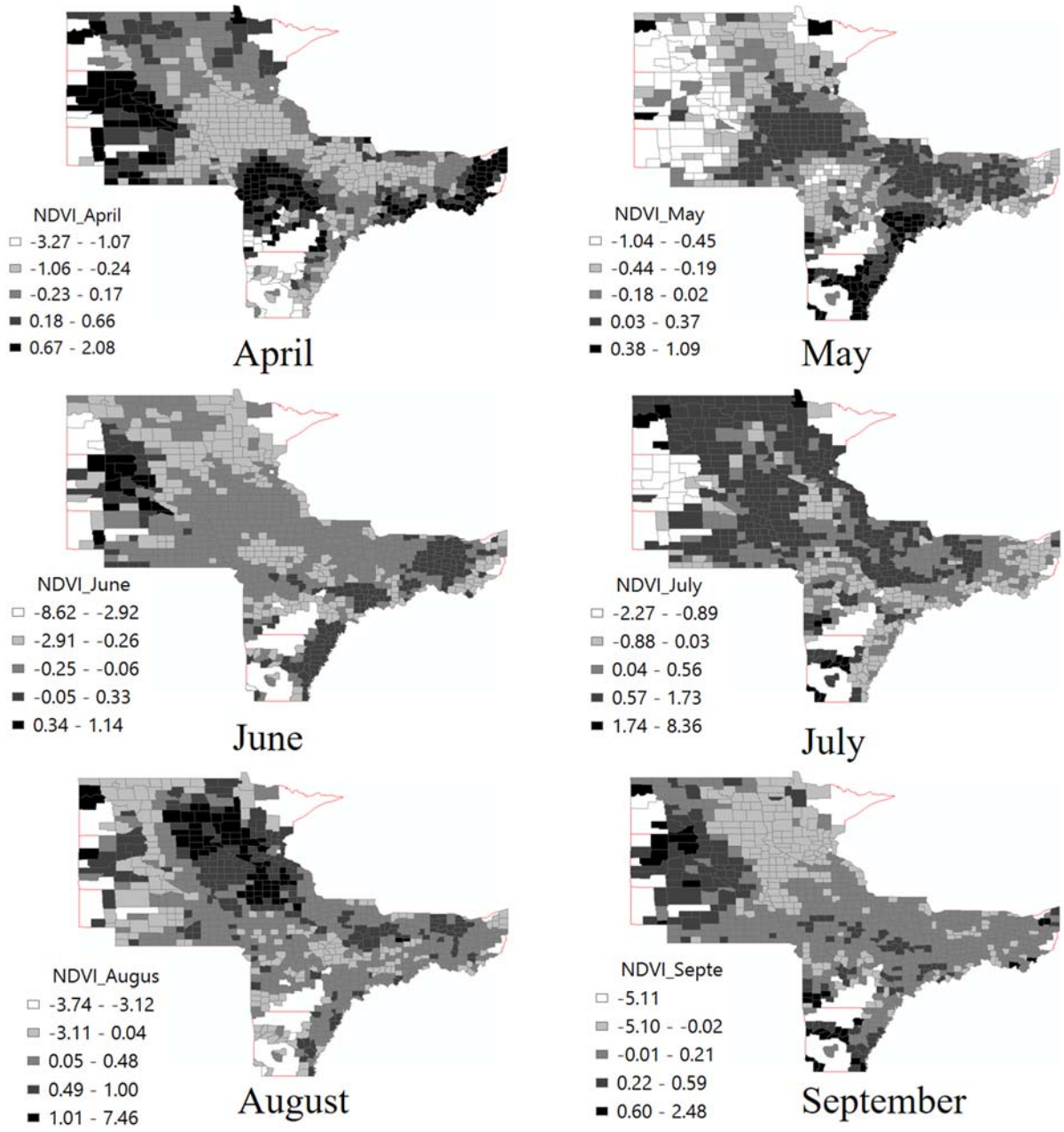

Figure 1. Geographical Distribution by State of Elasticity Estimates from FFT, April-September

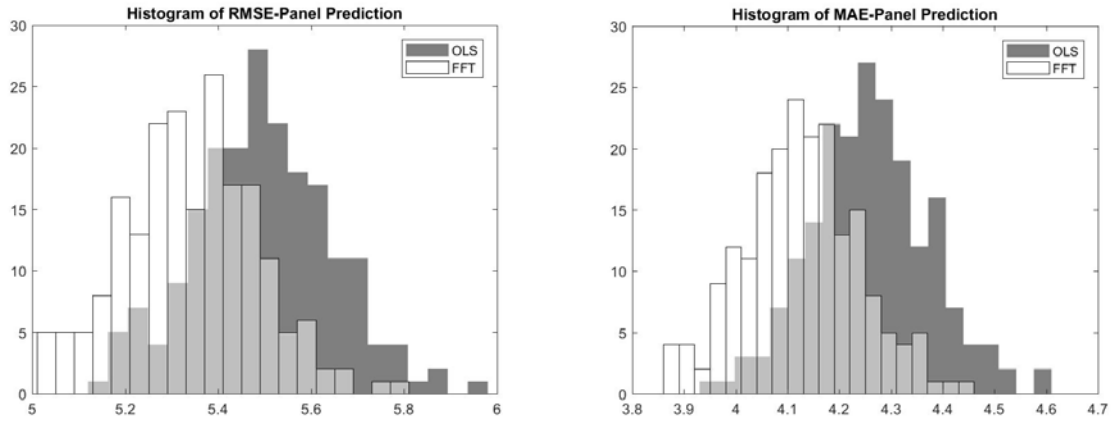

Figure 2. Histogram of RMSE and MAE between OLS and FFT