Modeling Wheat Yields in India: Describing & Exploiting spatiotemporal Variability with Panel Regression

The use of remote sensing for modeling and prediction of yields in developed countries has seen substantial academic and commercial interest. Here we aim to provide an open-source suite of algorithms to rapidly capture and summarize portions of the phenological process relevant to crop modeling while maintaining spatiotemporal structure for use in panel econometric models. Here we demonstrate the predictive power of both these algorithms for summarizing remotely sensed data, and that of panel econometric methods to estimate wheat yields at the district level in Punjab and Haryana India. We find that our algorithms coupled with panel econometric methods can accurately predict yields both across districts, and within any given district over time, even during periods of drought(>0.8).

Michael L. Mann, Berber Kramer, Francisco Ceballos, Miguel Robles, Josh Gray, Eli Melaas

Department of Geography, The George Washington University, Washington DC

Markets, Trade and Institutions Division, International Food Policy Research Institute, Washington DC

The World Bank, Washington DC

Department of Forestry and Environmental Resources, North Carolina State University, Raleigh NC

Department of Earth and Environment, Boston University, Boston MA

##### 

# Introduction

The ability to monitor and predict crop yields in developing countries is critical to the successful adaptation to changes in our climate. Increased temperatures and variability has already been linked to losses in maize and wheat yields (-3.8 and 5.5% respectively)and crop prices globally (David B. Lobell, Schlenker, & Costa-Roberts, 2011). Although much effort has been placed on modeling the spatial distribution of these shifts, less effort has been placed on how yields vary across space and time (Ray, Gerber, MacDonald, & West, 2015). Advances in remote sensing provide avenues to monitor agricultural crop health at high spatial and temporal resolution. However, our ability to monitor changes in plant productivity is still limited in the more complex environments common to many developing countries (ML L Mann & Warner, 2015).

Remote sensing based efforts to characterize the extent, cultivation practices, and productivity of global croplands has a long history. In fact, agricultural monitoring motivated much of the earliest work in remote sensing for example NASA's LACIE and AgRISTARS programs in the 1970s and 1980s and (R. B. MacDonald & Hall, 1980, J. Hatfield (1983), NASA (1984),Pinter Jr, Ritchie, Hatfield, & Hart (2003))). Since then, substantial progress has been made in mapping cropland extent, crop types, irrigation status, cropping intensity, and productivity from remotely sensed imagery. For example, the MODIS Land Cover Product MCD12Q1 (M. A. Friedl et al., 2002, M. A. Friedl et al. (2010b)) provides operationally produced, global scale maps of agriculture and agricultural-natural mosaics at an annual time step and 500 m spatial resolution from 2001-present. A finer resolution (~30 m) dataset is available for the conterminous United States which maps the annual extent and type for over 250 crops using primarily Landsat imagery: the Cropland Data Layer [NASS (2003). These are but two prominent examples out of a broad literature documenting a wide variety of efforts to map cropland extent and type from remotely sensed imagery (David B Lobell & Asner, 2004, Xiao et al. (2006), P. Thenkabail, GangadharaRao, Biggs, Krishna, & Turral (2007), Ramankutty, Evan, Monfreda, & Foley (2008), B. D. Wardlow & Egbert (2008), Biradar & Xiao (2011)). Remotely sensed imagery has also been employed to map irrigated areas (P. S. Thenkabail et al., 2009, Portmann, Siebert, & Döll (2010)), and cropping frequency/intensity (Biradar & Xiao, 2011, Gray et al. (2014), L. Li et al. (2014)).

Initial efforts (e.g. LACIE and AgRISTARS) primarily utilized remotely sensed imagery to characterize the spatial extent and growth stage of crops, but relied on models driven chiefly by meteorological information to predict crop yield (Idso, Jackson, & Reginato, 1977, Doraiswamy, Moulin, Cook, & Stern (2003)). However, the biophysical link between canopy spectral reflectance and net primary production has long been established (C. Tucker & Sellers, 1986); indicating that satellite measurements could play a role in determining crop yield directly. Indeed, early experimental work confirmed the usefulness of spectral measurements in predicting LAI and intercepted PAR in crops (Daughtry, Gallo, & Bauer, 1983, Asrar, Fuchs, Kanemasu, & Hatfield (1984), Clevers (1997)), a result that was later extended to satellite measurements of spectral reflectance (C. Tucker, Holben, Elgin Jr, McMurtrey III, & others, 1980, Groten (1993), Bartholome (1988)). Spectral measurements typically explain variability in LAI and intercepted PAR better than crop yields because a variety of factors other than net primary production (e.g. weather during critical crop growth stages) influence yield. Nevertheless, a wide number of studies have documented highly explanatory empirical relationships between satellite measures such as NDVI (in many forms: growing season maximum and mean, seasonally integrated, etc.) and yields for a variety of crops, particularly at regional scales (Rasmussen, 1992, Benedetti & Rossini (1993), Funk & Budde (2009), Becker-Reshef, Vermote, Lindeman, & Justice (2010b), Becker-Reshef et al. (2010a), Mkhabela, Bullock, Raj, Wang, & Yang (2011)). Because certain crop growth stages are particularly critical for final yield (Butler & Huybers, 2015), improved results are often seen when remotely sensed data are used to characterize crop phenology (Bolton & Friedl, 2013). More recently, methods for forecasting yields with remotely sensed variables at the field scale have been explored (David B Lobell, Thau, Seifert, Engle, & Little, 2015, Gao et al. (2017)). In addition to establishing a direct relationship between satellite measurements and crop yield, combining these observations with model output, through formal or ad hoc data assimilation techniques has also been demonstrated (Mo et al., 2005, Moriondo, Maselli, & Bindi (2007), De Wit & Van Diepen (2007)).

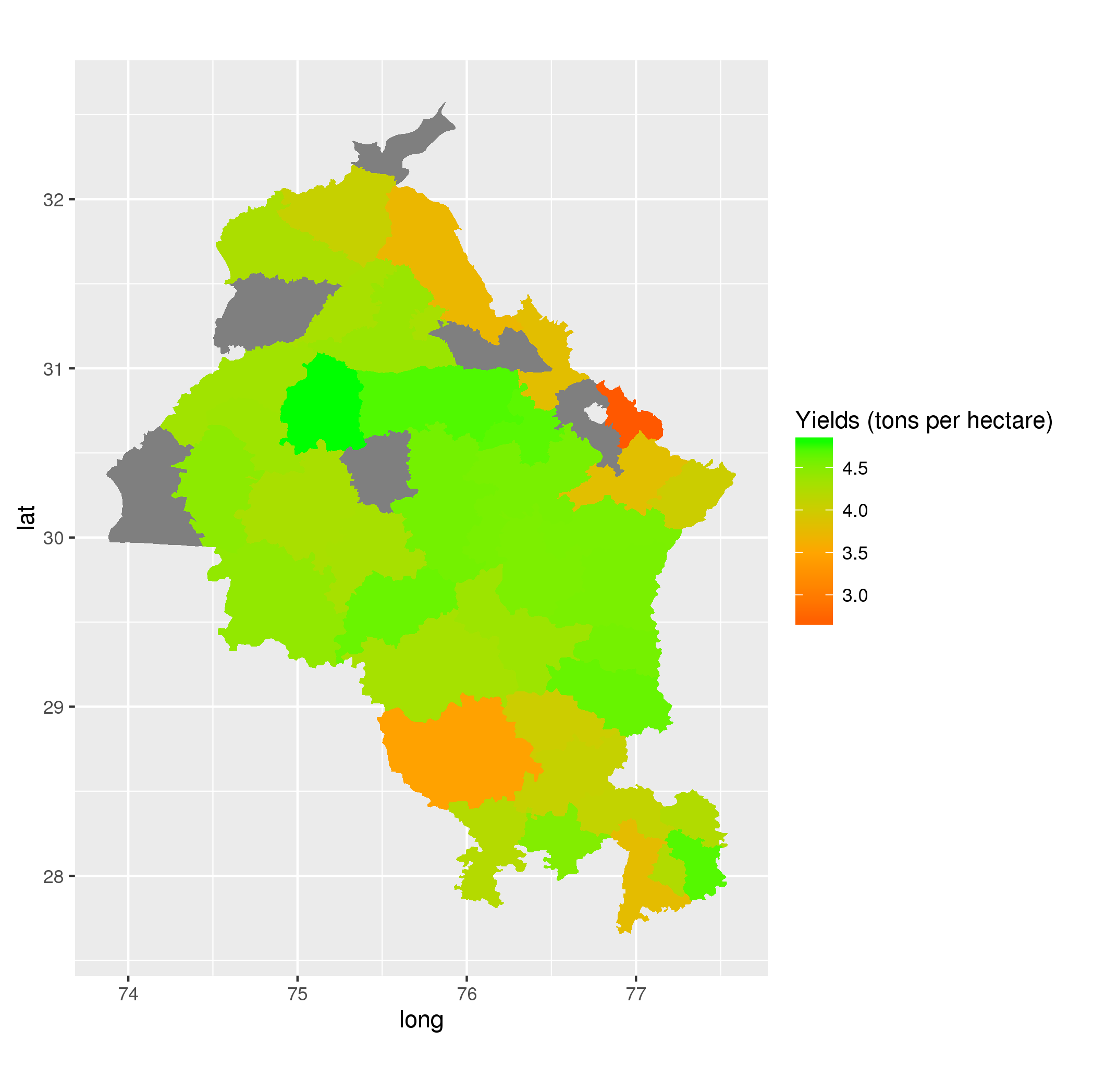
The main objectives of this paper is to compare statistical methods commonly applied these problems outside of the field of geography. In particular, we will focus on the application of spatial panel regression to monitor spatiotemporal variability in wheat yields for Punjab and Haryana India at the district level. We also created and present a suite of algorithms used to extract, summarize, and organize remotely sensed data and prepare it for spatiotemporal analysis.

All [algorithms](https://github.com/mmann1123/India-Index-Insurance-Code/blob/master/SummaryFunctions.R) as well as [the code](https://github.com/mmann1123/India-Index-Insurance-Code/blob/master/WriteUp/WheatYieldWriteUp.Rmd) used to generate the findings from this study are provided through an open github repository [here](https://github.com/mmann1123/India-Index-Insurance-Code).

# Methods

## Overview and Study Area

We examine wheat yields at the district level for Punjab and Haryana India for Rabi season (roughly Nov-Apr) for the period of 2002 to 2012. Both Punjab and Haryana are extensively cropped but are comprised of a large number of smaller heterogeneous plots. Both states are also extensively double-cropped with rice planting in the Kharif season (roughly May-Oct) and Wheat planted in the Rabi season. Rabi season wheat yield range from 1.88 to 5.68 metric tons per hectare (Figure 1).

*Figure 1: Mean Rabi Season Wheat Yields Metric Tons per Hectare by District*  

Here we develop a (non)spatial panel regression model to estimate wheat output per hectare using the open-source programming language R. This model utilizes historical data on plant phenology statistics obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) satellite. The objective is to develop a handful of metrics that can be used to accurate predict inter-annual variability in wheat yields at the district level.

## Data

The full model is comprised of 41 indicators of plant phenology. District level statistics are then generated from mean pixel level plant indicators.

### Focus Group Interviews

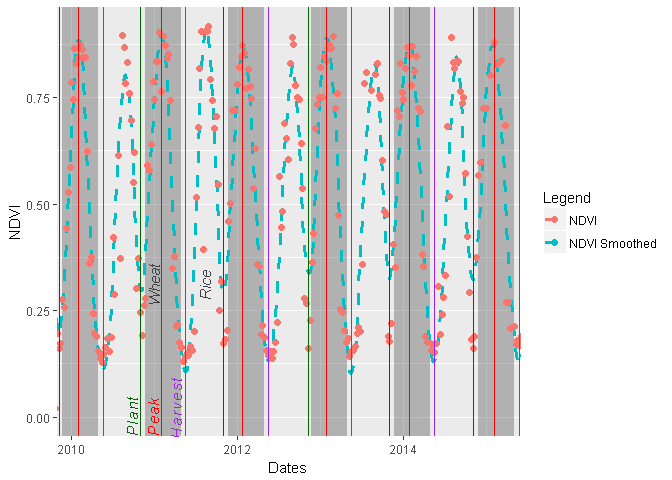
To help better characterize the physical properties and identify challenges, field visits and focus group interviews were conducted in the winter of 2015. These interviews were conducted in 12 villages with 71 participants in Haryana and Punjab states, in person with International Food Policy Research Institute (IFPRI) staff. Questions focused on farm characteristics, adopted technologies, Rabi crop calendar dates, and identifying the timing of risks to crops.

### Remote Sensing Data

Considering the relatively small scale of agriculture in this region (median plot sizes of 13.5 acres with a range of 2 to 17.2 acres) reported during focus groups (Robles, Ceballos, & Kramer, 2015), we utilized 250m vegetation products from the MODIS satellites. Vegetation indices are obtained from two 16 days MODIS products (MOD13Q1, MYD13Q1) from the Aqua and Terra satellites (Didan & Huete, 2006). Due to the staggered nature of acquisition these products are treated as partially overlapping windows representing 8 day periods (Doraiswamy, Stern, & Akhmedov, 2007). We find that the combination of these two products provides a stable and informative time series.

In particular, we examine the predictive power of the Normalized Difference Vegetation Index (NDVI) using panel econometric techniques. NDVI is sensitive to the amount of chlorophyll in any given pixel and is commonly used to estimate plant productivity and health in agricultural applications (ML L Mann & Warner, 2015, Funk & Budde (2009), Becker-Reshef et al. (2010b), Becker-Reshef et al. (2010a), Mkhabela et al. (2011)). After removal of snow, cloud and other flagged low quality cells, we remove all non-agricultural cells through the use of the 500m MODIS land cover product (MCD12Q1) for the appropriate year (M. A. Friedl et al., 2010a). The difference in resolutions is expected to have a minimal effect in this case because the extent of rural agriculture in these areas is extremely large. Therefore any cells include or excluded by omission or commission as agriculture should have a minimal effect at the district level. Moreover agricultural patterns are generally uniform over broad areas of Punjab and Haryana. This would likely no hold true in other areas of India.

In addition to cloud cover, MODIS data products suffer from four additional sources of error including atmospheric interference, georeferencing, bidirectional reflectance effect and differences in day of the year each pixel is observed (Doraiswamy et al., 2007). While indices such as NDVI minimize the effects of atmospheric distortion but will directly influence indices values. To minimize the effects of the artifacts described above we test the use of temporal smoothing splines and outlier removal (Hastie & Tibshirani, 1990, Gray et al. (2014)). Where outliers above three standard deviations are removed before applying a cubic smoothing spline. A visual example of the effects of the cubic smoothing and outlier removal procedure can be seen in Figure 2.

*Figure 2: (Un)smoothed 8-Day NDVI time signature for a dual cropped pixel in Punjab*  

*Time series of NDVI unsmoothed (red) and smoothed with outlier removal (blue) for both crop seasons of 2002 to 2016. Wheat growing season highlighted in dark grey, rice growing season in light grey. Date of estimated Rabi season maximum (red vertical line), estimated greenness onset (green vertical line), estimated harvest date (purple vertical line).*

### Agricultural Survey Data

Agricultural survey data at the district level was obtained from the Government of India (“District level agricultural yields,” n.d.). Yields are measured in tons per hectare. As described later, we limit the sample to the 2003-2012 period in order to obtain a balanced panel.

## Exploiting Time: Summarizing Remotely Sensed Data

One of the primary challenges in utilizing an 8-day time series to estimate annual wheat yields is the temporal mismatch in observations. Properties of the time series must be obtained to characterize and identify important components of plant phenology across the growing season correlated with wheat yields. Here we utilize 41 metrics to summarize phenology. These measures take two primary forms: first, growing season statistics, spanning the estimated planting date of wheat (mean DOY:315) until harvest date (mean DOY:135); and second monsoon season statistics, spanning end of the Rabi wheat season to end of the Kharif growing season (DOY 136-314). Two classes of statistics are estimated for these two periods: first, summary statistics (e.g. mean, max, variance), second, integrated summary statistics (e.g. area under the curve for the 1st 1/2 of the growing season), and third comparison to norms (e.g. comparisons to 95th percentile).

Pixels with in area of interest (AOI), district boundaries in this case, can be evaluated on a pixel by pixel basis in utilizing parallel processing or summarized by the AOI's mean value for each image. In this study, district-level mean values of NDVI for agricultural pixels are used to represent agricultural productivity for each 8-day period. The follow sections outline how these data are summarized for use in panel regression.

### Growing Season Metrics

Planting and harvest dates of are estimated for each growing season of interest. These dates are estimated through an iterative search algorithm finding the date of the global minimum NDVI value nearest to the *a priori* estimated date. A priori values were obtained from the focus group interviews described above. For wheat, sowing dates were reported to typically start in the last week of October, and harvest to begin in the 2nd week of April. For details on this see function 1 in the appendix. Basic growing season summary statics including minimum, maximum, mean, and standard deviation can be calculated using function 2 in the appendix below.

To estimate the cumulative impact of high or low vegetation indices across a season we calculate a variety of integration metrics. These include area under the curve (AUC) of the growing season, the AUC of the increasing portion of the curve from estimated planting date to growing season maximum ('Plant' to 'Peak' in Figure 2), and the AUC of the declining portion of the curve from growing season maximum to estimated harvest date ('Peak' to 'Harvest' in Figure 2). For comparison, these values are calculate using two methods, the first using integration using smoothing splines (appendix formula 3 ) and second using trapezoidal estimation (appendix formula 4 ).

We develop a series of metrics to test if modeled yields could be improved through comparisons to 'ideal' years. This includes calculating the 95th percentile (based on sample quantile where the resulting quantile estimates are approximately median-unbiased regardless of the distribution of x (Hyndman & Fan, 1996)) of all NDVI values, of maximum values, and of the integral (area under the curve) of NDVI values. These use the built in functionality of R's base stat function show formula 5 ) in the appendix.

Additional functions were developed to extract the timing of particular phenomena, for instance the date of the maximum value of NDVI. Figure 2 visually demonstrates the ability of this function to estimate the timing on greeness onset (referred to henceforth as planting date), seasonal maximums, and harvest dates. For details on these calculations see formula 6 in the appendix. Multiway ties are handled by preferring the middle most date or if an even number of ties the left middle most date. Another calculates the average value of NDVI for each day of the year, which can be used for graphing anomalies (see formula 7 in the appendix). Finally, some of the above codes have improved performance when run on smoothed time series while removing outliers. For this procedure we use a function developed by Joshua Gray at North Carolina State University (see function 8 in the appendix).

### Annual Metrics

Basic annual summary statics including minimum, maximum, mean, and standard deviation can be calculated using function 9 in the appendix below. Alternatively most functions described above can be used to calculate annual vegetation metrics.

### Variable Definitions

*Table 1: Variable Names & Descriptions*

|  |  |
| --- | --- |
| Name | Description |
| Basic Summary Statistics |  |
| Rabi(Kharif)\_mean | Rabi(Kharif) mean values of NDVI |
| Rabi(Kharif)\_min | Rabi(Kharif) min values of NDVI |
| Rabi(Kharif)\_max | Rabi(Kharif) maximum values of NDVI |
| Rabi(Kharif)\_sd | Rabi(Kharif) standard deviation in values of NDVI |
| Rabi(Kharif)\_mean | Rabi(Kharif) mean values of NDVI |
| Rabi(Kharif)\_max\_date | Date of Rabi(Kharif) maximum values of NDVI |
| Rabi(Kharif)\_plant\_date | Estimated date of Rabi(Kharif) planting based on NDVI phenology |
| Rabi(Kharif)\_harvest\_date | Estimated date of Rabi(Kharif) harvest based on NDVI phenology |
| All\_95th\_prct | The estimated 95th percentile values of all seasons in the NDVI historical record |
| Integrated Summary Statistics |  |
| Rabi(Kharif)\_AUC | Rabi(Kharif) area under the curve of NDVI |
| Rabi(Kharif)\_AUC\_v2 | Rabi(Kharif) area under the curve of NDVI estimated by splines |
| Rabi(Kharif)\_AUC\_leading | Area under the curve of NDVI for the ascending part of the curve during the Rabi(Kharif) season |
| Rabi(Kharif)\_AUC\_trailing | Area under the curve of NDVI for the decreasing part of the curve during the Rabi(Kharif) season |
| Rabi\_season\_length | Difference between Rabi plant\_dates and harvest\_dates in days |
| Comparison to Norms |  |
| Rabi(Kharif)\_95th\_diff\_mn | Difference between Rabi(Kharif) mean NDVI and Rabi(Kharif) estimated 95th percentile of historical values |
| Rabi(Kharif)\_95th\_diff\_mx | Difference between Rabi(Kharif) max NDVI and Rabi(Kharif) estimated 95th percentile of historical max values |
| Rabi(Kharif)\_95th\_diff\_AUC | Difference between Rabi(Kharif) AUC NDVI and Rabi(Kharif) estimated 95th percentile of historical AUC values |
| Rabi(Kharif)\_AUC\_diff\_mn | Difference between the Rabi(Kharif) mean area under the curve and Rabi(Kharif) area under the curve for NDVI |

## Summarizing Space & Time: Extraction and Aggregation of Remotely Sensed Data

One major hurdle for this study was the rapid extraction of raster values bases on vector data while maintaining meaningful spatial and temporal components. In response, we developed the function *extract\_value\_point\_polygon* (see formula 10 in the appendix) to enhance the performance of the the default raster::extract() function. User processing times were better that 1/6000th that of extract() with the use of a 16-core Linux server. Additionally the function can take a list of adjacent raster stacks to perform data extraction, thereby properly handling vector datasets that span more than the extent of one raster.

## Spatial & Panel Regression Methods and Models

### Regression Methods and Diagnostic results

Traditional ordinary least squares (OLS) approaches look at cross-sectional or time-series data, exploiting variance in one dimension. However most social and physical processes occur over both space and time. For this reason we focus on the use of panel data sets which increases the amount of observed heterogeneity by including information about individuals *i* over time *t*. Critically, despite its widespread use in the field Geography (e.g. Michael L Mann et al. (2010)), cross-sectional analysis should generally *not* be used for forecasting, or for modeling of phenomenon with strong temporal components.

Compared to cross-sectional approaches, panel analysis substantially increases the degree of observed variance over both space and time. When pooled together, the integration of two statewide data sets provides (n = 14) over the 2003-2012 sample period (t = 10) provides (N=140) observations. Due to issues with estimating spatial panel models, the input data must be balanced (no missing observations). As such the number of districts for this regression is limited to 14 out of 36 total. This loss however is compensated by the fact that we can now compare a variety of estimation strategies on an even playing ground. Future work will aim to avoid this complication.

### Diagnostic Tests

#### Tests: Spatial Autocorrelation

To avoid overstating the statistical significance of regression coefficients, we test the residuals from a pooled linear for spatial autocorrelation (Fotheringham, Brunsdon, & Charlton, 2002, Luc Anselin, Bera, Florax, & Yoon (1996)). Here we use splm::slmtest, a locally robust Lagrange Multiplier test for spatial dependence for panel data, we reject the null of spatial independence in favor of spatial autocorrelation of the residual for a variety of spatial neighborhood definitions including queens continuity, and K-nearest-neighbors 6 through 10 (p<0.01), see Table 2 below.

*Table 2: Robust Lagrange Multiplier test for spatial dependence of pooled regression residuals*

|  |  |  |
| --- | --- | --- |
| Neighborhood | Test Stat | P-Value |
| Polygon Continuity | 41.04 | 1.489e-10 |
| KNN 6 | 53.38 | 2.747e-13 |
| KNN 7 | 56.44 | 5.79e-14 |
| KNN 8 | 52.38 | 4.572e-13 |
| KNN 9 | 43.06 | 5.302e-11 |
| KNN 10 | 44.69 | 2.303e-11 |

#### Tests: Pooled, Fixed or Random

Testing is required to choose the proper estimation method for panel regression. We must choose between pooled, fixed effect (FE), and random effect (RE) models. First, we can test for poolability of our model. Pooled regression assumes a constant intercept and slopes between different districts and time periods. We can test if variance across districts is equal to zero using an F-test. Here we reject the use of pooled OLS (p <= 1.62e-02) in favor of a fixed effects model with unique intercepts for each district. We can then compare the use of the fixed effects and random effects models. The hausman test checks for exogeneity of the unobserved error component, if the null hypothesis is rejected, the random effects model is inconsistent, and the fixed effects model will be preferred. If individual effects are exogenous both fixed and random effects are asymptotically equivalent. Here we test if , where are coefficient vectors of time-varying explanitory variables. We reject the null hypothesis (p <= 1.00e+00) and choose to use the fixed effects estimator as it will be the only consistent estimator.

#### Multicolinearity & Principal Components Analysis Transform

Multicolinearity, high correlations between independent variables, can increase estimates of a variable's estimated variance. This can have the adverse effect of creating models in which the is high and no variables are statistically significant. Multicolinearity can also produce coefficients of the "wrong sign" and of unreasonable magnitude (Greene, 1997; O’Brien, 2007). Here we use a variance inflation factor (VIF) to quantify how much the variance is inflated for each coefficient[[1]](#footnote-45). VIF values over four for any variable are generally considered problematic and require further examination (Greene, 1997). Here we present summary statistics for VIFs on an ordinary least squares estimation of all model variable in Table (3) below:

*Table 3: Variance inflation factor summary table*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean | SD | Min | Max |
| **VIF** | 77,773 | 179,885 | 1.361 | 634,503 |

To avoid problems with multicolinearity between independent variables we apply a principal components analysis transformation (PCA) to a centered and scaled matrix of all independent variables . PCA allows for the replacement of with a new matrix whos variables are orthogonal to each other but span the multidimensional space of (Geladi & Kowalski, 1986). In this case, we generate 36 principal components for inclusion in a random effects panel regression. A table of PCA component imporance can be found in the appendix, Table(A1).

### Pooled Regression Estimation

Panel data can be treated as "pooled", "fixed effects" or "random effects" models. Pooled OLS estimation assumes that intercepts and slopes are fixed between individual districts *i* with no specific treatment of time *t*. Although simple to understand, pooled OLS sometimes fails to properly control for determinants of spatial and temporal heterogeneity (e.g. districts with different policies, or changes to policies over time). Although the choice between pooled and fixed or random effect estimators, must be tested (see section '*Tests: Pooled, Fixed or Random*'). For comparison we estimate three types of panel models in this paper. The first of which, pooled OLS, is estimated in equation 1:

Where is a vector of our dependent variable, district yields in tons per hectare (*yield\_tn\_ha*) for each observation *i* which includes all districts across all years, is an intercept term, is a vector of *K* coefficients, corresponding to the *K* principal components, and is the residual. This estimation strategy essentially treats panel data as cross-sectional OLS.

### Panel Regression Estimation

We use panel data to model wheat yields over time at the district level. The use of panel data in this study helps to alleviate two key problems, unobserved spatial and temporal dynamics, and homogeneity (lack of variance). For a fixed effect panel, we estimate equation 2:

Where is a vector of our dependent variable, district yields in tons per hectare (*yield\_tn\_ha*) for each district *i* for each year *t*, are *i* intercept terms that control for unobserved characteristics of each district *i*, is a vector of *K* coefficients, corresponding to the *K* principal components[[2]](#footnote-48), is the between entity error term, and is the within entity error term.

### Spatial Panel Regression Estimation

Spatial autocorrelation is a special case of cross-sectional dependence caused by similarities of neighboring districts, and creates a situation whereby data can no longer be considered independently generated (L Anselin, 1999; P. J. Elhorst, 2010). The inclusion of a spatial lag (spillovers) can also increase predictive accuracy, as neighboring regions are often effected by similar exogenous shocks (for instance drought or rust) (ML L Mann & Warner, 2015; Michael L Mann & Warner, 2017; M L Mann et al., 2014). A spatial lag model can be considered a specification identifying the equilibrium outcome of spatial or social interaction processes, where the dependent variable for an individual is jointly determined with that of its neighbors (P. Elhorst, 2017).

For these reasons a spatially lagged fixed effect panel model is developed where spatial dependence is controlled for using a spatially weighted dependent variable, in the following form in 3:

Formula 3 augments the specification in 3. The primary difference is the inclusion of where is called the spatial autoregressive coefficient, is a row standardized weights matrix based for each individual *i* for its neighbors *j* on polygon continuity. These weighted values can be considered the mean values of 'neighboring' districts. then is then a measure of how neighboring values of affect a District *i*.

## Accuracy Measures

### Within-Sample

To measure the ability of these models to accurately model wheat yields we calculate the Root Mean Squared Error (RMSE) of the residuals, as outlined in 4:

Where are observed values of *yield\_tn\_ha* and are the predicted values of *yield\_tn\_ha* for time .

### Out-of-Sample Cross Validation

In predictive applications is it also extremely important to provide relevant measures of predictive performance. For this application we are interested in the ability of the model to estimate wheat yields based on a new growing season's data. To validate the predictive accuracy of our models we complete a Leave-P-Out Cross Validation (LPOCV), where each of the models 1-3 are reestimated, withholding a single year of observations, and storing the residuals for the omitted year. This process is repeated *p* times until all years have been evaluated. We then calculate the Root Mean Squared Error (RMSE) of the retained residuals, providing an estimate of the predictive accuracy of our models. RMSE is defined by 4.

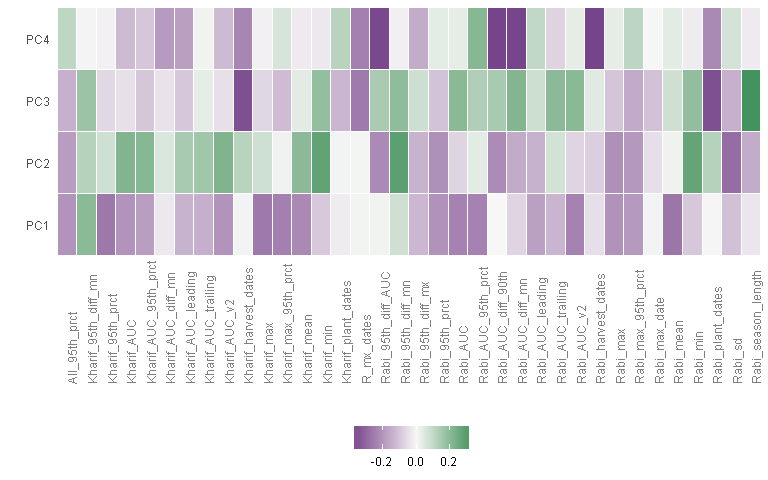
The following results section will outline the results from OLS, Panel, and Spatial Panel estimation (e.q. 1-3).

# Results

## Principal Component Loadings

We can evaluate how each variable contributes to the principal component used in later in this study. Large loadings indicate that a variable has a large effect on a principal component, this can be either in the positive or negative direction. We can see in Figure 3 below for principal component one (PC1), a large and positive loading for *Kharif\_95th\_diff\_mn*, which is a difference between mean NDVI for year *i* and the 95th percentile of historical values. This is relative to the large and positive loadings of a variety of variables largely relating to upper percentiles of the Kharif and Rabi seasons such as *Rabi\_max & Kharif\_max* and *Kharif\_95th\_prct*, a measure of the 95th percentile for that season. Combined this component emphasizes the variability between Kharif and Rabi growing seasons. The second principal component (PC2) is primarily positive loadings of the lower percentiles of the distribution for both growing seasons including *Kharif* and *Rabi\_min*. PC3 has strong loadings for variables relating to planting and harvest dates including *Kharif\_harvest\_dates*, *Rabi\_plant\_dates* and *Rabi\_season\_length*. PC4 emphasizes differences between the current year and the upper percentiles of the historic distribution for the wheat growing season. These include *Rabi\_95th\_diff\_AUC*, and *Rabi\_AUC\_diff\_mn*.

*Figure 3: Principal Component Analysis Loadings*



## Panel Regression

In this section we compare the results of three panel regression estimation techniques: pooled, fixed effects, and spatially lagged fixed effects. In particular, we are interested in the adjusted which controls for the loss of degrees of freedom, and the 'within' which evaluates the goodness of fit beyond what can be explained by fixed effects intercepts (or transform), (see "Assessing goodness of fit" in (Stata, 2016)). This 'within' is of particular interest because it is our best estimate of performance of the model in estimating year to year variations in crop yields. Since the objective of this study is to make predictions of wheat yields per hectare, it is also important to look at the accuracy of our estimation in these units. For this reason we also report the Root Mean Square Error (RMSE).

### Pooled OLS

Estimates of equation (1) are provide below:

Table 4: *District level pooled estimation of wheat yields in tons per hectare*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| **(Intercept)** | -199.5 | 26.66 | -7.484 | 8.598e-12 |
| **PC1** | -0.03061 | 0.01097 | -2.79 | 0.006046 |
| **PC2** | -0.1441 | 0.01194 | -12.06 | 3.841e-23 |
| **PC3** | -0.02384 | 0.017 | -1.402 | 0.1633 |
| **PC4** | 0.004502 | 0.02513 | 0.1792 | 0.8581 |
| **Years** | 0.1015 | 0.01328 | 7.642 | 3.658e-12 |

Residuals: Min 1st Qu. Median Mean 3rd Qu. Max -1.1 -0.264 -0.0318 -3.08e-19 0.312 0.998 RMSE: 0.402 on 134 degrees of freedom Multiple R2: 0.632, Adjusted R2: 0.618, Within R2:0.526 F-statistic: 45.9 on 5 and 100 DF

### Fixed Effects Panel

Estimates of equation (2) are provide below:

Table 5: *District level fixed effects estimation of wheat yields in tons per hectare*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t-value | Pr(>|t|) |
| **PC1** | -0.115 | 0.0533 | -2.17 | 0.0322 |
| **PC2** | 0.158 | 0.0319 | 4.97 | 2.24e-06 |
| **PC3** | 0.112 | 0.0273 | 4.11 | 7.3e-05 |
| **PC4** | -0.0634 | 0.0323 | -1.96 | 0.0517 |
| **Years** | 0.0774 | 0.0114 | 6.8 | 4.25e-10 |

Residuals: Min 1st Qu. Median Mean 3rd Qu. Max -0.54 -0.162 0.00707 -9.63e-18 0.174 0.74 RMSE: 0.256 on 121 degrees of freedom Multiple R2: 0.649, Adjusted R2: 0.597, Within R2: 0.808 F-statistic: 44.8 on 5 and 100 DF

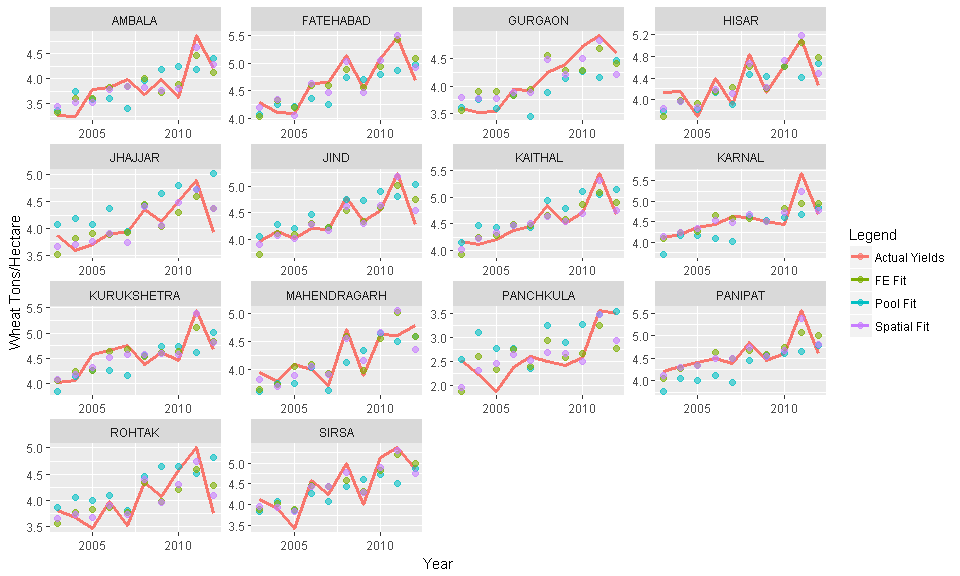
## Spatial Panel Regression

Estimates of equation (3) are provide below: Table 6: *District level spatial lag fixed effects estimation of wheat yields in tons per hectare*

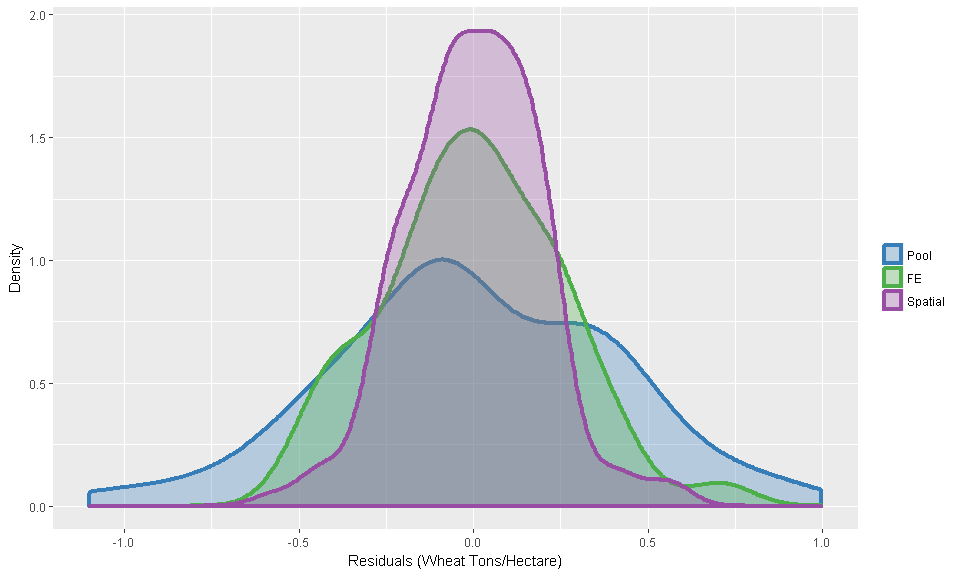
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t-value | Pr(>|t|) |
| **lambda** | 0.574 | 0.0581 | 9.88 | 4.94e-23 |
| **PC1** | -0.0617 | 0.0374 | -1.65 | 0.0993 |
| **PC2** | 0.0582 | 0.023 | 2.54 | 0.0112 |
| **PC3** | 0.0297 | 0.0194 | 1.53 | 0.126 |
| **PC4** | -0.0416 | 0.0229 | -1.82 | 0.0688 |
| **Years** | 0.0306 | 0.00939 | 3.26 | 0.00112 |

Residuals: Min 1st Qu. Median Mean 3rd Qu. Max -0.573 -0.112 0.00519 -5.08e-17 0.134 0.566 RMSE: 0.192 on 134 degrees of freedom Multiple R2: 0.916, Within R2: 0.803

*Figure 4: District Wheat Yield Estimates (2003-2012)*



*Figure 5: Residuals from District Wheat Yield Estimates (2003-2012)*



## Out-Of-Sample Cross Validation

Here we present the results of measurements of predictive accuracy outlined in the methods section. Within-sample refers to the RMSE associated with a single regression estimation, and out-of-sample refers to the Leave-P-Out Cross Validation (LPOCV) described in the *Methods* section, and the within- evaluates the goodness of fit beyond what can be explained by fixed effects intercepts .

*Table 7: LPOCV Performance Metrics for Within and Out Of Sample Prediction*

|  |  |  |  |
| --- | --- | --- | --- |
|  | Within Sample RMSE | Out of Sample RMSE | Within R2 |
| **Pool Fit** | 0.4 | 0.5 | 0.53 |
| **FE Fit** | 0.26 | 0.39 | 0.81 |
| **Spatial Fit** | 0.19 | 0.24 | 0.8 |

# Discussion

## Algorithms

A broad suite of algorithms for crop modeling using the open-source language R are provided in the *Functions* section of Appendix A as well as in an open github repository [here](https://github.com/mmann1123/India-Index-Insurance-Code), and were used to create variables described in Table 1. These functions have been designed to capture phenologically relevant components of the a growing season. They are designed to be flexibly applied to a wide set of geographic areas and crop seasons. For instance, similar or better results to those presented here were created for rice production in Punjab and Haryana. The functions are also able to reasonably estimate critical management indicators such as planting and harvest dates by mixing expert opinion with optimization routines.

Due to the idiosyncratic and heterogeneous nature of crop production in many developing countries these statistics are calculated independently for each area of interest and time-period. Therefore, metrics can accurately reflect spatial and temporal variability due to differences in management practices, climatic zone, weather, and edaphic properties. Importantly, the spatiotemporal characteristics of the data are maintained for use in panel econometric applications.

## Principal Component Loadings

Here we use principal components to avoid issues of multicolinearity across NDVI metrics for both the Rabi and Kharif seasons. An unusual feature of this study is the explicit integration of statistics from the preceeding Kharif season (typically rice) for Rabi wheat estimation. These variables were included because nutrients and water availability, although rarely modeled as such, are part of longer-term integrative processes. Water availability during a growing season is a function of rainfall, temperatures, wind speeds, and edaphic properties more closely related to metrics such as climatic water deficit (Michael L Mann et al., 2016). As such, the inclusion of Kharif season statistics allow us to control for spillovers from the preceding growing season.

Considering the longer-term nature of moisture availability, we also must consider the inability of Rabi NDVI to capture all meaningful characteristics of the season. Rainfall in the months leading up to a growing season will be important for maintaining soil moisture across the growing season, but is particularly important during the germination process, when the NDVI signal for the Rabi season is minimal. Plant health in the preceding season is therefore a good indicator of current growing conditions, especially in the early parts of the season. Looking at Figure 4 we can see that effects of a failed monsoon in 2004-2005 on yeilds. Despite the extensive use of irrigation in Punjab and Haryana, deficits in soil moisture proved difficult to make up for. For more details on this please see the later section on droughts.

## Regression Results

Preliminary tests rule out the use of pooled OLS in favor of fixed effects methods. Looking at Table 5 results for PC1, which largely comprises comparisons of the current years mean value and the historic 95th percentile in the Kharif season, we can see that low current values relative to the upper percentiles are significantly correlated with reductions in Rabi wheat yields (p<0.0322). For PC2, we see that increases in the minimum values of NDVI correspond to significantly higher wheat yields (p<2.24e-06). For PC3, we see that increases in season length correspond (although with a relatively small response) to significantly higher yields (p<7.3e-05). For PC4, we see significantly lower yields corresponding to greater differences between the current year AUC values and the upper percentiles of the observed record (p<0.0517). The spatial lag model from eq. 3 follows a similar pattern of sign and significance, however the relative size of the coefficients are smaller due to the influence of , the spatial autoregressive coefficient (Table 6). Here the large and positive coefficient indicates strong spatial spillover effects, with uncontrolled factors affecting yield not easily captured by NDVI (e.g. pest, disease, management, ground water availability) influencing their neighboring district.

## Spatiotemporal Model Accuracy

Metrics for wheat yield prediction at the district level are promising, with within- 0.53 to 0.81[[3]](#footnote-68) and out of sample RMSEs as low as 0.24 metric tons per hectare using only information extracted from NDVI (Table 7). The improvements in accuracy from leveraging the spatiotemporal properties of panel data is apparent with predictive accuracy increasing nearly 2 fold relative to traditional OLS approaches. We can also see the improvements of accuracy obtained when spatially lagged values of yields () are included in the regression. Note however that the inclusion of on the right side of the equation precludes true forecasting applications because it requires contemporaneous information on yields in neighboring districts. This particular issue could be at least partially addressed by integrated spatially and lagged values of Y for neighbors *j* (), which require no contemporaneous information.

These differences are clearly exhibited in Figure 4 where the fixed effect and spatial lag models clearly out-perform the pooled OLS methodology. This is particularly true during years exhibiting high variance such as 2005 and 2011 where unusually low and high values are prevalent across the sample. This characteristic of panel data reflects the underlying principal of the methodology which is able to model both the time-series properties of a single district, as well as the cross-sectional (spatial) properties of a particular year and or district. These improvements are also expressed in the distribution of residuals presented in Figure 5.

## Observations of Drought

Much of western India has experience regular and often intense droughts over the past decades. In Punjab prior to 2010, annual rainfall had declined nearly 450 mm, and experienced drought conditions in 1987, 1997, 2002, 2007 and 2009 (Sidhu & Vatta, 2012). These declines, coupled with increasing temperatures, has made both Punjab and Haryana increasingly dependent on irrigation to maintain yields, with over 95% of arable land irrigated (Sidhu & Vatta, 2012, Ambika, Wardlow, & Mishra (2016)). Declining precipitation and increasing temperatures has necessitated the ever increasing use of groundwater, calling into question the economic and environmental sustainability of the practice. For instance in 2009 alone, the water table in the central valley of Punjab declined by more than one meter (Sidhu & Vatta, 2012). Rapid declines in the water table substantially increases the costs of irrigation by necessitating the deepening of wells, and increasing the costs of pumping water to the surface. Because the government nearly fully subsidies the fuel and electricity costs of groundwater extraction, it seems broadly perceived that "yield levels are seldom affected by even a significant fall in rainfall" (Sidhu & Vatta, 2012).

Our time series of wheat yields shows strong resilience to drought but still demonstrates declines during periods of extended drought (Figure 4). For instance, yields declined up to 1/2 a ton per hectare during 2009. These declines, and other years like it, may be caused by the lag between observations of drought and investment in wells. Supporting this, a recent remotely sensed study of irrigated land reports significant portions both states coming into and out of irrigated production during times of drought (Ambika et al., 2016). Regardless of access to irrigation water, proper timing of irrigation and management can be difficult during periods of drought especially in areas with sandy soils or even those with compacted soils (Utt, 2017, Kumar (2017)).

From Figure 4 we can see that the variables derived from NDVI as estimated by fixed effects and spatial panel can realistically track these shifts. Meanwhile the pooled regression approach, which ignores the spatiotemporal nature of the data, is largely insensitive the shifts in NDVI caused by drought. This finding reinforces the critical importance of panel regression to predictive modeling, especially in studies with spatiotemporal components.

# Conclusions

In this study we demonstrate and provide open access to a new suite of algorithms designed to rapidly extract and summarize remotely sensed data for use in panel regression. We then demonstrate the desireability of panel regression techniques above traditional crossectional or pooled approaches. Traditional cross sectional and pooled regression approaches mistreat or underutilize the spatiotemporal variation they wish to explain. All code use to estimate these models and to write this text is provided [online](https://github.com/mmann1123/India-Index-Insurance-Code/blob/master/WriteUp/WheatYieldWriteUp.Rmd) through github.

##### 

# Appendix A

## Yield Data

*Table A2: Rabi Season Wheat Yields Metric Tons per Hectare by State*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| State | district | Min | Mean | Max |
| HARYANA | AMBALA | 3.24 | 3.82 | 4.86 |
| HARYANA | BHIWANI | 3.28 | 3.78 | 4.47 |
| HARYANA | FARIDABAD | 3.66 | 4.22 | 4.92 |
| HARYANA | FATEHABAD | 4.08 | 4.62 | 5.46 |
| HARYANA | GURGAON | 3.52 | 4.09 | 4.92 |
| HARYANA | HISAR | 3.7 | 4.3 | 5.1 |
| HARYANA | JHAJJAR | 3.59 | 4.07 | 4.89 |
| HARYANA | JIND | 3.96 | 4.35 | 5.24 |
| HARYANA | KAITHAL | 4.11 | 4.53 | 5.45 |
| HARYANA | KARNAL | 4.13 | 4.55 | 5.68 |
| HARYANA | KURUKSHETRA | 4.03 | 4.54 | 5.44 |
| HARYANA | MAHENDRAGARH | 3.73 | 4.2 | 4.78 |
| HARYANA | MEWAT | 3.07 | 3.78 | 4.37 |
| HARYANA | PALWAL | 4.15 | 4.69 | 5.09 |
| HARYANA | PANCHKULA | 1.88 | 2.63 | 3.57 |
| HARYANA | PANIPAT | 4.22 | 4.57 | 5.56 |
| HARYANA | REWARI | 4.01 | 4.51 | 4.97 |
| HARYANA | ROHTAK | 3.45 | 4.01 | 5 |
| HARYANA | SIRSA | 3.43 | 4.42 | 5.36 |
| HARYANA | SONIPAT | 4.28 | 4.63 | 5.51 |
| HARYANA | YAMUNANAGAR | 3.39 | 4.01 | 5.38 |
| PUNJAB | AMRITSAR | 4.05 | 4.27 | 4.43 |
| PUNJAB | BATHINDA | 3.89 | 4.28 | 4.79 |
| PUNJAB | FARIDKOT | 4.04 | 4.35 | 4.81 |
| PUNJAB | FATEHGARH SAHIB | 4.09 | 4.67 | 5.18 |
| PUNJAB | FIROZPUR | 3.98 | 4.33 | 4.92 |
| PUNJAB | GURDASPUR | 3.57 | 4.06 | 4.48 |
| PUNJAB | HOSHIARPUR | 3.4 | 3.71 | 4.29 |
| PUNJAB | JALANDHAR | 4.14 | 4.38 | 4.69 |
| PUNJAB | KAPURTHALA | 3.97 | 4.28 | 4.64 |
| PUNJAB | LUDHIANA | 4.39 | 4.7 | 4.96 |
| PUNJAB | MANSA | 3.75 | 4.29 | 4.88 |
| PUNJAB | MOGA | 4.14 | 4.52 | 5.01 |
| PUNJAB | MUKTSAR | 3.94 | 4.45 | 4.98 |
| PUNJAB | PATIALA | 4.12 | 4.55 | 4.83 |
| PUNJAB | RUPNAGAR | 3.31 | 3.81 | 4.51 |
| PUNJAB | SANGRUR | 4.23 | 4.57 | 5.13 |

## Principal Components Analysis

*Table A1: Importance of PCA components*

Table continues below

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| **Standard deviation** | 3.47 | 2.83 | 2.28 | 1.56 | 1.27 | 1.23 | 1.13 |
| **Proportion of Variance** | 0.335 | 0.222 | 0.144 | 0.0678 | 0.0451 | 0.0418 | 0.0353 |
| **Cumulative Proportion** | 0.335 | 0.557 | 0.701 | 0.769 | 0.814 | 0.855 | 0.891 |

Table continues below

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | PC8 | PC9 | PC10 | PC11 | PC12 | PC13 | PC14 |
| **Standard deviation** | 1.05 | 0.889 | 0.762 | 0.739 | 0.577 | 0.416 | 0.367 |
| **Proportion of Variance** | 0.0305 | 0.022 | 0.0161 | 0.0152 | 0.00925 | 0.00481 | 0.00375 |
| **Cumulative Proportion** | 0.921 | 0.943 | 0.959 | 0.974 | 0.984 | 0.988 | 0.992 |

Table continues below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PC15 | PC16 | PC17 | PC18 | PC19 | PC20 |
| **Standard deviation** | 0.275 | 0.218 | 0.199 | 0.177 | 0.158 | 0.136 |
| **Proportion of Variance** | 0.00211 | 0.00132 | 0.0011 | 0.00087 | 0.00069 | 0.00052 |
| **Cumulative Proportion** | 0.994 | 0.996 | 0.997 | 0.998 | 0.998 | 0.999 |

Table continues below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PC21 | PC22 | PC23 | PC24 | PC25 | PC26 |
| **Standard deviation** | 0.123 | 0.109 | 0.066 | 0.0618 | 0.0529 | 0.0493 |
| **Proportion of Variance** | 0.00042 | 0.00033 | 0.00012 | 0.00011 | 8e-05 | 7e-05 |
| **Cumulative Proportion** | 0.999 | 1 | 1 | 1 | 1 | 1 |

Table continues below

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | PC27 | PC28 | PC29 | PC30 | PC31 | PC32 |
| **Standard deviation** | 0.0359 | 0.0227 | 0.00273 | 0.0018 | 0.00119 | 0.000975 |
| **Proportion of Variance** | 4e-05 | 1e-05 | 0 | 0 | 0 | 0 |
| **Cumulative Proportion** | 1 | 1 | 1 | 1 | 1 | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PC33 | PC34 | PC35 | PC36 |
| **Standard deviation** | 1e-15 | 6.9e-16 | 3e-16 | 2.85e-16 |
| **Proportion of Variance** | 0 | 0 | 0 | 0 |
| **Cumulative Proportion** | 1 | 1 | 1 | 1 |

## Functions

All functions used in the creation of this paper above are provided in the section below. Additionally, all [algorithms](https://github.com/mmann1123/India-Index-Insurance-Code/blob/master/SummaryFunctions.R) as well as [the code](https://github.com/mmann1123/India-Index-Insurance-Code/blob/master/WriteUp/WheatYieldWriteUp.Rmd) used to generate the findings from this study are provided through a github repository [here](https://github.com/mmann1123/India-Index-Insurance-Code).

*Functions 1: Planting/Harvest date functions*

PlantHarvestDates = function(start\_date, end\_date, PlantingMonth, PlantingDay,   
 HarvestMonth, HarvestDay) {  
 # this function takes in date range and returns planting and harvest date  
 # for time series as a data.frame for all years of interest. Handles growing  
 # periods overlaping a new year properly. NOTE: This is used to create  
 # dataframe of planting / harvest dates for many other functions e.g.  
 # PlantHarvest =  
 # PlantHarvestDates('2002-01-01','2016-02-02',PlantingMonth=11,  
 # PlantingDay=23,HarvestMonth=4,HarvestDay=30)  
   
 start\_end\_years = c(strptime(start\_date, "%Y-%m-%d"), strptime(end\_date,   
 "%Y-%m-%d"))  
 names(unclass(start\_end\_years[1]))  
 start\_end\_years[1]$mon = PlantingMonth - 1  
 start\_end\_years[1]$mday = PlantingDay  
 planting = as.Date(seq(start\_end\_years[1], length = strptime(dates[2], "%Y-%m-%d")$year -   
 strptime(dates[1], "%Y-%m-%d")$year, by = "year"))  
 # set harvest  
 start\_end\_years[2]$year = start\_end\_years[1]$year + 1 # set year equal to start year +1  
 start\_end\_years[2]$mon = HarvestMonth - 1  
 start\_end\_years[2]$mday = HarvestDay  
 harvest = as.Date(seq(start\_end\_years[2], length = strptime(end\_date, "%Y-%m-%d")$year -   
 strptime(start\_date, "%Y-%m-%d")$year, by = "year"))  
 return(data.frame(planting = planting, harvest = harvest))  
}  
  
SearchMinumumBeforeAfterDOY = function(x, dates\_in, DOY\_in, days\_shift, dir) {  
 # calculates the global minimum for days before,after,both of expected  
 # planting date best to set DOY as the last expected date of planting x =  
 # vegetation index, dates\_in = dates of observation POSIX, DOY\_in = expected  
 # planting or harvest date days\_shift = # days to search around DOY\_in,  
 # dir='before' 'after' 'beforeafter'  
   
 if (days\_shift <= 8) {  
 print("Using less than 8 days is dangerous, 15-30 stable")  
 }  
   
 # avoid problems with time class  
 if (is.na(DOY\_in[1])) {  
 print("ERROR: convert date format to %Y%j")  
 break  
 }  
 if (class(dates\_in)[1] != "POSIXlt") {  
 dates\_in = as.POSIXlt(dates\_in)  
 }  
   
 # limit to fixed # of days before/after DOY  
 DOY\_in = as.POSIXlt(DOY\_in)  
 DOY\_before = DOY\_in  
   
 # names(unclass(DOY\_before[1]))  
 if (dir == "before")   
 DOY\_before$mday = DOY\_before$mday - days\_shift # set days before to doy - days\_before  
 if (dir == "after")   
 DOY\_before$mday = DOY\_before$mday + days\_shift # set days before to doy - days\_before  
 if (dir == "beforeafter") {  
 DOY\_before$mday = DOY\_before$mday - days\_shift  
 DOY\_in$mday = DOY\_in$mday + days\_shift  
 }  
 DOY\_table = data.frame(DOY\_before = DOY\_before, DOY\_in = DOY\_in) #join start end search dates  
   
 # list all days 'days\_before' DOY\_in  
 if (dir == "before" | dir == "beforeafter") {  
 DOY\_interest = as.POSIXlt(unlist(lapply(1:dim(DOY\_table)[1], function(h) {  
 format(seq(DOY\_table[h, 1], DOY\_table[h, 2], by = "day"), "%Y-%m-%d")  
 })), tz = "UTC")  
 }  
 if (dir == "after") {  
 DOY\_interest = as.POSIXlt(unlist(lapply(1:dim(DOY\_table)[1], function(h) {  
 format(seq(DOY\_table[h, 2], DOY\_table[h, 1], by = "day"), "%Y-%m-%d")  
 })), tz = "UTC")  
 }  
   
 # find all local minima, and match with DOY  
 x\_DOY\_interest = x[dates\_in %in% DOY\_interest]  
 dates\_DOY\_interest = dates\_in[dates\_in %in% DOY\_interest]  
 # get min value for this period for each year  
 sort(AnnualMaxima(x\_DOY\_interest \* -1, as.Date(dates\_DOY\_interest)))  
}

*Functions 2: Flexible growing season vegetation metrics*

PeriodAggregator = function(x,dates\_in,date\_range\_st, date\_range\_end,by\_in='days',FUN){  
 # returns a summary statistic of x for any function FUN, over the period defined by date\_range\_st, date\_range\_end  
 # x = vegetation index data, dates\_in = dates of observation POSIX, dates\_in,date\_range\_st = start end dates of period, FUN = function  
 # E.g. PeriodAggregator(x=plotdatasmoothed$EVI,dates\_in = plotdatasmoothed$dates,date\_range\_st=plotdatasmoothed$dates[1],date\_range\_end=plotdatasmoothed$dates[20], FUN = function(y){mean(y,na.rm=T)})  
 if(class(dates\_in)[1]== "POSIXct"|class(dates\_in)[1]== "POSIXlt" )dates\_in = as.Date(dates\_in)  
 if(class(date\_range\_st)[1]== "POSIXct" ){date\_range\_st = as.Date(date\_range\_st)  
 date\_range\_end = as.Date(date\_range\_end)}  
 #Avoid problems with missing plant or harvest dates  
 if(length(date\_range\_st)!=length(date\_range\_end)){print('number of elements in start end dates dont match'); break}  
 dataout=lapply(1:length(date\_range\_st),function(z){  
 DateRange = seq(date\_range\_st[z],date\_range\_end[z],by=by\_in)  
 x=x[dates\_in %in% DateRange]  
 dates\_in=dates\_in[dates\_in %in% DateRange]  
 FUN(x)})  
 dataout = do.call(c,dataout)  
 names(dataout)=format(date\_range\_st,'%Y')  
 dataout  
 }

*Functions 3: Area under the curve estimation - smoothing splines*

PeriodAUC = function(x\_in,dates\_in,DOY\_start\_in,DOY\_end\_in){  
 # calculate area under the curve by period of the year using spline estimation  
 # x = data, dates\_in=asDate(dates),DOY\_start\_in=asDate(list of start periods),DOY\_end\_in=asDate(list of end per  
 # x = plotdatasmoothed$EVI,dates\_in = plotdatasmoothed$dates , DOY\_start\_in= plant\_dates ,DOY\_end\_in=harvest\_dates)  
 if(class(dates\_in)[1]== "POSIXct"|class(dates\_in)[1]== "POSIXlt" )dates\_in = as.Date(dates\_in)  
 dates\_group = rep(0,length(dates\_in)) # create storage for factors of periods  
 # get sequences of periods of inerest  
 seq\_interest = lapply(1:length(DOY\_start\_in),function(z){seq(DOY\_start\_in[z],DOY\_end\_in[z],by='days')})  
 # switch dates-group to period group  
 years\_avail = sort(as.numeric(unique(unlist(  
 lapply(seq\_interest,function(z) format(z,'%Y'))))))  
 for(z in 1:length(seq\_interest)){ #assigns year for beginging of planting season  
 dates\_group[dates\_in %in% seq\_interest[[z]]]=years\_avail[z]  
 assign('dates\_group',dates\_group,envir = .GlobalEnv) } # assign doesn't work in lapply using for loop instead  
 # calculate AUC for periods of interest  
 FUN = function(q,w){auc(q,w,type='spline')}  
 datesY = format(dates\_in,'%Y')  
 data.split = split(x\_in,dates\_group)  
 d = do.call(c,lapply(2:length(data.split),function(z){ # start at 2 to avoid group=0  
 FUN(q=1:length(data.split[[z]]),w=data.split[[z]]) }))  
 names(d) = names(data.split)[2:length(data.split)]  
 d  
 }

*Functions 4: Area under the curve estimation - trapazoidal estimation*

PeriodAUC\_method2 = function(x\_in,dates\_in,DOY\_start\_in,DOY\_end\_in){  
 #NOTE SPLINE METHOD 1 SEEMS to WORK BETTER  
 # calculate area under the curve by period of the year  
 # x = data, dates\_in=asDate(dates),DOY\_start=asDate(list of start periods),DOY\_end=asDate(list of end per$  
 # x = plotdatasmoothed$EVI,dates\_in = plotdatasmoothed$dates , DOY\_start=annualMinumumBeforeDOY(x = plotd$  
 if(class(dates\_in)[1]== "POSIXct"|class(dates\_in)[1]== "POSIXlt" )dates\_in = as.Date(dates\_in)  
  
 dates\_group = rep(0,length(dates\_in)) # create storage for factors of periods  
 # get sequences of periods of inerest  
 seq\_interest = lapply(1:length(DOY\_start\_in),function(z){seq(DOY\_start\_in[z],DOY\_end\_in[z],by='days')})  
 # switch dates-group to period group  
 years\_avail = sort(as.numeric(unique(unlist(  
 lapply(seq\_interest,function(z) format(z,'%Y'))))))  
 for(z in 1:length(seq\_interest)){ #assigns year for beginging of planting season  
 dates\_group[dates\_in %in% seq\_interest[[z]]]=years\_avail[z]  
 assign('dates\_group',dates\_group,envir = .GlobalEnv) } # assign doesn't work in lapply using for loop instead  
   
 # calculate AUC for periods of interest  
 FUN = function(q,w){ sum(diff(q)\*rollmean(w,2))}  
 datesY = format(dates\_in,'%Y')  
 data.split = split(x\_in,dates\_group)  
 d = do.call(c,lapply(2:length(data.split),function(z){ # start at 2 to avoid group=0  
 FUN(q=1:length(data.split[[z]]),w=data.split[[z]]) }))  
 names(d) = names(data.split)[2:length(data.split)]  
 #print(cbind(names(data.split)[2:length(data.split)], d))  
 d  
 }

*Functions 5: Base function used for estimating sample quantiles*

quantile\_type8 = function(x){  
 quantile(x ,p=Quant\_percentile,type=8,na.rm=T)  
}

*Functions 6: Function to return date of any given phenomenon*

PeriodAggregatorDates = function(x,dates\_in,date\_range\_st, date\_range\_end,by\_in='days',FUN){  
 # returns a date of summary statistic defined by FUN  
 # like the date of the maximum value of x for the period defined by date\_range\_st, date\_range\_end  
 # other parameters identical to other functions show above  
 if(class(dates\_in)[1]== "POSIXct"|class(dates\_in)[1]== "POSIXlt" )dates\_in = as.Date(dates\_in)  
 if(class(date\_range\_st)[1]== "POSIXct" ){date\_range\_st = as.Date(date\_range\_st)  
 date\_range\_end = as.Date(date\_range\_end)}  
 #Avoid problems with missing plant or harvest dates  
 if(length(date\_range\_st)!=length(date\_range\_end)){print('number of elements in start end dates dont match');break}  
  
 dataout=lapply(1:length(date\_range\_st),function(z){  
 DateRange2 = seq(date\_range\_st[z],date\_range\_end[z],by=by\_in)  
 x2 = x[dates\_in %in% DateRange2]  
 dates\_in2 = dates\_in[dates\_in %in% DateRange2]  
 which\_max = which(FUN(x2) == x2)  
 if(length(which\_max)>1){  
 which\_max = c(which\_max[1],which\_max[length(which\_max)]) # limit to only 2   
 if((which\_max[2]-which\_max[1])==1){  
 which\_max=which\_max[1] # favor the first instance of maximum  
 } else if((which\_max[2]-which\_max[1])==2){  
 which\_max=which\_max[1]+1 # is seperated by 2 choose middle left  
 } else if((which\_max[2]-which\_max[1])==3){  
 which\_max=which\_max[1]+2} # is seperated by 3 choose middle  
 }  
 max\_dates = dates\_in2[which\_max]  
 })  
 dataout = do.call(c,dataout)  
 names(dataout)=format(date\_range\_st,'%Y')  
 dataout  
 }

*Functions 7: Mean day of the year values*

AnnualAverageDOYvalues = function(x,dates\_in){  
 # calculates the average value for DOY for the whole series  
 datesj = format(dates\_in,'%j')  
 do.call(c,lapply(split(x,datesj),function(y){mean(y,na.rm=T)}))}

*Functions 8: Smoothing splines with outlier removal*

#---------------------------------------------------------------------  
# This function takes a time series w/ dates (x, dates) and returns a spline smoothed time series with outliers removed.  
# Outliers are identified as points with absolute value more than out\_sigma \* sd, where sd is the residual  
# standard deviation between the input data and the initial spline fit, and out\_sigma is a variable  
# coefficient. The spline smoothing parameter spline\_spar controls the smoothness of the fit (see spline.smooth help)  
# and out\_iterations controls the number of times that outliers are checked and removed w/ subsequent spline refit  
# pred\_dates is a vector of dates where spline smoothed predictions of x are desired. If NA, then a daily series spanning  
# min(dates)-max(dates) is returned  
SplineAndOutlierRemoval <- function(x, dates, out\_sigma=3, spline\_spar=0.3, out\_iterations=1,pred\_dates){  
 dates <- as.numeric(dates) # spline doesn't work with dates  
 pred\_dates = as.numeric(pred\_dates)  
 # if prediction dates aren't provided, we assume we want daily ones  
 if(is.na(pred\_dates[1])){  
 pred\_dates <- min(dates, na.rm=T):max(dates, na.rm=T)}  
 # eliminate outliers and respline  
 for(i in 1:out\_iterations){  
 # fit a smoothing spline to non-missing data  
 spl <- try(smooth.spline(dates[!is.na(x)], x[!is.na(x)], spar=spline\_spar), silent=T)  
 if(inherits(spl, 'try-error')){  
 print("Failed to fit smoothing spline")  
 return(NA)  
 }  
 smooth\_x <- try(predict(spl, dates)$y, silent=T) # calculate spline smoothed values  
 if(inherits(smooth\_x, 'try-error')){  
 print("Failed to predict with spline")  
 return(NA)  
 }  
 smooth\_x\_resid <- x - smooth\_x # calculate residuals from spline  
 smooth\_x\_resid\_sd <- try(sd(smooth\_x\_resid, na.rm=T), silent=T) # standard dev of absolute value of residuals  
 if(inherits(smooth\_x\_resid\_sd, 'try-error')){  
 print("Failed to get sd of residuals")  
 return(NA)  
 }  
 outliers <- abs(smooth\_x\_resid) > out\_sigma \* smooth\_x\_resid\_sd  
 outliers[is.na(outliers)] <- F  
 if(sum(outliers) > 0){  
 # if we found outliers, eliminate them in x and refit up to iterations  
 x[outliers] <- NA  
 }else{  
 # if we didn't find any outliers, we abandon the iteration and return the smoothed values  
 smooth\_x\_return <- try(predict(spl, pred\_dates)$y, silent=T)  
 if(inherits(smooth\_x\_return, 'try-error')){  
 print("No outliers, but failed to predict with final spline")  
 return(NA)  
 }else{  
 return(smooth\_x\_return)  
 }  
 }  
 }  
 # fit the spline to the outlier screened data, then return the predicted series  
 spl <- try(smooth.spline(dates[!is.na(x)], x[!is.na(x)], spar=spline\_spar), silent=T)  
 if(inherits(spl, 'try-error')){  
 print("Failed to predict with final spline")  
 return(NA)  
 }else{  
 smooth\_x\_return <- try(predict(spl, pred\_dates)$y, silent=T)  
 if(inherits(smooth\_x\_return, 'try-error')){  
 return(NA)  
 }else{  
 return(smooth\_x\_return)  
 }  
 }  
}

*Function 9: Flexible annual vegetation metrics*

AnnualAggregator = function(x,dates\_in,FUN){  
 # returns an annual summary statistic of any function  
 # x = vegetation index data, dates\_in = dates of observation POSIX,  
 # E.g. AnnualAggregator(x= plotdatasmoothed$EVI,dates\_in = plotdatasmoothed$dates, FUN = function(y){mean(y,na.rm=T)})  
 datesY = format(dates\_in,'%Y')  
 do.call(c,lapply(split(x,datesY),FUN))}

*Function 10: Rapid multicore extract raster data by point or polygon*

extract\_value\_point\_polygon = function(point\_or\_polygon, raster\_stack, num\_workers){  
 # Returns list containing values from locations of spatial points or polygons  
 # if polygons are too small reverts to centroid   
 if(class(raster\_stack)!='list'){raster\_stack=list(raster\_stack)}  
 lapply(c('raster','foreach','doParallel'), require, character.only = T)  
 registerDoParallel(num\_workers)  
 ptm <- proc.time()  
 # iterate between points or polygons  
 ply\_result = foreach(j = 1:length(point\_or\_polygon),.inorder=T) %do%{  
 print(paste('Working on feature: ',j,' out of ',length(point\_or\_polygon)))  
 get\_class= class(point\_or\_polygon)[1]  
 # switch rasterstack according to which point or polygon is %over%  
 for(z in 1:length(raster\_stack)){  
 # set raster to use  
 raster\_stack\_use = raster\_stack[[z]]  
 # get cell numbers of point of polygon, repeat if missing  
 if(get\_class=='SpatialPolygons'|get\_class=='SpatialPolygonsDataFrame'){  
 cell = as.numeric(cellFromPolygon(raster\_stack\_use, point\_or\_polygon[j,], weights=F)[[1]])  
 # if polygon is too small to find cells, convert to centroid and get cellfromXY  
 if(length(cell)==0){ #coord(poly) returns centroid  
 cell = as.numeric(na.omit(cellFromXY(raster\_stack\_use, coordinates(point\_or\_polygon[j,]) )))}}  
 if(get\_class=='SpatialPointsDataFrame'|get\_class=='SpatialPoints'){  
 cell = as.numeric(na.omit(cellFromXY(raster\_stack\_use, point\_or\_polygon[j,])))}  
 # if cells found keep raster\_stack\_use = raster\_stack[[z]]  
 if(length(cell)!=0){break}  
 # if cells not found repeat for different stack or return NA  
 if(length(cell)==0 & z!=length(raster\_stack)){next}else{return(NA)}  
 }  
 # create raster mask from cell numbers  
 r = rasterFromCells(raster\_stack\_use, cell,values=F)  
 result = foreach(i = 1:dim(raster\_stack\_use)[3],.packages='raster',.inorder=T) %dopar% {  
 crop(raster\_stack\_use[[i]],r)  
 }  
 result=as.data.frame(getValues(stack(result)))  
 return(result)  
 }  
 print( proc.time() - ptm)  
 endCluster()  
 return(ply\_result)  
 }

# References

*Figure 6: Fitted vs actual for estimation of equation (5)*

(Sidhu & Vatta, 2012)

The growth in productivity and production was ushered by the trio of high yielding seeds, irrigation, and fertilizers supported by farm mechanization and institutional and infrastructural development. Currently, 98 per cent of the cultivated area in Punjab is under assured irrigation.

The yield levels are seldom affected by even a significant fall in rainfall due to larger dependence on ground water which mitigates the negative impacts of any such fall.

The impact of rise in temperature has started appearing in wheat productivity, which fell from 2001 to 2005 continuously

The declining water table in the Punjab has raised serious doubts about the sustainability of the rice-wheat crop rotation, which has been manifested in terms of stagnating productivity in the state.

However, during the decade of 2000, average annual rainfall has declined less than 450 mm during 1998-2005. Not only has the precipitation fallen, it has also become irregular with uneven spread over space and time. The frequency of drought has gone up; we had drought years in 1987, 1997, 2002, 2007 and 2009, which clearly demonstrates that the frequency of droughts has gone up lately. It has however been observed that the droughts do not have any major direct impact on crop production and productivity in the state. It is usual while the production in the rest of the country declines during the drought years, the production in Punjab increases owing to assured access to the ground water.

Record production of 8.8 million tones of paddy during 2002 and 9.1 million tones in the last year strengthens this argument. The respective production of wheat during these years was also 15.5 and 15.6 million tones. However, the droughts have significant social and economic adverse impacts on the Punjab economy. The drought conditions necessitate increased use of diesel to pump out the ground water which pushes up the cost of cultivation of paddy. Consequently, the expenditure on irrigation increases.

Due to deficient rainfall during the months of July, August and September 2009, additional power estimated to be worth Rs 450 crore and additional diesel oil amounting to Rs 300 crore were used to lift groundwater for saving paddy crops. Deficient rainfall further engenders the sustainability of ground water resources which is already being over-exploited to maintain the current levels of rice production for the food security of the nation. In the drought of 2009, the water table went down more than one metre in central districts f the state (compared to around 70 cms in the previous year during the same months). The droughts, therefore, necessitate the increased investments in irrigation.

Zone I (Sub-mountainous region), known as the Kandi region, has undulating topography comprising about 17 percent of the total area of the state. This zone has particularly abundant rainfall that averages more than 1,100 mm per annum. As the water table is deep and the soil is rocky, the sinking of tube wells and pumping out of water is very costly. Because of heterogeneity in agro-climatic conditions, the cropping pattern of the area is more diverse relative to the other zones and comprises crops like wheat, rice, basmati rice, maize, oilseeds, fruits, and vegetables.

central region, is also known as the "sweet water" region and comprises about 47 percent of the area of the state. Average annual rainfall in this region is about 760 mm. It is highly productive and has a tight-knit system of irrigation, mainly through the use of tube-wells. The main cropping system in Zone II is the rice-wheat rotation. The water table in this region has been falling at an alarming average rate of 0.94 meters per year during 2004-07. The steep fall in groundwater is due to the massive increase in the number of tube wells from 192,000 in 1970-71 to 1,276,200 in 2008-09 that has inexorably been fuelled by power subsidies. The falling water table and declining soil fertility threaten the sustainability of the production environment of this region.

Zone III, the south-western region and popularly known as the cotton belt, comprises almost 36 percent of the cultivated area of the state. This region is endowed with deep and brackish groundwater and sandy soil. It is much dryer than the other two zones. The average annual rainfall of this region is 360 mm. Over the last decade, there has been a fall in the area under cotton that is attributed to a decline in its productivity. At the same time, the increase in area under rice has increased salt accumulation on the soil surface due to the continuous use of underground water which is brackish and has led to water logging of the soil.

ambika2016remotely LAI rarely approaches such a value for crop and this would be no issue for irrigation classification therefore NDVI alone can be used to differentiate irrigated from non-irrigated areas as well as crop types.

smoothing techniques typically result in only subtle increases in overall classification accuracy and some of these techniques lead to large inconsistencies in previous classification efforts2,25

while 95% of arable land in the state of Punjab is irrigated.

Additionally, states like Tamil Nadu, Punjab, Rajasthan, Maharashtra, West Bengal and Haryana have already developed over 70% of its major and moderate irrigation potential.

Irrigation is estimated to use about 70% of world's total available freshwater for food production using 18% of cultivated area globally1,2.

precipitation3. An increase in winter temperature, erratic monsoon season rainfall, extensive use of ground water resources, and absence of effective adaptation strategies are likely to negatively affect crop productivity4,5. In the future, the Indian region may experience stress to meet its water demand due to extreme weather and climate events such as droughts and heat waves, specifically in arid and semi-arid regions, where groundwater extraction is prominent for irrigation3-5.

Declines in irrigated area were fairly constant for some years with values ranging between ???1.46 and ???2.60 mha, however, other years 2003-2004, 2005-2006 and 2012-2013 showed a larger decrease in irrigated area with values ranging from ???6.21 to ???13.38 mha. One reason for the latter case is that the preceding year had a pronounced rainfall deficit as illustrated in Supplementary Fig. 3c.

Terrestrial and ground water storages withdrawal were also less during 2012 drought year (may be because of less sown area), which could be a possible reason to decrease irrigation for this time period.

Ambika, A. K., Wardlow, B., & Mishra, V. (2016). Remotely sensed high resolution irrigated area mapping in india for 2000 to 2015. *Scientific Data*, *3*.

Anselin, L. (1999). Spatial Econometrics. Richardson TX: School of Social Sciences- University of Texas Dallas.

Anselin, L., Bera, A. K., Florax, R., & Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, *26*(1), 77–104.

Asrar, G., Fuchs, M., Kanemasu, E., & Hatfield, J. (1984). Estimating absorbed photosynthetic radiation and leaf area index from spectral reflectance in wheat. *Agronomy Journal*, *76*(2), 300–306.

Bartholome, E. (1988). Radiometric measurements and crop yield forecasting some observations over millet and sorghum experimental plots in mali. *International Journal of Remote Sensing*, *9*(10-11), 1539–1552.

Becker-Reshef, I., Justice, C., Sullivan, M., Vermote, E., Tucker, C., Anyamba, A., … others. (2010a). Monitoring global croplands with coarse resolution earth observations: The global agriculture monitoring (glam) project. *Remote Sensing*, *2*(6), 1589–1609.

Becker-Reshef, I., Vermote, E., Lindeman, M., & Justice, C. (2010b). A generalized regression-based model for forecasting winter wheat yields in kansas and ukraine using modis data. *Remote Sensing of Environment*, *114*(6), 1312–1323.

Benedetti, R., & Rossini, P. (1993). On the use of ndvi profiles as a tool for agricultural statistics: The case study of wheat yield estimate and forecast in emilia romagna. *Remote Sensing of Environment*, *45*(3), 311–326.

Biradar, C. M., & Xiao, X. (2011). Quantifying the area and spatial distribution of double-and triple-cropping croplands in india with multi-temporal modis imagery in 2005. *International Journal of Remote Sensing*, *32*(2), 367–386.

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.

Butler, E. E., & Huybers, P. (2015). Variations in the sensitivity of us maize yield to extreme temperatures by region and growth phase. *Environmental Research Letters*, *10*(3), 034009.

Clevers, J. (1997). A simplified approach for yield prediction of sugar beet based on optical remote sensing data. *Remote Sensing of Environment*, *61*(2), 221–228.

Daughtry, C., Gallo, K., & Bauer, M. E. (1983). Spectral estimates of solar radiation intercepted by corn canopies. *Agronomy Journal*, *75*(3), 527–531.

De Wit, A. de, & Van Diepen, C. (2007). Crop model data assimilation with the ensemble kalman filter for improving regional crop yield forecasts. *Agricultural and Forest Meteorology*, *146*(1), 38–56.

Didan, K., & Huete, A. (2006). MODIS Vegetation Index Product Series Collection 5 Change Summary.

District level agricultural yields. (n.d.). *District level agricultural yields*. Department of Agriculture, Cooperation; Farmers Welfare: Ministry of Agriculture; Farmers Welfare, Govt. of India. Retrieved from <http://apy.dacnet.nic.in/>

Doraiswamy, P. C., Moulin, S., Cook, P. W., & Stern, A. (2003). Crop yield assessment from remote sensing. *Photogrammetric Engineering & Remote Sensing*, *69*(6), 665–674.

Doraiswamy, P. C., Stern, A. J., & Akhmedov, B. (2007). Crop classification in the us corn belt using modis imagery. In *2007 ieee international geoscience and remote sensing symposium* (pp. 809–812). IEEE.

Elhorst, P. (2017). Spatial Panel Data Models. Retrieved from <http://regroningen.nl/elhorst/doc/Spatial Panel Data Models.pdf>

Elhorst, P. J. (2010). Spatial Panel Data Models. In M. M. Fischer & A. Getis (Eds.), *Handbook of applied spatial analysis: Software tools, methods and applications*. New York: Springer.

Fotheringham, S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression‚ The analysis of spatially varying relationships*. (Anonymous, Ed.). Hoboken: Wiley.

Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X., Muchoney, D., Strahler, A. H., … others. (2002). Global land cover mapping from modis: Algorithms and early results. *Remote Sensing of Environment*, *83*(1), 287–302.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. (2010a). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*(1), 168–182. Retrieved from <http://www.sciencedirect.com/science/article/B6V6V-4XFNCW6-1/2/926cb3ceff79aa59819a94aa7fa4d00e>

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. (2010b). MODIS collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*(1), 168–182.

Funk, C., & Budde, M. E. (2009). Phenologically-tuned modis ndvi-based production anomaly estimates for zimbabwe. *Remote Sensing of Environment*, *113*(1), 115–125.

Gao, F., Anderson, M. C., Zhang, X., Yang, Z., Alfieri, J. G., Kustas, W. P., … Prueger, J. H. (2017). Toward mapping crop progress at field scales through fusion of landsat and modis imagery. *Remote Sensing of Environment*, *188*, 9–25.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, *185*, 1–17. <https://doi.org/http://dx.doi.org/10.1016/0003-2670(86)80028-9>

Gray, J., Friedl, M., Frolking, S., Ramankutty, N., Nelson, A., & Gumma, M. K. (2014). Mapping asian cropping intensity with modis. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(8), 3373–3379.

Greene, W. (1997). H.(1997): Econometric analysis. New York: Prentice Hall.

Groten, S. (1993). NDVI—crop monitoring and early yield assessment of burkina faso. *TitleREMOTE SENSING*, *14*(8), 1495–1515.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models* (Vol. 43). CRC Press.

Hatfield, J. (1983). Remote sensing estimators of potential and actual crop yield. *Remote Sensing of Environment*, *13*(4), 301–311.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*(4), 361–365.

Idso, S. B., Jackson, R. D., & Reginato, R. J. (1977). Remote-sensing of crop yields. *Science*, *196*(4285), 19–25.

Kumar, U. (2017, June). Discussion with wheat expert. CIMMYT.

Li, L., Friedl, M. A., Xin, Q., Gray, J., Pan, Y., & Frolking, S. (2014). Mapping crop cycles in china using modis-evi time series. *Remote Sensing*, *6*(3), 2473–2493.

Lobell, D. B., & Asner, G. P. (2004). Cropland distributions from temporal unmixing of modis data. *Remote Sensing of Environment*, *93*(3), 412–422.

Lobell, D. B., Schlenker, W., & Costa-Roberts, J. (2011). Climate trends and global crop production since 1980. *Science*, *333*(6042), 616–620. <https://doi.org/10.1126/science.1204531>

Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333.

MacDonald, R. B., & Hall, F. G. (1980). Global crop forecasting. *Science*, *208*(4445), 670–679.

Mann, M. L., & Warner, J. (2015). *Ethiopian Wheat Yield and Yield Gap Estimation: A Small Area Integrated Data Approach*. Addis Ababa, Ethiopia: International Food Policy Research Institute.

Mann, M. L., & Warner, J. M. (2017). Ethiopian wheat yield and yield gap estimation: A spatially explicit small area integrated data approach. *Field Crops Research*, *201*, 60–74.

Mann, M. L., Batllori, E., Moritz, M. A., Waller, E. K., Berck, P., Flint, A. L., … Dolfi, E. (2016). Incorporating anthropogenic influences into fire probability models: Effects of human activity and climate change on fire activity in california. *PloS One*, *11*(4), e0153589.

Mann, M. L., Berck, P., Moritz, M. A., Battllori, E., Baldwin, J. G., Gately, C., & Cameron, D. R. (2014). Modeling residential development in California from 2000-2050: Integrating wildfire risk, wildland and agricultural encroachment. *Land Use Policy*, *41*, 438–452. Retrieved from <https://www.researchgate.net/profile/Michael{\_}Mann6>

Mann, M. L., Kaufmann, R. K., Bauer, D., Gopal, S., Vera-Diaz, M. D. C., Nepstad, D., … Amacher, G. (2010). The economics of cropland conversion in Amazonia: The importance of agricultural rent. *Ecological Economics*, *69*, 1503–1509.

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.

Mo, X., Liu, S., Lin, Z., Xu, Y., Xiang, Y., & McVicar, T. (2005). Prediction of crop yield, water consumption and water use efficiency with a svat-crop growth model using remotely sensed data on the north china plain. *Ecological Modelling*, *183*(2), 301–322.

Moriondo, M., Maselli, F., & Bindi, M. (2007). A simple model of regional wheat yield based on ndvi data. *European Journal of Agronomy*, *26*(3), 266–274.

NASA. (1984). *AgRISTARS: Agriculture and resources inventory surveys through aerospace remote sensing. technical report research report*. *AgRISTARS: Agriculture and resources inventory surveys through aerospace remote sensing. Technical Report Research Report*. NASA.

NASS, U. (2003). USDA-national agricultural statistics service, cropland data layer. *United States Department of Agriculture, National Agricultural Statistics Service, Marketing and Information Services Office, Washington, DC [Available at Http//Nassgeodata. Gmu. Edu/Crop-Scape, Last Accessed September 2012.]*.

O’Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, *41*(5), 673–690.

Pinter Jr, P. J., Ritchie, J. C., Hatfield, J. L., & Hart, G. F. (2003). The agricultural research service’s remote sensing program. *Photogrammetric Engineering & Remote Sensing*, *69*(6), 615–618.

Portmann, F. T., Siebert, S., & Döll, P. (2010). MIRCA2000—Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling. *Global Biogeochemical Cycles*, *24*(1).

Ramankutty, N., Evan, A. T., Monfreda, C., & Foley, J. A. (2008). Farming the planet: 1. geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, *22*(1).

Rasmussen, M. S. (1992). Assessment of millet yields and production in northern burkina faso using integrated ndvi from the avhrr. *International Journal of Remote Sensing*, *13*(18), 3431–3442.

Ray, D. K., Gerber, J. S., MacDonald, G. K., & West, P. C. (2015). Climate variation explains a third of global crop yield variability. *Nature Communications*, *6*, 5989. Retrieved from <http://dx.doi.org/10.1038/ncomms6989 http://10.1038/ncomms6989 http://www.nature.com/articles/ncomms6989{\#}supplementary-information>

Robles, M., Ceballos, F., & Kramer, B. (2015). Focus Group Interviews: Haryana and Punjab India. India: IFPRI.

Sidhu, R., & Vatta, K. (2012). Risk in punjab agriculture: Current status and emerging issues. *Unpublished Paper, PAU Ludhiana*.

Stata. (2016). XTREG Manual. Retrieved from <http://www.stata.com/manuals13/xtxtreg.pdf>

State, P. (2016). Detecting Multicollinearity Using Variance Inflation Factors. Retrieved from <https://onlinecourses.science.psu.edu/stat501/node/347>

Thenkabail, P. S., Biradar, C. M., Noojipady, P., Dheeravath, V., Li, Y., Velpuri, M., … others. (2009). Global irrigated area map (giam), derived from remote sensing, for the end of the last millennium. *International Journal of Remote Sensing*, *30*(14), 3679–3733.

Thenkabail, P., GangadharaRao, P., Biggs, T., Krishna, M., & Turral, H. (2007). Spectral matching techniques to determine historical land-use/land-cover (lulc) and irrigated areas using time-series 0.1-degree avhrr pathfinder datasets. *Photogrammetric Engineering & Remote Sensing*, *73*(10), 1029–1040.

Tucker, C., & Sellers, P. (1986). Satellite remote sensing of primary production. *International Journal of Remote Sensing*, *7*(11), 1395–1416.

Tucker, C., Holben, B., Elgin Jr, J., McMurtrey III, J., & others. (1980). Relationship of spectral data to grain yield variation. *Photogrammetric Engineering and Remote Sensing*, *46*(5), 657–666.

Utt, N. (2017, March). Discussion with irrigation expert. University of Minnesota.

Wardlow, B. D., & Egbert, S. L. (2008). Large-area crop mapping using time-series modis 250 m ndvi data: An assessment for the us central great plains. *Remote Sensing of Environment*, *112*(3), 1096–1116.

Xiao, X., Boles, S., Frolking, S., Li, C., Babu, J. Y., Salas, W., & Moore, B. (2006). Mapping paddy rice agriculture in south and southeast asia using multi-temporal modis images. *Remote Sensing of Environment*, *100*(1), 95–113.

1. A nice synopsis of VIF can be found here (State, 2016). [↑](#footnote-ref-45)
2. The first four principal components in this PCA comprises 77 % of total variance in [↑](#footnote-ref-48)
3. The within- evaluates the goodness of fit beyond what can be explained by fixed effects intercepts [↑](#footnote-ref-68)