Responsible Land Governance:
Towards an Evidence Based Approach
ANNUAL WORLD BANK CONFERENCE ON LAND AND POVERTY
WASHINGTON DC, MARCH 20-24, 2017

PIXEL LEVEL LAND USE AND THE IMPACTS OF BIOPHYSICAL FACTORS

Jingyu Song[*], Michael S. Delgado[*], Paul V. Preckel[*], and Nelson B. Villoria[**]

[*] Department of Agricultural Economics, Purdue University, West Lafayette, IN 47907, USA
[**] Department of Agricultural Economics, Kansas State University, Manhattan, KS 66506, USA

Corresponding author: Jingyu Song (song173@purdue.edu)

Paper prepared for presentation at the
"2017 WORLD BANK CONFERENCE ON LAND AND POVERTY"
The World Bank - Washington DC, March 20-24, 2017

Responsible Land Governance:
Towards an Evidence Based Approach
Annual World Bank Conference on Land and Poverty
Washington DC, March 20-24, 2017

Abstract

Despite substantial research and policy interest in pixel level cropland allocation data, few sources are available that span a large geographic area. The data used for much of this research are often derived from complex modeling techniques that may include model simulation and other data processing. We develop a transparent econometric framework that uses pixel level biophysical measurements and aggregate cropland statistics to develop pixel level cropland allocation predictions. Such pixel level land use data can be used to investigate the impact of human activities on the environment. Our framework also provides marginal effects of changes in climatic and biophysical factors on cropland allocation at the pixel level, which can be used in a variety of research and policy contexts such as to assess the impact of global warming on cropland use. Validation exercises show that our approach is effective at predicting cropland allocation at multiple levels of resolution.

Key Words: fractional regression, global cropland allocation, biophysical impacts, pixel level measurements, quasi-maximum likelihood

1. Introduction

Agricultural productivity and environmental sustainability are central focuses for policymakers and academics. Understanding the interaction between agricultural systems and economic and environmental systems is critical for enhancing public policies related to economic development, food security, and human well-being. A major component of this interaction is climate change and adaptation, for which agriculture is central (e.g., Table 2 in Burke et al., 2015), especially for developing nations where the agricultural sector is more important and the impacts of climate change are most severe (Auffhammer and Schlenker, 2014; Hertel and Lobell, 2014; Kyle et al., 2015; Mendelsohn, 2009). Agricultural adaptation is inevitable as climates change and global crop production is affected, and the greatest pressure is likely to be on developing countries in which the agricultural sector can be highly vulnerable and the capacity to adapt may be the lowest (Hertel and Lobell, 2014). As all countries face challenges in adaptation, understanding patterns and changes in cropland allocation is central to policy designs geared toward promoting agricultural productivity while ensuring environmental sustainability.[1]

An important challenge for applied research is a lack of data on cropland allocation below the national or subnational level over a wide geographic area. Typically, cropland allocation data are collected via national census or survey instruments, and measures the total amount of land under crop in a particular state or province. These data, however, do not usually indicate the distribution of cropland within that state. Figure 1 illustrates the issue using Mexico as an example. In the figure, the color of each state indicates the fraction of the harvested land area for maize in the state, with a darker shade of green indicating a greater harvested area in maize. The inset image shows a 5-arc minute pixelated grid placed over part of Mexico; researchers and policymakers often desire to know the distribution of the total harvested area for maize within a state over a set of pixels in the grid in order to understand the effects of changes in economic or biophysical

[1] Cropland allocation is also closely linked to regional, national, and international food security (Mueller et al., 2014), international trade in agricultural commodities (Polasky et al., 2004), international development (Schaldach et al., 2011; Henderson et al., 2012), environmental issues such as domestic and trans-boundary pollution, groundwater quality and consumption, and pest management (Rosegrant et al., 2002; Lapola et al., 2010; Mallory et al., 2011; Fezzi and Bateman, 2011), and it is important for urban and regional planning (Burchfield et al., 2006; Klein Goldewijk et al., 2010).

conditions. Simply assuming that the pixel level allocation of cropland in a state is equal to the aggregate state level allocation is *ad hoc*, and may mask pixel level heterogeneity that bears implications for national and international research and policy.[2]

Previous economic analyses of issues related to land use have demonstrated the benefits of using cropland allocation data measured at a fine geographic scale, relative to using the state/provincial or national aggregates. For instance, Auffhammer et al. (2013) describe how aggregate climate measures mask important spatial heterogeneity that is measureable using pixelated data. Hendricks et al. (2014) demonstrate that county averaged data in the United States leads to qualitatively significant statistical bias in estimates of crop acreage response to price shocks, and advocate using spatially explicit pixel level data to avoid this bias. In short, pixel level data is important for research and policy related to cropland allocation because it reflects heterogeneity that is germane to land use decisions.

Yet, much of the evidence indicating advantages of pixel level data is based on analysis for the United States, where pixel level measurements are relatively abundant and are generally reliable. Researchers have lamented the dearth of pixel level measurements over a wider geographic scale; moreover, data availability and data quality problems are more severe in developing countries and regions such as Sub-Saharan Africa, where lack of reliable data limits statistical analysis (Auffhammer and Schlenker, 2014; Lobell et al., 2008; Schlenker and Lobell, 2010). In some countries, observed cropland allocation data (i.e., the data measured by national census and survey reports) is simply not available below a national or subnational (e.g., state/provincial) level. This lack of widespread availability of pixel level cropland allocation data prohibits analysis of land allocation at the pixel resolution across broad regions that researchers desire; it also limits the reliability of high spatial resolution environmental models that use land use data as inputs. To address these issues, researchers have turned to satellite imagery or simulation methods to obtain pixel level measurements of cropland allocation. These data, however, have been criticized for being too highly processed and may be difficult to reproduce (see discussion below); the contribution in this paper is a statistical approach that simultaneously

[2] Various terms including gridded, parcel, pixel, and fine-scale have been used interchangeably in the literature. To avoid confusion, we use the term pixel.

allows us to understand the allocation of harvested area in a particular crop at a pixel level, as well as ascertain how marginal changes in climate and other biophysical measurements impact that allocation.

Specifically, we develop a quasi-maximum likelihood estimation framework that can be used to predict cropland allocation at a pixel resolution. We first use available information on built-up land and protected areas to eliminate urban centers and protected areas such as national parks, and then use pixel level variation in biophysical factors and the observed aggregate land allocation data to determine where in each state/province the crops are located. We apply the framework to the Americas, focusing on application of the proposed approach to two cases: a single crop model for maize and a multi-crop model for maize, soybeans, and wheat.

The method we develop has the following four virtues. First, in contrast to other methods that are currently in use, our approach uses a (modified) fractional regression that is relatively simple and transparent. We believe simplicity and transparency are critical; the research that uses these data supports policy development with deep international implications. Transparency in the data and its handling leads to clearer insights from the research that uses the data. Second, our method is relatively parsimonious, and is not dependent on the availability of specialized measurements that typically require *ad hoc* assumptions and complex data manipulations that are not motivated by the estimation framework.[3] Yet, the method is flexible and can accommodate additional variables in the event that they become available. Third, we can recover the marginal effects of the variables in the model on the allocation of cropland across the pixels within each state. These marginal effects are allowed to be heterogeneous across pixels, and these can be used to explore the impacts of perturbations in the climate and land attribute variables on pixel level cropland allocation.[4] Models that rely on satellite imagery are unable to estimate marginal impacts; that is, they can only observe the current situation and cannot project the impact of a change in the underlying determinants of cropland allocation. Fourth, the regression framework is developed

[3] For example, Li et al. (2014) use travel time to major cities calculated based on transport survey data, speed, routes, and road surface conditions, as a measure of transportation costs. This kind of information may not be generally available over a large geographic area.

[4] Establishing causality in the climate-cropland context is a difficult and well-known challenge. See Section 2 for further discussion.

generally, and can accommodate any number of cropping options – i.e., our model allows for either a single-crop or multi-crop structure.[5]

Our approach for downscaling aggregated data from censuses and other national statistics is purposefully different from the methods used by other researchers who have focused on land allocation. For instance, Ramankutty and Foley (1998), Ramankutty et al. (2008) and Monfreda et al. (2008) combine satellite-derived land cover data with agricultural inventory data to develop a global land use database measured at the 5 arc-minute pixel level. The Monfreda et al. (2008) dataset is the most comprehensive, spanning 175 distinct crops across the world, and is regarded as the standard in economic models of cropland allocation (Hertel et al., 2009). An important contribution of our work is that we identify cropland allocations without relying heavily on satellite imagery, thus reducing the uncertainties associated with discrepancies across reported land use patterns in different sources of satellite images (Ramankutty et al., 2008). This also distinguishes our work from You and Wood (2006), who use a cross entropy approach to predict cropland allocation that relies on existing satellite-based land cover data, a broad set of biophysical and socioeconomic factors, as well as model-generated indicators of land suitability for specific crops (FAO, 1981; Fischer et al., 2000). These indicators act as prior information of the most likely crop to be found in a given pixel. In contrast, our multi-crop model identifies land suitability based solely on the variability of biophysical attributes. Li et al. (2014) estimate land allocation as a function of geo-referenced biophysical factors – some of which include crop-specific land suitability variables – and spatially explicit producer prices for the Democratic Republic of Congo. Relative to our model, their approach has two main limitations. First, they use a binary outcome logit regression model that restricts the allocation within each pixel to a single land use. Second, their model adopts pixel (1 km by 1 km) measurements that are not widely available, rendering application of their model to a larger geographic area challenging. We believe our approach offers several important advantages over these other techniques. As with the data generated using these other methods, our data can be used to support counterfactual scenarios with different climate or

[5] Though we focus on cropland allocation, the model is also applicable to any empirical setting in which the outcome variable is a share observed at an aggregate level and the conditioning variables are observed at a disaggregated level.

soil settings, bridging the gap between the integrated assessment models (IAMs) and reduced form studies (Auffhammer and Schlenker, 2014).

The rest of the paper proceeds as follows. Section 2 describes our approach. We formalize the econometric problem and develop a robust quasi-maximum likelihood framework based around a modified fractional logit regression. In Section 3, we illustrate our approach by developing empirical models of cropland allocation for maize as a single crop and for multiple crops (maize, soybeans, and wheat) simultaneously across North, Central and South America. Section 3 also describes the data. Section 4 presents our statistical results including the marginal effects of pixelated data on predicted cropland allocation. Section 5 conducts both in-sample and out-of-sample validation exercises to assess the predictive performance of the model. We also provide evidence that our approach is capable of reliably predicting cropland allocation at the pixel level. In Section 6 we demonstrate how the model can be used to develop the pixel level cropland allocation data, and Section 7 provides conclusions.

2. Theoretical Framework

Model Preliminaries

We define a land parcel measured at the 5 arc-minute longitude/latitude level as a pixel. Basic biophysical land attributes such as temperature, precipitation, and soil pH are available at the pixel level; one important reason why these data are more readily available is that they are collected from globally positioned monitoring stations. Agricultural land use is observed at a more aggregated level – typically, the total area of cropped land and the share of cropped land devoted to a particular crop, observed at the subnational administrative unit level (e.g., the state or provincial level).[6] These data come directly from national census or survey instruments. These subnational administrative units are composed of pixels (as in Figure 1). Our goal is to estimate the fraction of each individual pixel that is cropped in a particular crop given the available pixel level biophysical measurements and aggregate land shares.

[6] Formally, we will call the state or provincial level Administrative Unit Level 1, and we will call the district or county level Administrative Unit Level 2.

Papke and Wooldridge (1996) propose a quasi-maximum likelihood estimator to analyze models with fractional response; Mullahy (2015) extends the model to the case of multivariate fractional response. Maximizing a Bernoulli log-likelihood function produces consistent estimates of the structural parameters, and a logistic function can ensure that the fitted values are restricted to the unit interval (Gourieroux et al., 1984; Wooldridge, 1991). Our case is similar to Papke and Wooldridge (1996) and Mullahy (2015), but not identical. The difference is that our conditioning variables are measured at the pixel level while the outcome is observed at the administrative level, which is more aggregated than the pixel level. In this structure, the outcome does not vary at the same level as the regressors, and aggregation is required.

To address this issue, we develop an aggregated fractional response model to accommodate the structure of the problem. The added value is that the model is capable of predicting cropland allocation at any level of spatial resolution at which the conditioning variables are measured or aggregated, regardless of whether the outcome is observed at that level.

Econometric Model and Estimation

Formally, let j index administrative units for $j = \{1,2,\ldots,J\}$ and k represent crops for $k = \{1,2,\ldots,K\}$.[7] Let y_{jk} be the observed fraction of land area in administrative unit j that is in crop k, such that $0 \leq y_{jk} \leq 1$. Let Z_{ijk} be the (unobserved) fraction of cropped land in pixel i in administrative unit j that is cropped in crop k, where $i = \{1,2,\ldots,I_j\}$ is the pixel index that allows the total number of pixels in each administrative unit to vary. Define $\boldsymbol{X}_{ij}$ to be an N-dimensional vector of observable biophysical attributes for pixel i in administrative unit j.

We are interested in estimating the parameters β in the conditional mean for pixel level share Z_{ijk}:

$$E\left[Z_{ijk} \middle| \boldsymbol{X}_{ij}\right] = G_{ijk}(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij}), \beta_k) \tag{1}$$

where $\boldsymbol{W}(\cdot): \mathbb{R}^N \to \mathbb{R}^M$ reflects transformations of the fundamental explanatory variables (linear, quadratic, interaction, etc.), $G(\cdot): \mathbb{R}^M \to \mathbb{R}$, $0 \leq G(\cdot) \leq 1$ is a function that maintains the unit

[7] We derive the model in terms of crops, but note that this econometric model generally fits any context with a similar data structure. In our empirical models, the Kth crop is for all crops not included in $K = \{1,\ldots,K-1\}$.

interval restriction on the conditional mean. We parameterize $G(\cdot)$ using a logistic function, and the predicted fraction of land in crop k in pixel i in administrative unit j becomes:

$$G_{ijk}\big(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij}), \beta_k\big) = \frac{\exp(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij})\beta_k)}{\sum_{i=1}^{K}\exp(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij})\beta_k)} \quad \text{where } \beta_1 = 0. \tag{2}$$

The $\beta_1 = 0$ normalization facilitates parameter identification relative to a base cropland allocation. We extend equation (2), which is defined at the pixel level, to the administrative unit level via aggregation – i.e., the predicted fraction of land in crop k in administrative unit j is equal to the average pixel fraction weighted by area. The predicted fraction of land in crop k in administrative unit j is

$$H_{jk} = \frac{\sum_{i\in I_j} G_{ijk}\big(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij}), \beta_k\big) A_{ij}}{\sum_{i\in I_j} A_{ij}} \tag{3}$$

where A_{ij} is the area of pixel i in state j. Function (3) aggregates our predicted land shares for pixel i to the administrative level, hence converting pixel level information to the administrative level so that the pixel level land attribute (biophysical) data can be used to explain cropland allocation. Given H_{jk} the quasi-log-likelihood function to be maximized with respect to the parameters β_k is:

$$\mathcal{L} = \sum_{j=1}^{J}\sum_{k=1}^{K} y_{jk} \ln H_{jk}. \tag{4}$$

In addition to being relatively simple to optimize to obtain parameter estimates, the quasi-maximum likelihood estimator is consistent regardless of the conditional distribution of y given x as long as the conditional mean is correctly specified and the distribution is a member of the linear exponential family (Gourieroux et al., 1984; Wooldridge and Papke, 1996). This makes the estimator relatively robust. This method is generally applicable to other cases in which only aggregate level data is available for the outcome, but pixel level estimates are desired. This framework may also be adapted to a panel data context using a correlated random effects approach (Papke and Wooldridge, 2008).

Gourieroux et al. (1984) and Wooldridge (1991, 1997) derive the asymptotic variance of the quasi-maximum likelihood estimator using the estimated information matrix and the outer product of the score. Papke and Wooldridge (1996) and Mullahy (2105) provide expressions for the univariate and multivariate fractional logit models; we follow the same asymptotic variance calculation but base the formulation at the aggregated state level H_{jk}. The estimated asymptotic variance of β_k is the diagonal of:

$$F^{-1}BF^{-1}/J \quad (5)$$

where F^{-1} denotes the inverse Hessian and B denotes the outer product of the score.[8,9]

Cropland Allocation, Marginal Effects, and Policy Analysis

An important difference in our approach to predicting cropland allocation is that we use variation in pixel level biophysical measurements to generate our predictions, rather than relying on satellite image derived crop shares for each pixel (e.g., Monfreda et al., 2008). This gives us two unique advantages. First, we project cropland allocation based on pixel level biophysical factors and aggregate cropland statistics, which allows for estimation when satellite imagery is not available or is unable to distinguish between crop types.

Second, the econometric approach allows us to derive the effects of marginal changes in pixel level biophysical factors on cropland allocation. The marginal effect on the fraction of cropland in crop k in pixel i in state j with respect to regressor X_{ijm} is

$$\frac{\partial G_{ijk}(\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij}),\beta_k)}{\partial X_{ijm}}. \quad (6)$$

The exact form of this marginal effect depends on the transformations in $\boldsymbol{W}_{ij}(\boldsymbol{X}_{ij})$ and the specification of the index function.

Given the logistic parameterization of the likelihood model, we also calculate the odds ratio. Holding everything else constant, an odds ratio greater than one means that a one unit change in $\boldsymbol{X}$ would make the fraction of land in crop k larger, while an odds ratio less than one means that a one unit change in $\boldsymbol{X}$ leads to a decrease in the fraction of land in the crop. For the simple case that a variable enters linearly into the index function, the odds ratio is $\exp(\beta_{km})$.

8 An alternative is to use a block bootstrap that preserves the pixel to state ratio across the bootstrap replications (i.e., sample all pixels from each bootstrap sampled states). The authors' calculations have shown this bootstrap procedure yields standard errors that are virtually identical to those from the asymptotic variance-covariance formulation.

9 An open access tool deploying this framework is available at https://mygeohub.org/tools/flat. Source code and data used in this paper are free to download via the tool user interface.

Remark 2.1 We have described the pixel level measurements as biophysical factors, though clearly, economic and institutional forces also have a substantial impact on land allocation and crop choice. Our model structure does not preclude the inclusion of additional variables. Currently, we are operating under the constraint that measurements of these other factors are not available at a pixel level across a broad geographic area, either because the data do not exist, or because factors do not vary at a pixel level (such as national or state laws, institutional structures, and in many cases, prices). This constraint is not unique to our work. We control for these additional factors to the best of our ability using national level indicators (see Section 3).

Remark 2.2 The parameter estimates and pixel level predictions are obtained under the assumption that factors related to political and economic differences (e.g., laws, agricultural policies, consumer preferences, etc.) are held constant via the indicators. Likewise, the marginal effects should be interpreted as a (potential) re-allocation of cropland induced by changes in biophysical measurements, holding the country level factors constant.

3. Data and Empirical Models

We develop two empirical applications: a single-crop model for maize, and a multi-crop model of maize, soybeans, and wheat. For both models, we focus on harvested crop area spanning North, Central and South America. The 5 arc-minute resolution is a commonly used pixel measurement (Ramankutty and Foley, 1999; Erb et al., 2007), and yields pixels of about 100 square kilometers at the equator, 60 square kilometers in Minnesota, and nearly zero square kilometers near the north/south pole. Table A.1 in the appendix provides a comprehensive list of all data we employ, including the units of measurement and source.

Harvested Land Area

We use data series of built-up land and protected areas to exclude pixels that are in urban centers or protected areas such as national parks and forests. Total area in a pixel, A_{ij}, is calculated based on longitude and latitude. The total land area in an administrative unit is collected from Statoids (see Table A.1). We restrict the sample of states to countries that have administrative units with at least 0.5 percent total land area in maize for the maize model and at least 0.5 percent total land area in each of the three crops for the multi-crop model. This excludes administrative units where area in these crops is negligible. We also require that the sum over the shares of cropland in maize, soybeans, and wheat does not exceed 1 for the multi-crop model.[10]

The total areas of land harvested in maize, soybeans, and wheat at Administrative Unit Level 1 are collected from FAO Agro-MAPS (see Table A.1), which is the largest source of subnational agricultural harvested land area data (Monfreda et al., 2008). These data come from national or subnational census or survey statistics. In these data, the years for which harvested area data is available vary across countries; to build our sample we choose the years that are closest to the year 2000 to form a circa 2000 dataset (Ramankutty et al., 2008 also use a circa 2000 dataset). For example, state level harvested maize area data for the United States is not available for the year 2000 from Agro-MAPS, so instead we use 2001 data. For most of the countries in our analysis the data measurements come either from the year 2000/2001 or from the mid-late 1990s; the largest difference is for Costa Rica, for which the data is measured in 1984. The gaps in these data arise because the national censuses and surveys are typically done at five year intervals. Further, not every administrative unit has data for maize, soybeans, and wheat, and not every unit is reported in Agro-MAPS. Hence we only include units with reported FAO data. Since there are a few cases where the names of the units have changed in recent years or a unit has been divided into separate units, we verify the administrative divisions using the CIA World Factbook and Global Administrative Areas (GADM) database (see Table A.1) to ensure a consistent set of administrative units for circa 2000.

Pixel Level Biophysical Data

The land attribute data comes from Villoria and Liu (2015) and includes biophysical variables to measure biophysical conditions that influence crop choice. These variables include the average temperature over the growing season in degrees Celsius, average annual precipitation in meters, elevation in thousands of meters, the soil pH level (defined over the range 0 - 14), soil carbon content (kg per square meter, 0 to 1 meter depth), land slope (from almost flat to steep, 0.0025 -

[10] There are a few states for which the sum of land shares in maize, soybeans, and wheat is greater than 1. This can be a result of multiple cropping seasons, or data inconsistencies. We leave these issues for future research.

0.725), and latitude. Elevation and latitude data are at the 5-minute degree level; other variables are available at a 30 arc-minute resolution and are downscaled assuming all the 5-minute degree cells within a 30-minute degree cell have the same value. Our preferred specification is quadratic in temperature, precipitation, and latitude. The interaction between temperature and precipitation is also included; this specification of temperature and precipitation is similar to that used by Lobell et al. (2011), Lobell et al. (2013), Schlenker et al. (2006), and Schlenker and Roberts (2009). We define soil pH to be the soil pH deviation from pH6.5 as pH6.5 is approximately optimal for maize, soybeans, and wheat (Lerner and Dana, 2001; Mallarino et al., 2011).[11] We expect the fraction of a pixel that is cropped in maize to decrease with an increase in the deviation (above or below) of soil pH from the optimal pH6.5 level. To allow the response above pH6.5 and below pH6.5 to be asymmetric, we include two variables max(pH6.5-pH, 0) and max(pH-pH6.5, 0). Table 1 reports descriptive statistics for the variables, divided into the North, Central and South American regions.

Indicator Variables and Final Sample

We add binary indicators for each country included in the analysis (see Table 2 for included countries), with a group of South American countries (Ecuador, Peru, Paraguay, and Uruguay) excluded as the base group for the maize model. These indicators allow us to account for political and economic factors that influence crop production but are unobservable and/or do not vary within each country. For the multi-crop model, we include country indicators for Argentina and the United States and use the rest of the countries as the base.[12]

Based on our selection criterion, 196 administrative units from 18 countries are included in the maize model. These countries are listed in Table 2 with the number of states included from each country in parentheses. For the multi-crop model, a total of 40 states from 5 countries are included.

4. Parameter Estimates and Marginal Effects

[11] http://www.cropnutrition.com/efu-soil-ph#soil-acidity. Accessed on Aug 30, 2015.

[12] Five countries are included in the multi-crop model based on the sample selection criteria. Different model specifications show that the model with Argentina and United States indicators yields the lowest root mean squared error.

The Single-Crop Model

The coefficient estimates and standard errors for the maize model are reported in Table 3, and the implied marginal effects and odds ratios are reported in Table 4.[13] These coefficient estimates and marginal effects can be immediately deployed in a variety of policy analysis contexts to understand how changes in biophysical factors might influence cropland allocation. Temperature, precipitation, soil pH (both deviations below and above pH6.5), soil carbon, slope, and latitude are all statistically significant. Temperature has a positive impact on the fraction of maize in a pixel. The average temperature of the growing season in our sample is 19.34 degrees Celsius. All else equal, a one degree Celsius increase in temperature above this average increases the fraction of maize by about 0.35 percent (Table 4). The average precipitation in our sample is 1 meter. All else equal, a one meter increase in precipitation increases the maize fraction by about 1.03 percent. In Figure 2a we plot the relationship between temperature, precipitation, and the predicted fraction of maize, evaluating all observations at the base group for the country indicators while holding other variables constant at their mean. The average growing season temperature ranges from -0.17 to 28.67 degrees Celsius and the average precipitation is between 0 and 5.67 meters. At low levels of precipitation, as temperature increases the maize fraction first increases and peaks around 10 degrees Celsius, then decreases to about 0. Holding temperature constant, the maize fraction increases first, then drops to about 0. At low levels of precipitation, an increase in precipitation leads to a rapid increase in the maize fraction, while at higher levels of precipitation, a further increase in precipitation does not affect the maize fraction.

Deviation from the optimal soil pH of 6.5 decreases the fraction of maize. A decrease in pH below pH6.5 reduces the maize fraction by about 7.87 percent, and an increase in pH above pH6.5 reduces the fraction of maize by about 13.83 percent. Average soil carbon content is 5.83 kg per square meter; the maize fraction increases by about 1.34 percent as soil carbon content increases by 1 kg per square meter. Slope has a negative impact on the fraction of maize. As land becomes steeper, the maize fraction decreases. Latitude has a positive impact on maize fraction for the studied area. As latitude increases, the maize fraction increases.

[13] Following Greene (2010), we assess statistical significance in our model via the t-values on the parameters, and then report and draw economic conclusions directly from the implied marginal effects.

To further investigate the relationships among the variables and the fraction of maize, we plot the maize fraction as a function of each variable, holding the other variables constant at their means (Figure 2b). Pixel level soil pH ranges from 4.20 to 8.17, and deviation from above and below the optimal soil pH of 6.5 leads to a decrease in the maize fraction. Latitude ranges between -40.96 and 56.79 degrees. We see an increase moving from South to North America.

The Multi-Crop Model

Similar to the maize model, we report the coefficient estimates for each regressor in the multi-crop model in Table 5. The significance of these variables varies across crops. Temperature and precipitation are significant for maize and soybeans. Temperature squared and the interaction between temperature and precipitation are significant for all three crops. Precipitation squared is significant for soybeans, but not for maize and wheat. The land attribute variables – slope and soil pH above pH6.5 – have a significant impact on the crop fraction for all three crops. Elevation, soil pH below pH6.5, and latitude, have a significant impact on the maize and soybean fractions, Implied marginal effects and odds ratios are displayed in Table 6.

Temperature has a negative impact on the fraction of all three crops on average within the studied area. The average temperature of the growing season in our sample of administrative regions is 18.39 degrees Celsius. A one degree Celsius increase in temperature above this average decreases the fraction of maize by 1.19 percent, soybeans by 2.48 percent, and wheat by 0.51 percent, indicating that maize and soybeans are more sensitive to changes in temperature than wheat. Precipitation has a negative impact on the fractions of soybeans and wheat, but a positive impact on maize. Specifically, a one meter increase in precipitation from the average level of 0.95 meters leads to an increase in the fraction of maize of 20.41 percent, and a decrease in the fraction of soybeans of 22.57 percent, and wheat of 1.34 percent. Deviation in soil pH from pH6.5 has a negative impact on maize, soybeans, and wheat, and slope has a negative relationship with soybeans and wheat. Latitude, on average, has a positive relationship with the fractions of all three crops, meaning that as we move north to states in the Northern United States, crop fractions tend to increase. The calculated odds ratios imply similar results as the marginal effects.

Figure 3(a) shows a 3-dimensional plot of the relationship between temperature, precipitation, and the soybean fraction. The lowest average temperature in the growing season across all included states is 8.33 degrees Celsius, and the highest is 27.00 degrees Celsius. As temperature increases, the soybean fraction first increases and decreases slightly, then increases and decreases; the minimum level of average annual precipitation is 0.07 meters, and the maximum is 1.94 meters; as precipitation increases, the soybean fraction decreases then increases.

Graphical illustrations for the other variables that have statistically significant impacts on the soybean fraction are shown in Figure 3(b). Elevation has a negative impact on soybean fraction. The minimum soil pH is 4.90. Deviation from below the optimal soil pH of 6.5 leads to a decrease in the soybean fraction. The maximum soil pH is 8.05. Soybean fraction increases in the studied area when soil pH deviates from 6.5. Slope has a negative impact on the soybean fraction: as land gets steeper the soybean fraction decreases. Latitude has a positive overall impact on the soybean fraction. Moving towards the Northern Hemisphere, the soybean fraction first increases, then stays high around the Central American countries and Mexico, then decreases moving further north passing the United States Corn Belt.

5. Model Validation

In-Sample and Out-of-Sample Validation

In general, the estimated relationship between the biophysical variables and the crop fraction within each pixel is consistent with our expectation. We now validate the predictive power of our model, considering both in-sample and out-of-sample validation. By out-of-sample validation, we mean that the estimation relies on Level 1 data, and we validate our results using Level 2 data. Ideally, Level 2 data should add up to Level 1 values. However, we do not have complete Level 2 data for all counties/districts, but using the available Level 2 data enables us to validate prediction at a spatial level that is finer than the level of data that we use to estimate the model. To validate, we compare the crop fraction predicted by our model to the actual crop fraction reported by FAO Agro-MAPS at both Administrative Levels 1 and 2. Level 2 FAO data are available for only 108 states for the maize model, and 32, 35, and 35 states for the multi-crop model for maize, soybean, and wheat. USDA NASS Quick Stats includes more Level 2 units (counties) in their database for the United States; for Level 2 validation for the United States, we rely on USDA NASS data instead of Agro-MAPS Level 2 data.

The predicted fraction of a crop in each pixel comes from equation (2) given the estimated parameters. The predicted harvested area for a crop in a pixel equals the total land area in the pixel times the fraction of the pixel that is in that crop. Summing over the predicted crop area in all the pixels in each Level 1 or 2 unit yields the total predicted area in that crop for the unit. To create the validation plot at the Administrative Unit Level 2, we first scale the predicted fractions at the pixel level so that the predicted total maize fraction at Level 1 matches the FAO fraction. Then we plot the estimation results at Level 2 based on the scaled fractions.

Figure 4 graphically displays the prediction results at Level 1 and Level 2 for the single-crop maize model – the horizontal axis shows the model predicted maize fraction and the vertical axis shows the FAO maize fraction. The dashed diagonal line in each plot represents the 45 degree line; the closer the points are to the 45 degree line, the better the prediction. As is illustrated, the points generally cluster around the 45 degree line at both levels, indicating that the model generally predicts well. To more precisely measure the correlation between the predicted Level 1 fraction of maize and the observed FAO Level 1 maize fraction, we regress the predicted fraction on the observed FAO fraction. The regression line is shown by the solid line (in red) in each plot. In the case of ideal prediction, we expect that the intercept coefficient is equal to zero, and the slope coefficient is equal to one. At the Administrative Unit Level 1, the estimated intercept coefficient is -0.003, and is not statistically different from zero; the estimated slope coefficient is 1.06, and is not statistically different from 1. In addition, we compute the squared correlation between the predicted and FAO maize fraction as an R-squared, which is 0.72.

At Level 2, the intercept coefficient is 0.04, and is significantly different from zero; the slope parameter is 0.96, and is significantly different from one. The squared correlation between the predicted and FAO maize fraction is 0.15.

We report the results from similar validation exercises for the multi-crop model in Figures 5(a) and 5(b). The figures show that at both Administrative Unit levels the points are all close to the 45 degree line, which indicates that the multi-crop model predicts well by both in-sample and out-of-sample metrics. The squared correlation is 0.8930 for maize, 0.8569 for soybeans, and 0.6383 for wheat.

From these validation exercises, we find that at Level 1, for which we have crop harvested area data, the prediction results are closer to the FAO numbers than those at Level 2. However, considering our parsimonious set of land attribute variables, and that we do not have any crop specific variables, all the fractions are identified from the variation of crop shares at Level 1.

Relative Performance of the Single and Multi-Crop Models

To assess the relative performance of the single and multi-crop models, we calculate the root mean squared error (RMSE):

$$RMSE = \left[\frac{1}{J}\sum_{j=1}^{J}\left(predicted\ crop\ fraction_j - FAO\ crop\ fraction_j\right)^2\right]^{\frac{1}{2}}. \qquad (8)$$

Table 7 shows the RMSE values for the maize model (with 196 states) and the multi-crop model (with 40 states), both at Administrative Levels 1 and 2. For these samples, the multi-crop model predicts slightly better at Level 1 for the maize fraction compared to the maize model. It is possible that the relatively better performance of the multi-crop model arises because it incorporates three types of crops; it is also possible that the differences in predictive performance arises because the samples are different. Both models predict relatively worse at Level 2 compared to each model's own performance at Level 1; yet the Level 2 prediction is out-of-sample.

We also consider the relative predictive performance of the single and multi-crop models using an identical sample of observations. With the same 40 states, the RMSE values for the maize model are 0.0396 at Level 1 and 0.2585 at Level 2. Compared to the RMSE values for the multi-crop model in Table 7, the maize RMSE is slightly worse at both Level 1 and Level 2, which indicates that when the sample is the same, the performance of the multi-crop model is better.

Validation against Alternative Data Sources

To provide additional insight into the reliability of our cropland predictions, we validate our predictions against two additional sources. The first is the Monfreda et al. (2008) predictions, given their widespread use in applied research. The second source is the USDA Cropland Data Layer (CDL), which is built on high resolution satellite images and agricultural surveys for several states in the United States and estimates crops growing in the field in June. For the available states, we take the CDL to be the benchmark (ground truth). For the year 2001 (the year of the United States data), the CDL has 30 by 30 meter pixel data for Illinois, Indiana, Iowa and North Dakota. To ensure comparability we aggregate the CDL pixels to match our pixel size and calculate the corresponding fractions. We compare the CDL crop shares to those from our maize model, multi-crop model (for maize), and the Monfreda et al. (2008) predictions.

Table 8 shows the correlations at the pixel level for all four states across the four measures of predicted maize fractions. The correlations between different model estimates are relatively close for states where maize is the primary crop (Illinois, Indiana and Iowa). For North Dakota, other crops (wheat, soybeans, sunflowers, canola, and barley) all take higher percentages of cropland than maize;[14] our models do not perform as well as the other models. We also check the performance of Monfreda et al. (2008) for maize for all 196 states included in our estimation, at both Level 1 and Level 2. The Level 1 RMSE is 0.02, and the squared correlation is 0.88; the Level 2 RMSE is 0.20, and the squared correlation is 0.15. Considering that both the CDL data and Monfreda et al. (2008) approach take into account a much larger number of crops, both are based on satellite imagery, and Monfreda et al. (2008) also incorporate Level 2 if data is available, it is not surprising that their predictions are closer to each other. However, the results indicate that in general, we are able to achieve comparably good estimates based on a rather parsimonious set of dependent and independent variables, without using satellite imagery to isolate crop location.

6. Illustrating the Pixel-Level Predictions

We provide two illustrations of the pixel-level cropland predictions generated by our model. In each case, we compare predictions from a naïve approach that assumes that the fraction of maize in each pixel is equal to the aggregate FAO Level 2 fraction of maize. That is, a constant (pixel level) fraction within each Level 2 area. The second approach uses our model and the pixel-specific shares. The first illustration is for Hidalgo, Mexico, and is shown in Figures 6(a) and 6(b). Colors from yellow to dark red indicate an increasing share of maize. We pick Hidalgo as our first example for two reasons. First, FAO Level 2 data is available for all 84 districts of Hidalgo, enabling us to fully compare the prediction results with FAO data. Second, there is relatively high variation in the fractions of maize within Hidalgo, making it an interesting case for comparison.

Figure 6(a) shows that the constant fraction approach predicts that the pixels in the north-central part of Hidalgo have a lower fraction of maize, and the pixels in the north-eastern and part of the south-eastern regions have higher fractions of maize. The model results in Figure 6(b) are quite different: the model predicts that a larger number of pixels in the northern part of Hidalgo

[14] https://nassgeodata.gmu.edu/CropScape/

have a low fraction of maize, and a larger number of southern pixels have a higher fraction of maize. As a means of validation, we compare the FAO Agro-MAPS state level data with data collected from the Agrifood and Fisheries Information Service of Mexico (*Servicio de Información Agroalimentaria y Pesquera*, SIAP México). The correlation between FAO and SIAP data is 0.85, which provides reasonable assurance that the Level 1 FAO data for Mexico is reliable. However, district level harvested area data for maize is not available from SIAP, so we are not able to validate FAO district level data for Hidalgo. Nonetheless, further investigation confirms the credibility of our downscaled predictions: the south-eastern part of Hidalgo is home to the Valley of Tulancingo which is known for being one of the most fertile parts of the Valley of Mexico – a largely agricultural area. Further, the north-eastern part of Hidalgo is largely forest. Our model correctly identifies that the south-eastern part is the part of Hidalgo where most of the maize is grown; and there is less maize in the north-eastern area. These spatial patterns further indicate the predictive capability of our model, but more importantly comparison of Figure 6(a) and 6(b) highlights the dangers in simply applying a uniform share to all pixels based on an aggregate share when downscaling.

Similarly, we generate two sets of downscaling predictions for harvested maize area in the United States – one with the average USDA Level 2 fraction applied to all pixels within a county where Level 2 data is available, and the other with our scaled predicted pixel level fractions. Figure 7(a) shows the USDA maize fraction plot at the pixel level assuming that all the pixels within each Level 2 unit have the same maize fraction that equals to the USDA Level 2 maize fraction for that unit and 7(b) presents the pixel level predicted maize fraction plot scaled using the FAO Level 1 maize total harvested area. The plots are alike for most areas. The northwestern corner of Minnesota stands out as our model predicts a higher maize fraction than FAO. Given that USDA NASS data indicates the primary crops in that area are wheat and sugar beets, our over-prediction of maize is not surprising since our empirical model focuses on three major crops across a large geographical area only, rather than incorporating locally important crops. But with the flexibility offered by our framework, a potential study is to build a more comprehensive model including more crops.

These illustrations show that our model predicts well over a large geographic scale.

7. Conclusion

We develop a statistical method for predicting pixel level cropland allocation across a (large) geographic area in which pixel-level measurements are not available. Specifically, we develop a fractional response model that combines measurements of pixel level land attributes with observable aggregate land use patterns to predict the share of cropland allocated to a certain crop at the pixel level. We formulate the likelihood function and demonstrate application to a single-crop model for maize and a multi-crop model for maize, soybeans, and wheat. We show that both the single-crop and multi-crop models at the Administrative Unit Levels 1 and 2 are reasonably precise in predicting cropland allocation at the pixel level.

Our statistical model and land allocation predictions provide applied scientists with measurements of land allocation at a pixel level, and with important advantages over previous measurements. First, the statistical framework is straightforward and transparent. This allows users of the model to gain a clear understanding of the relationship between the explanatory variables and the allocations. Second, while the model returns fractional estimates of cropland allocation in each pixel, the model also provides estimates of the marginal impacts of the variables on the pixel level allocations, which can be used in a variety of different contexts, such as in the case of climate change. Third, the framework is flexible, and variables can be easily added if reliable data are available. For instance, if a variable such as travel time to markets is available, the model could be used to assess the impact of road infrastructure investments on cropping patterns. Finally, while we focus on North, Central and South America, our model can be readily applied to any continental or geographic area, such as Africa, where pixel level cropland allocation data is critically needed but not readily available.

Acknowledgements

The authors are grateful for comments and suggestions from Timothy Baker, Thomas Hertel, David Laborde, Navin Ramankutty, Juan Sesmero, and participants at the 2016 AAEA Annual Meeting, 2015 AAEA and WAEA Annual Meeting, IFPRI, Washington, DC, and the Purdue University SHaPE brownbag seminar. This research is supported in part by USDA grant

Agreement #58300010058, Purdue grant #105651; computational resources provided by Information Technology at Purdue – the Carter Cluster, Purdue University, West Lafayette, Indiana.

References

Auffhammer, M., Hsiang, S. M., Schlenker, W., and Sobel, A. (2013). Using weather data and climate model output in economic analyses of climate change. *Review of Environmental Economics and Policy*, 7(2), 181-198.

Auffhammer, M., and Schlenker, W. (2014). Empirical studies on agricultural impacts and adaptation. *Energy Economics*, 46(2014), 555-561.

Burchfield, M., Overman, H. G., Puga, D., and Turner, M. A. (2006). Causes of sprawl: A portrait from space. *Quarterly Journal of Economics*, 121(2), 587-633.

Burke, M., Dykema, J., Lobell, D. B., Miguel, E., and Satyanath, S. (2015). Incorporating climate uncertainty into estimates of climate change impacts. *The Review of Economics and Statistics*, 97(2), 461-471.

Erb, K.-H., Gaube, V., Krausmann, F., Plutzar, C., Bondeau, A., and Haberl, H. (2007). A comprehensive global 5 min resolution land-use data set for the year 2000 consistent with national census data. *Journal of Land Use Science*, 2(3), 191-224.

FAO, 1981. Report of the agro-ecological zones project. World Soil Resources Report No. 48 (1-4). FAO, Rome, Italy.

Fezzi, C. and Bateman I. J. (2011). Structural agricultural land use modeling for spatial agro-environmental policy analysis. *American Journal of Agricultural Economics*, 93(4), 1168-1188.

Fischer, G., Shah, M., van Velthuizen, H., Nachtergaele, F. (2000). Global agro-ecological assessment for agriculture in the 21st century. International Institute for Applied Systems Analysis, Laxenburg, Austria.

GAMS Development Corporation. General Algebraic Modeling System (GAMS) Release 24.2.1. Washington, DC, USA, 2013.

Greene, W. (2010). Testing hypotheses about interaction terms in nonlinear models. *Economics Letters*, 107(2), 291-296.

Gourieroux, C., Monfort, A., and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, 52(3), 681-700.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review*, 102(2), 994-1028.

Hendricks, N. P., Smith, A., and Sumner, D. A. (2014). Crop supply dynamics and the illusion of partial adjustment. *American Journal of Agricultural Economics*, 96(5), 1469-1491.

Hertel, T. W., Rose S., and Tol R. (eds.) (2009). *Economic Analysis of Land Use in Global Climate Change Policy*. Abingdon: Routledge.

Hertel, T. W., and Lobell, D. B. (2014). Agricultural adaptation to climate change in rich and poor countries: Current modeling practice and potential for empirical contributions. *Energy Economics*, 46(2014), 562-575.

IGBP-DIS. SoilData (V.0) a program for creating global soil-property databases, IGBP global soils data task, France, 1998.

IIASA/FAO. Global Agro-ecological Zones (GAEZ v3.0). IIASA, Laxenburg, Austria and FAO, Rome, Italy, 2012.

Klein Goldewijk, K., Beusen, A., and Janssen, P. (2010). Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *The Holocene*, 20(4), 565-573.

Kyle, P., Thomson, A., Wise, M., and Zhang, X. (2015). Assessment of the importance of spatial scale in long-term land use modeling of the Midwestern United States. *Environmental Modelling & Software*, 72(2015), 261-271.

Lapola, D. M., Schaldach, R., Alcamo, J., Bondeau, A., Koch, J., Koelking, C., et al. (2010). Indirect land-use changes can overcome carbon savings from biofuels in Brazil. *Proceedings of the National Academy of Sciences*, 107(8), 3388-3393.

Lerner, B. R., and Dana, M. N. (2001). Growing sweet corn: Department of Horticulture, Purdue University Cooperative Extension Service.

Li, M., De Pinto, A., Ulimwengu, J., You, L., and Robertson, R. (2014). Impacts of road expansion on deforestation and biological carbon loss in the Democratic Republic of Congo. *Environmental and Resource Economics*, 1-37.

Lobell, D. B., Banziger, M., Magorokosho, C., and Vivek, B. (2011). Nonlinear heat effects on African maize as evidenced by historical yield trials. *Nature Climate Change*, 1(1), 42-45.

Lobell, D. B., Burke, M. B., Tebaldi, C., Mastrandrea, M. D., Falcon, W. P., and Naylor, R. L. (2008). Prioritizing climate change adaptation needs for food security in 2030. *Science*, 319(5863), 607-610.

Lobell, D. B., Hammer, G. L., McLean, G., Messina, C., Roberts, M. J., and Schlenker, W. (2013). The critical role of extreme heat for maize production in the United States. *Nature Climate Change*, 3(5), 497-501.

Mallarino, A. P., Pagani, A., and Sawyer, J. E. (2011). Corn and soybean response to soil pH level and liming. Paper presented at the 2011 Integrated Crop Management Conference, Iowa State University.

Mallory, M., Hayes, D. J., and Babcock, B. A. (2011). Crop-based biofuel production with acreage competition and uncertainty. *Land Economics*, 87(4): 610-627.

Mendelsohn, R. (2009). The impact of climate change on agriculture in developing countries. *Journal of Natural Resources Policy Research*, 1(1), 5-19.

Monfreda, C., Ramankutty, N., and Foley, J. A. (2008). Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, 22(1), GB1022.

Mueller, V., Quisumbing, A., Lee, H. L., and Droppelmann, K. (2014). Resettlement for food security's sake: Insights from a Malawi land reform project. *Land Economics*, 90(2), 222-236.

Mullahy, J. (2015). Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods*, 4(1), 71-100.

National Geophysical Data Center/NESDIS/NOAA/U.S. Department of Commerce. 1995. TerrainBase, global 5 arc-minute ocean depth and land elevation from the U.S. National Geophysical Data Center (NGDC). Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds759.2/. Accessed on May 25, 2015.

New, M., Hulme, M., and Jones, P. (1999). Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, 12(3), 829-856.

Papke, L. E., and Wooldridge, J. M. (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 11, 619-632.

Papke, L. E., and Wooldridge, J. M. (2008). Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics*, 145, 121-133.

Polasky, S., Costello, C., and McAusland, C. (2004). On trade, land-use, and biodiversity. *Journal of Environmental Economics and Management*, 48(2), 911-925.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ramankutty, N., and Foley, J. A. (1998). Characterizing patterns of global land use: An analysis of global croplands data. *Global Biogeochemical Cycles*, 12(4), 667-685.

Ramankutty, N., and Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles*, 13(4), 997-1027.

Ramankutty, N., Evan, A. T., Monfreda, C., and Foley, J. A. (2008). Farming the planet: 1. Geographic distribution of global agricultural lands in the year 2000. *Global Biogeochemical Cycles*, 22(1), GB1003.

Rosegrant, M. W., Cai, X., and Cline, S. A. (2002). *World Water and Food to 2025: Dealing with Scarcity*. Washington, DC and Battaramulla, Sri Lanka: International Food Policy Research Institute (IFPRI).

Schaldach, R., Alcamo, J., Koch, J., Kölking, C., Lapola, D. M., Schüngel, J., et al. (2011). An integrated approach to modelling land-use change on continental and global scales. *Environmental Modelling and Software*, 26(8), 1041-1051.

Schlenker, W., Hanemann, W. M., and Fisher, A. C. (2006). The impact of global warming on U.S. agriculture: An econometric analysis of optimal growing conditions. *The Review of Economics and Statistics*, 88(1), 113-125.

Schlenker, W., and Lobell, D. B. (2010). Robust negative impacts of climate change on African agriculture. *Environmental Research Letters*, 5(1), 014010.

Schlenker, W., and Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. *Proceedings of the National Academy of Sciences*, 106(37), 15594-15598.

Villoria, N. B., and Liu, J. (2015). Using continental grids to improve our understanding of global land supply responses and land use change. Unpublished manuscript.

Wooldridge, J. M. (1991). Specification testing and quasi-maximum likelihood estimation. *Journal of Econometrics*, 48(1–2), 29-55.

Wooldridge, J. M. (1997). Quasi-likelihood methods for count data. M. Pesaran, P. Schmidt (Eds.), *Handbook of Applied Econometrics*, Volume II: Microeconometrics, Blackwell Publishers Ltd., Malden, MA.

You, L., and Wood, S. (2006). An entropy approach to spatial disaggregation of agricultural production. *Agricultural Systems*, 90(1–3), 329-347.

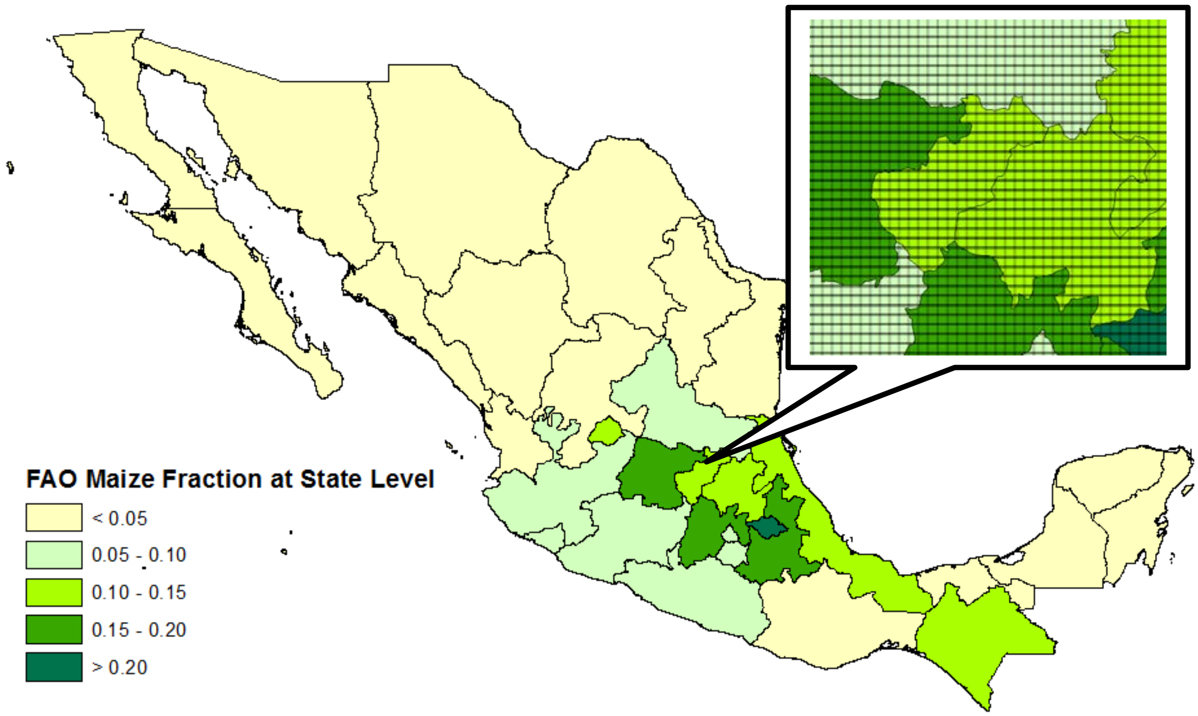

Figure 1. Observed FAO maize fraction data for Mexico at Administrative Unit 1. The inset image illustrates the pixelated land grid.

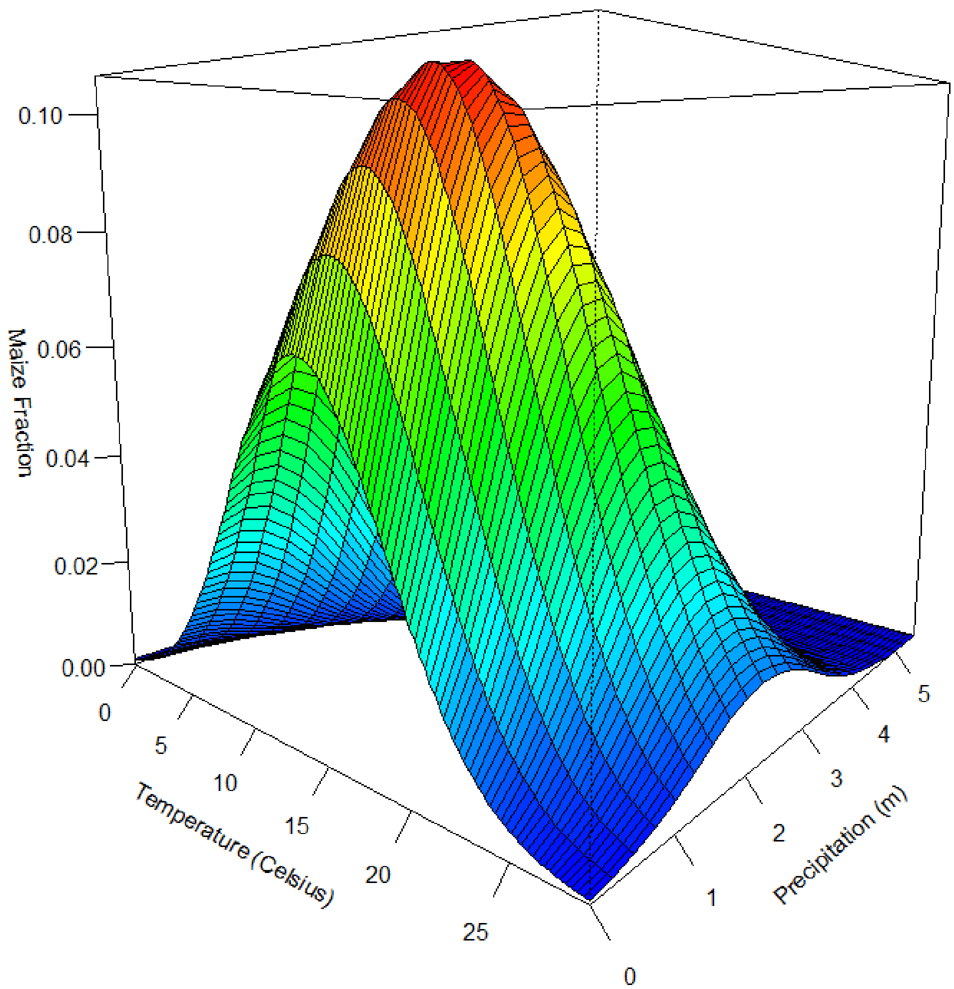

Figure 2(a). Estimated relationship between temperature, precipitation, and the fraction of maize for the single-crop maize model.

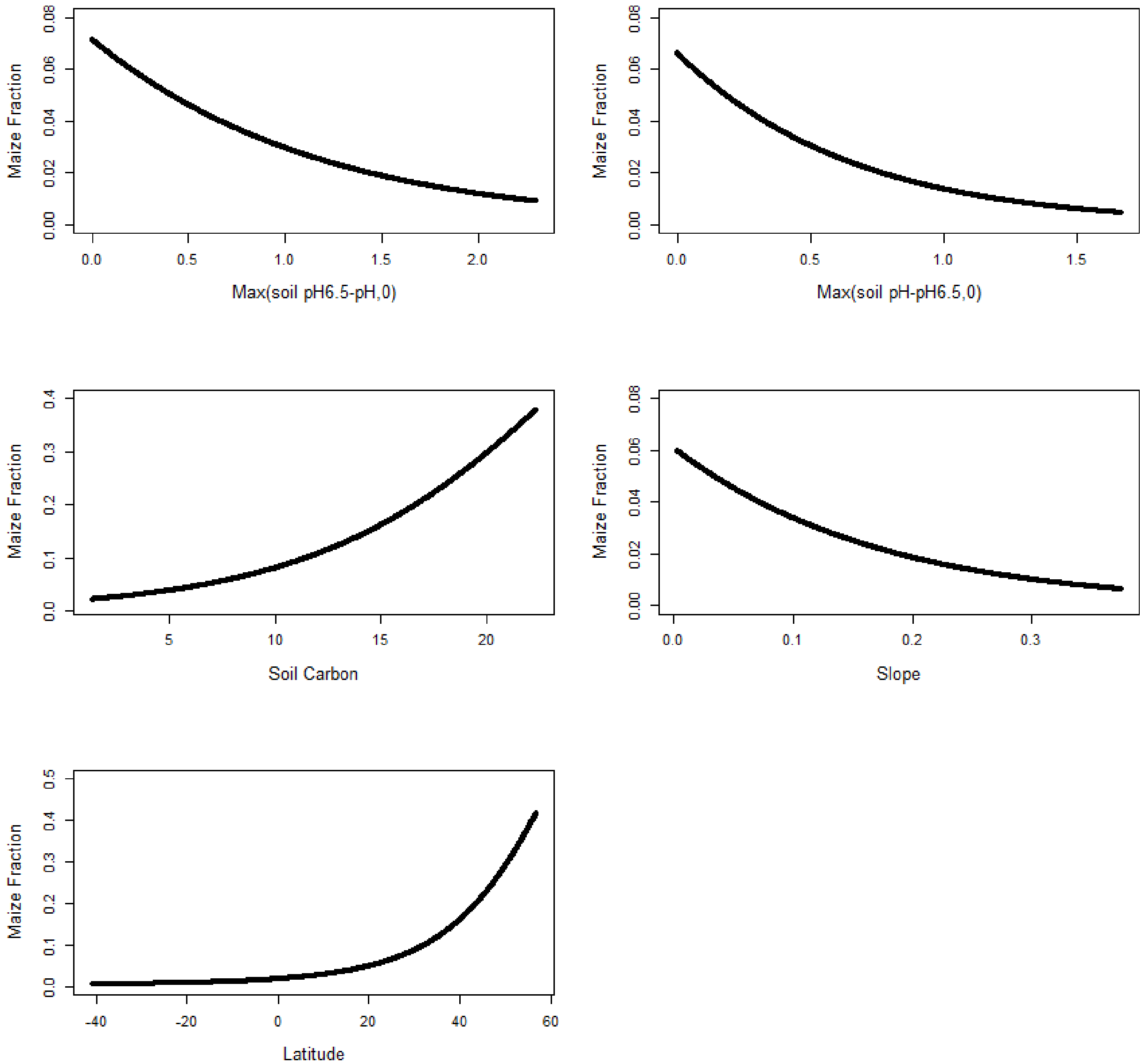

Figure 2(b). Estimated relationship between elevation, soil pH, soil carbon, slope, latitude and the fraction of maize for the single-crop maize model.

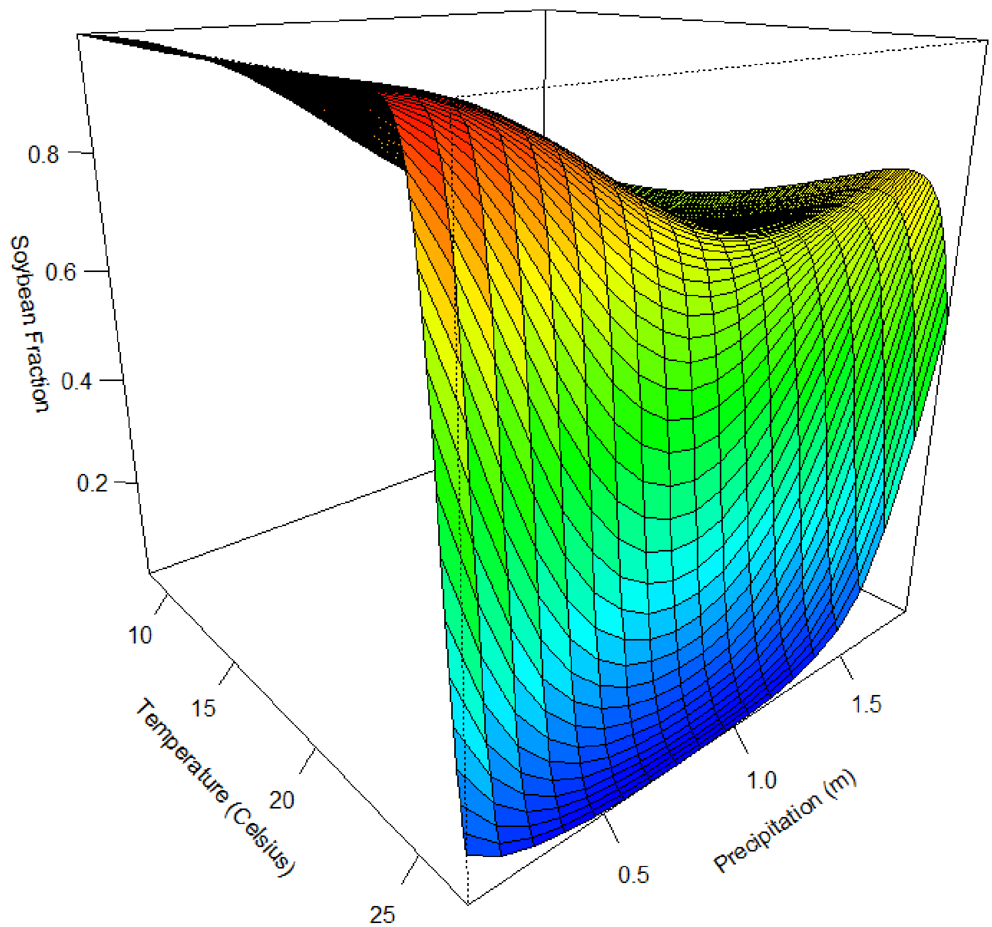

Figure 3(a). Estimated relationship between temperature, precipitation, and the soybean fraction for the multi-crop model.

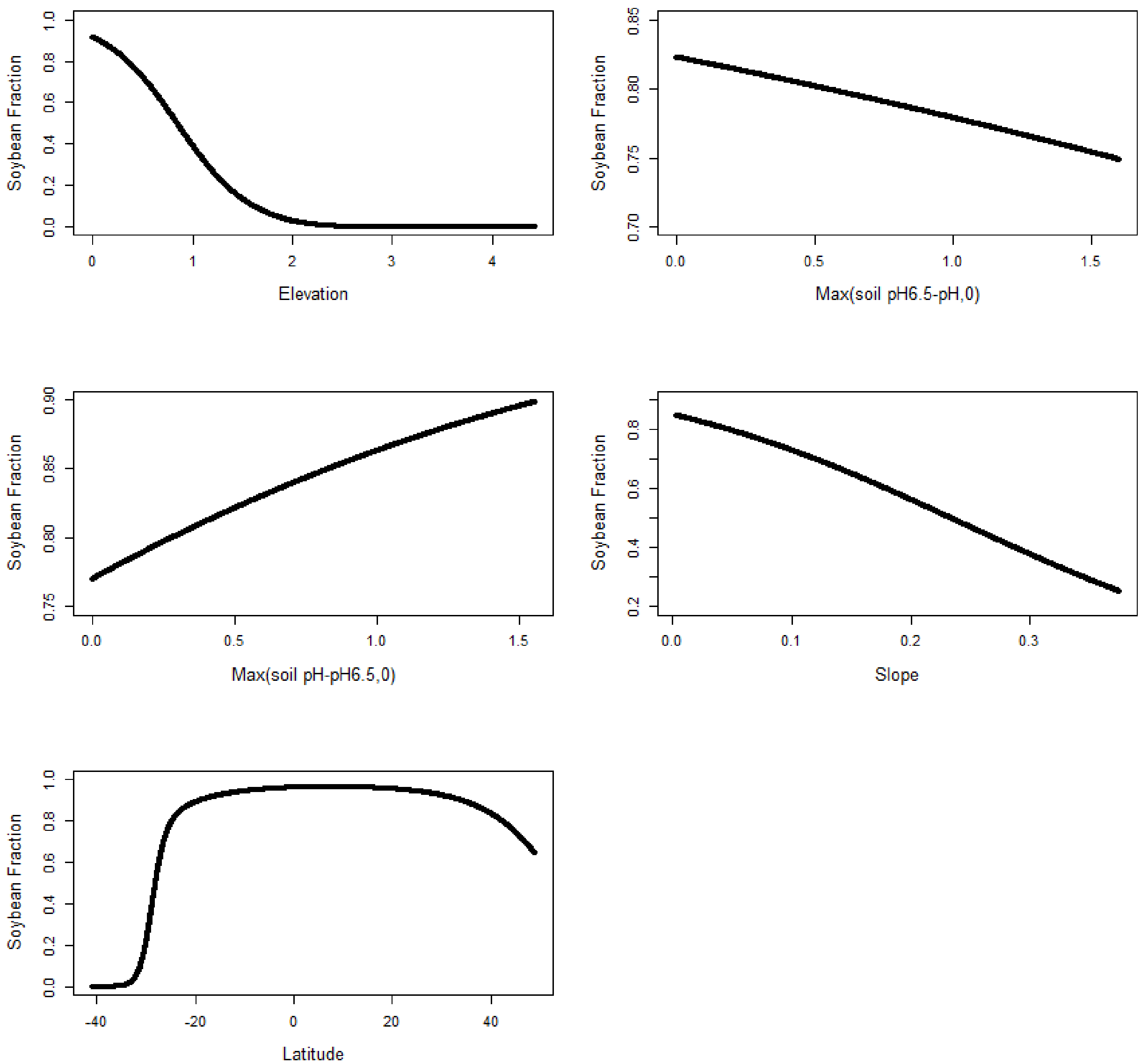

Figure 3(b). Estimated relationship between elevation, soil pH, slope, latitude, and the soybean fraction for the multi-crop model.

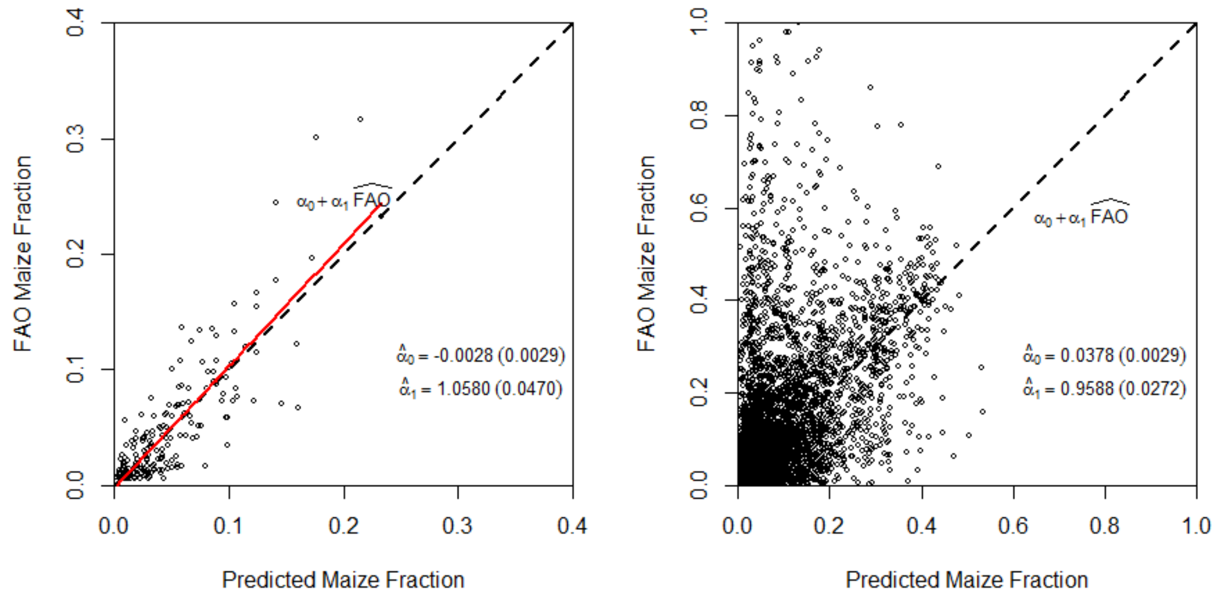

Figure 4. Comparison between the predicted maize area fraction versus observed FAO maize area fraction at Administrative Unit Levels 1 (left) and 2 (right).

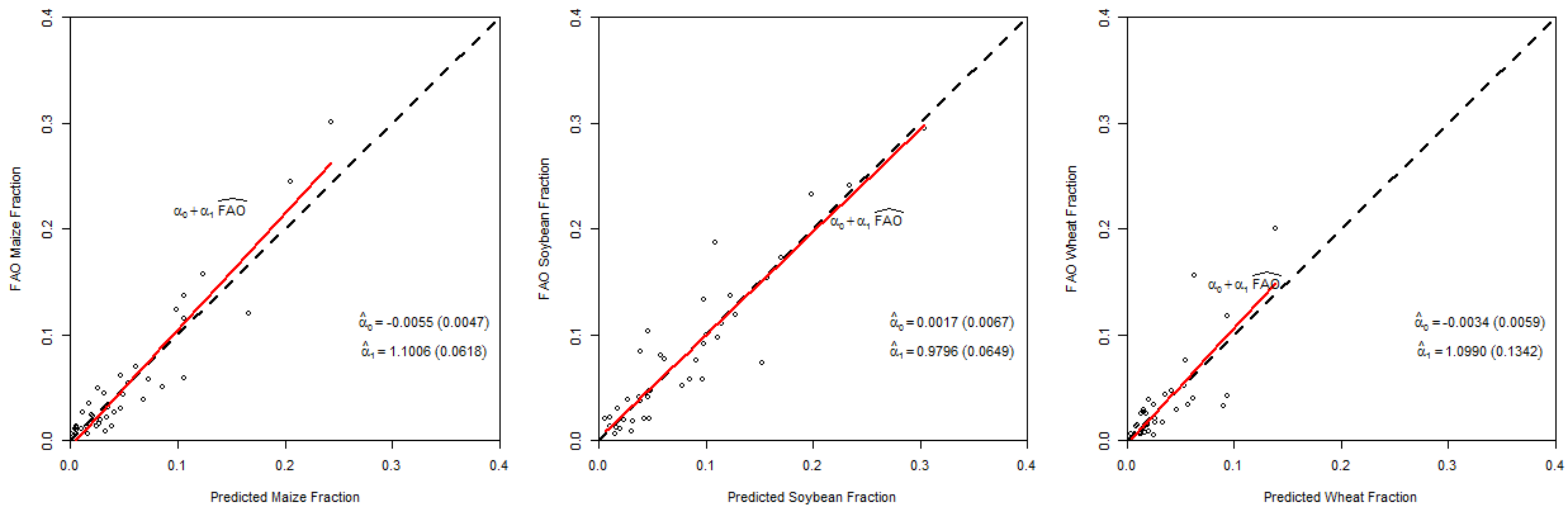

Figure 5(a). Comparison between the predicted area fraction versus the observed FAO area fraction for maize, soybeans, and wheat at Administrative Unit Level 1 for the multi-crop model.

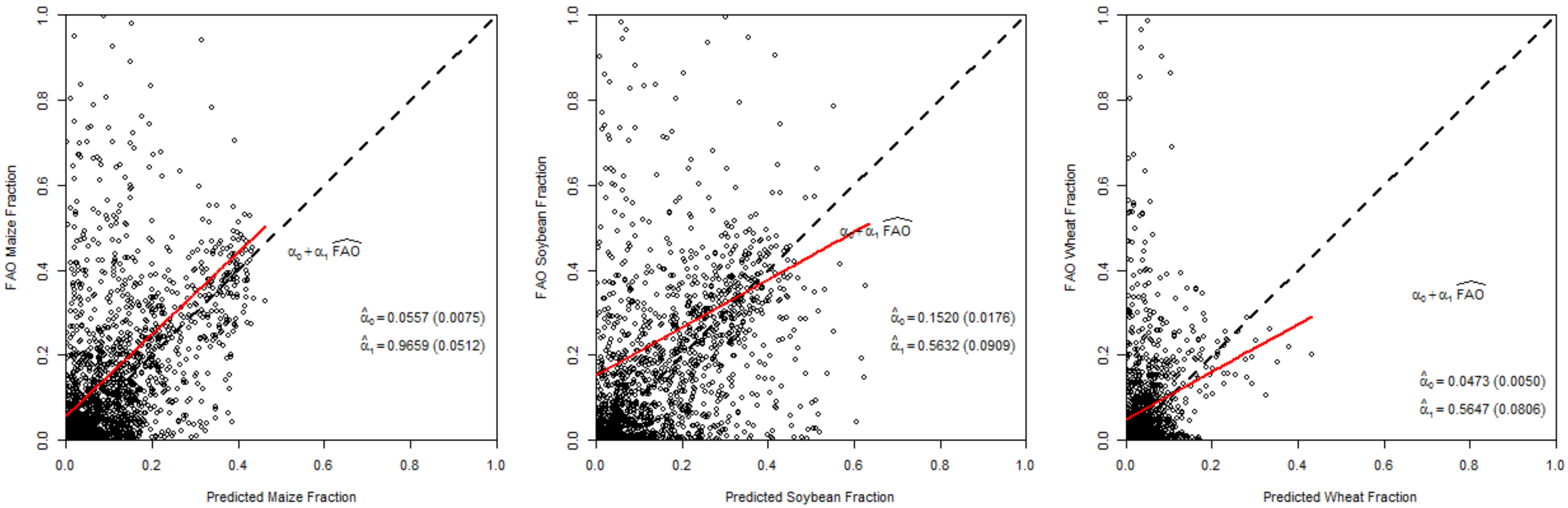

Figure 5(b). Comparison between the predicted area fraction versus the observed FAO area fraction comparison for maize, soybeans, and wheat at Administrative Unit Level 2 for the multi-crop model.

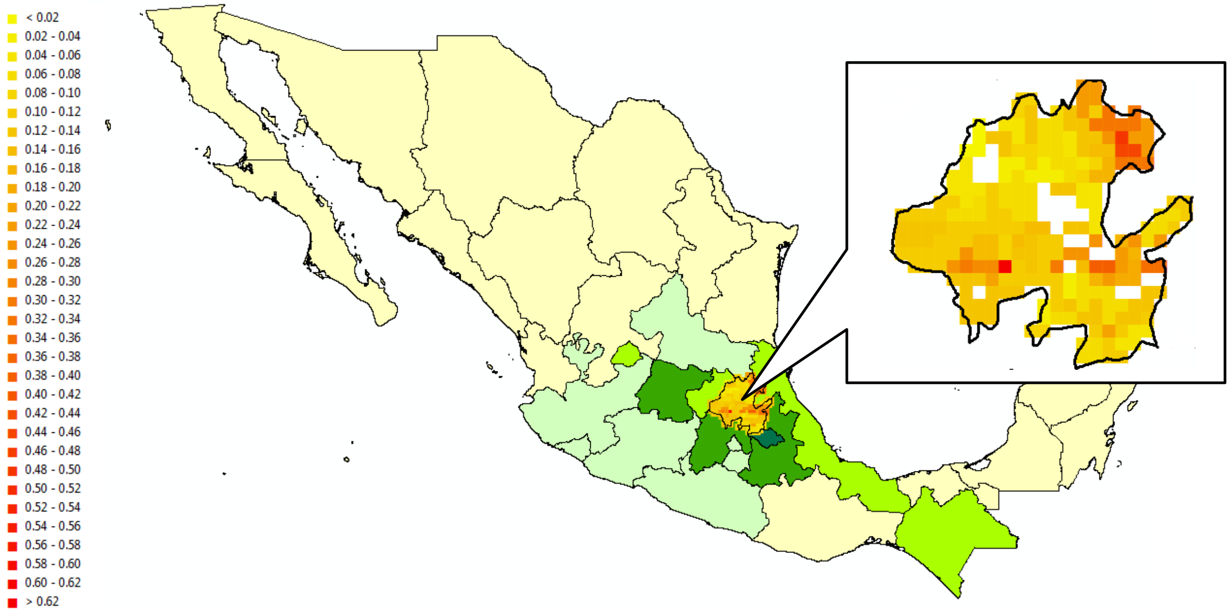

Figure 6(a). Pixel level maize fraction predictions for Hidalgo, Mexico using constant FAO Level 2 shares for all pixels within each Level 2 area.

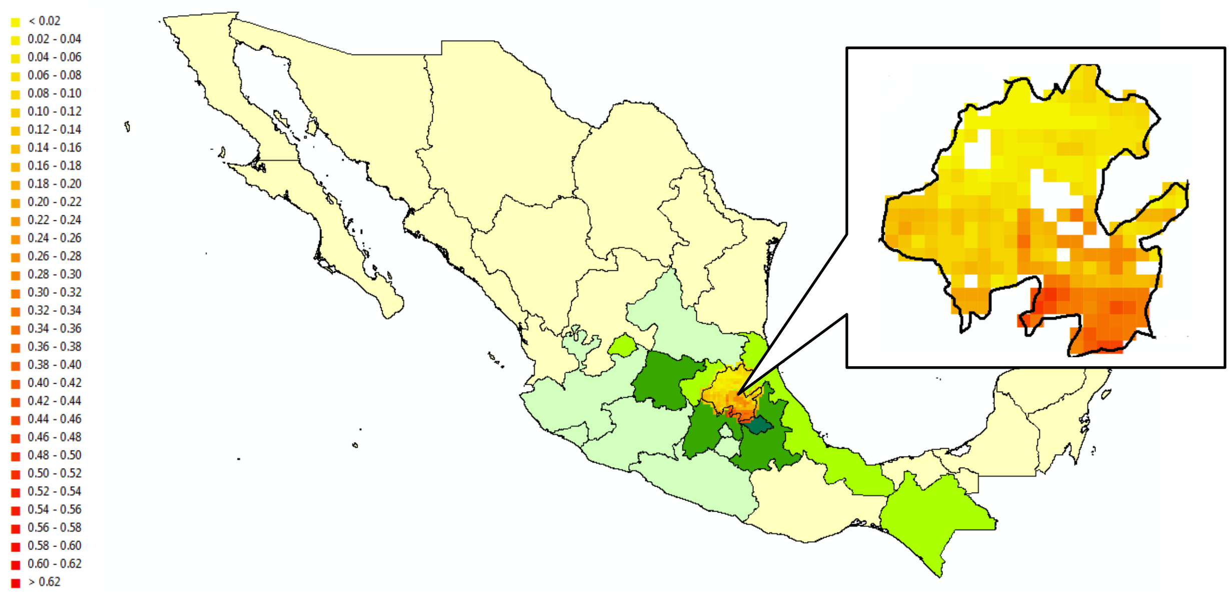

Figure 6(b). Pixel level maize fraction predictions for Hidalgo, Mexico using the estimated pixel-specific maize shares.

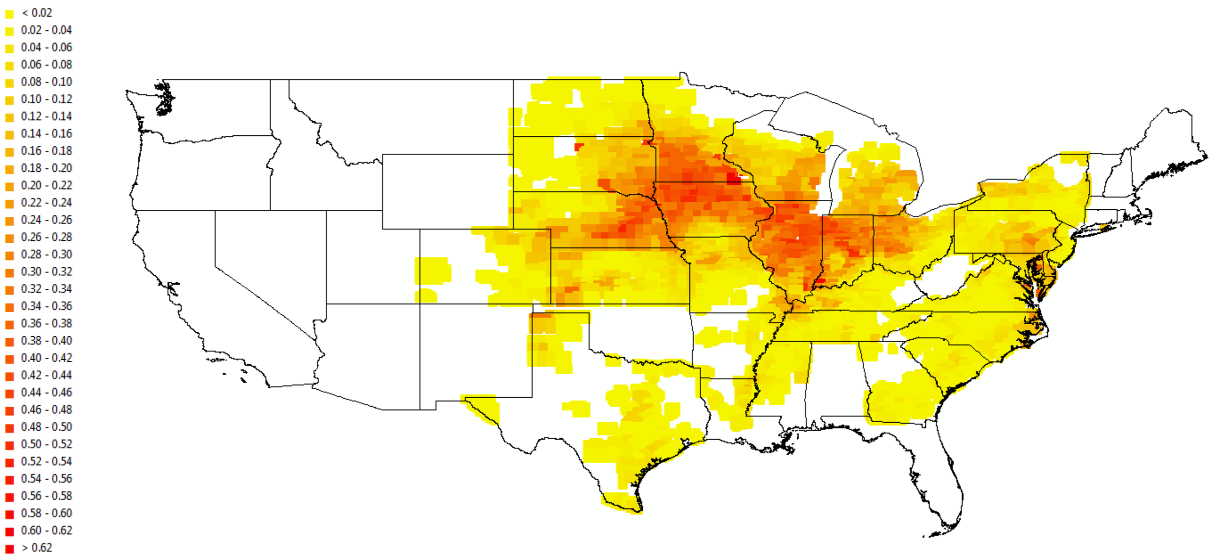

Figure 7(a). Pixel level maize fraction predictions for the United States using constant FAO Level 2 shares for all pixels within each Level 2 area.

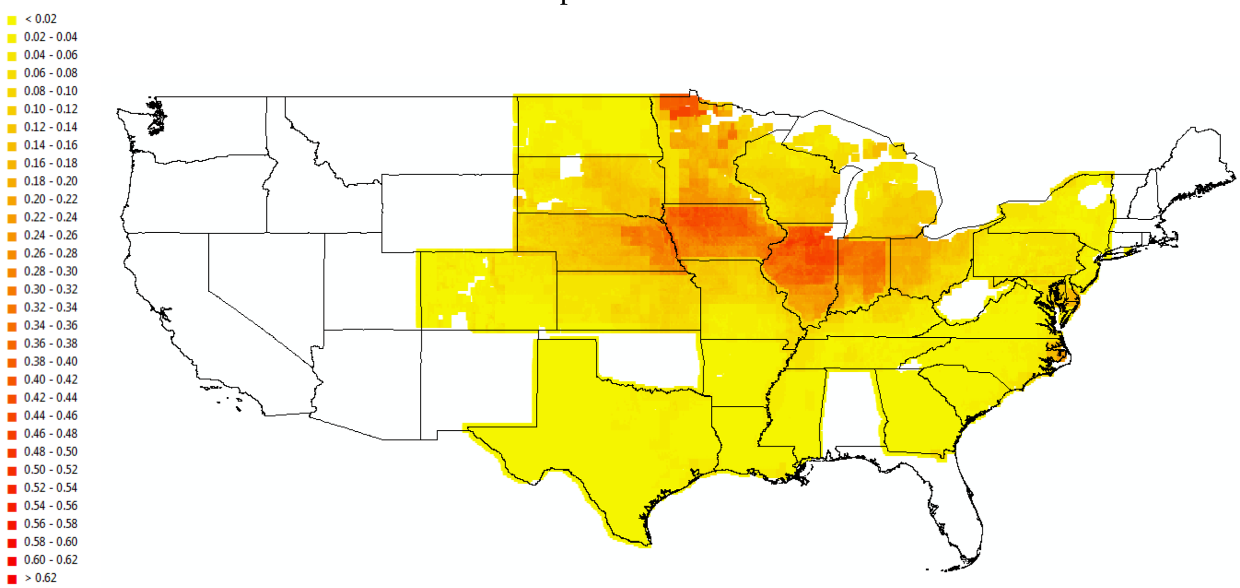

Figure 7(b). Pixel level maize fraction predictions for the United States using the estimated pixel-specific maize shares.

Table 1. Descriptive statistics of the biophysical variables measured at the pixel level

	Variable	Mean	Standard Deviation	Min	Max
North America	Temperature	16.6598	6.0868	1.8333	28.6667
	Precipitation	0.8131	0.3612	0.0680	3.2580
	Elevation	0.5685	0.6012	-0.2270	3.7040
	Max(pH6.5-pH,0)	0.4662	0.6172	0.0000	2.3000
	Max(pH-pH6.5,0)	0.3423	0.4661	0.0000	1.6650
	Soil Carbon	6.3510	2.8951	1.6080	22.3560
	Slope	0.0562	0.0649	0.0025	0.3750
	Latitude	37.3264	9.6123	14.6250	56.7917
Central America	Temperature	23.4744	2.5139	15.8333	27.5000
	Precipitation	2.2560	0.7393	1.1420	4.6780
	Elevation	0.6743	0.6086	-0.1970	3.3000
	Max(pH6.5-pH,0)	0.6111	0.4224	0.0000	1.4000
	Max(pH-pH6.5,0)	0.0304	0.1011	0.0000	0.6400
	Soil Carbon	6.8998	1.7396	3.9840	12.7240
	Slope	0.1340	0.0876	0.0025	0.3750
	Latitude	13.2632	2.1562	7.2917	16.0417
South America	Temperature	22.4933	3.9667	-0.1667	28.5000
	Precipitation	1.1810	0.4937	0.0000	5.6670
	Elevation	0.5644	0.7089	-0.1320	4.7830
	Max(pH6.5-pH,0)	0.6903	0.6087	0.0000	1.8690
	Max(pH-pH6.5,0)	0.1940	0.3918	0.0000	1.5630
	Soil Carbon	5.1467	1.6458	1.3250	13.1830
	Slope	0.0535	0.0647	0.0025	0.3750
	Latitude	-18.5371	11.9075	-40.9583	11.4583

Note: We follow the CIA World Factbook (https://www.cia.gov/library/publications/the-world-factbook/) division of the Americas. Our data includes the North American countries Canada, United States and Mexico; the Central American countries Costa Rica, Guatemala, Honduras, Nicaragua and Panama; the South American countries Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Paraguay, Peru, Uruguay and Venezuela.

Table 2. List of countries in each model with the number of administrative units in the sample

Maize Model			
Argentina (9)	Bolivia (3)	Brazil (18)	Canada (1)
Chile (3)	Colombia (6)	Costa Rica (2)	Ecuador (10)
Guatemala (12)	Honduras (14)	Mexico (29)	Nicaragua (17)
Panama (5)	Peru (8)	Paraguay (13)	Uruguay (5)
United States (28)	Venezuela (13)		
Multi-crop Model			
Argentina (8)	Brazil (3)	Mexico (2)	Paraguay (5)
United States (22)			

Note: In the maize model, we include only the states in which there is at least 0.5 percent cropland in maize. In the multi-crop model, we require each state to have at least 0.5 percent cropland in each crop. The numbers indicated in parentheses after the country names are the number of states included for the country.

Table 3. Quasi-maximum likelihood estimates and standard errors for the maize model

Variable	Coefficient	Standard Error
Intercept	-7.1552***	1.9488
Temperature	0.5818***	0.1575
Temperature Squared	-0.0212***	0.0041
Precipitation	-0.5780	0.7133
Precipitation Squared	-0.5841***	0.1407
Temperature·Precipitation	0.1142***	0.0278
Elevation	-0.0623	0.2810
Max(pH6.5-pH,0)	-0.9227***	0.2031
Max(pH-pH6.5,0)	-1.6218***	0.3832
Soil Carbon	0.1566***	0.0467
Slope	-6.1423**	2.2885
Latitude	0.0442**	0.0148
Latitude Squared	0.0004	0.0005
Argentina Indicator	0.3924	0.3001
Bolivia Indicator	0.6112	0.5155
Brazil Indicator	1.4320***	0.3004
Canada Indicator	-3.9650***	1.2771
Chile Indicator	0.3771	0.5265
Colombia Indicator	-0.3332	0.3274
Costa Rica Indicator	-0.7427*	0.4387
Guatemala Indicator	1.1551*	0.5781
Honduras Indicator	0.1313	0.5390
Mexico Indicator	0.4500	0.6413
Nicaragua Indicator	0.7408	0.4616
Panama Indicator	0.0518	0.4615
United States Indicator	-1.8866	1.1502
Venezuela Indicator	0.4996	0.4798

Note: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 4. Implied marginal effects and odds ratios for the single-crop maize model

Variable	Marginal Effect	Odds Ratio
Temperature	0.0035	0.8900
Precipitation	0.0103	1.1195
Elevation	-0.0053	0.9396
Max(pH6.5-pH, 0)	-0.0787	0.3974
Max(pH-pH6.5, 0)	-0.1383	0.1976
Soil Carbon	0.0134	1.1695
Slope	-0.5240	0.0022
Latitude	0.1055	2.5666

Note: The marginal effects and odds ratios are averages over all included pixels.

Table 5. Quasi-maximum likelihood estimates and standard errors for the multi-crop model

Variable	Maize	Soybeans	Wheat
Intercept	9.6632	18.7251*	5.8832
	(11.0798)	(8.9678)	(14.4399)
Temperature	2.8774***	2.7152***	1.0892
	(0.8463)	(0.6468)	(0.7522)
Temperature Squared	-0.1183***	-0.1135***	-0.0575*
	(0.0317)	(0.0263)	(0.0313)
Precipitation	-20.7500**	-26.1529***	-15.3136
	(8.6990)	(7.1030)	(9.5744)
Precipitation Squared	2.8250	4.4937**	0.5055
	(2.0965)	(1.6356)	(2.0077)
Temperature·Precipitation	0.6872*	0.7069	0.6722*
	(0.4006)	(0.3521)	(0.3946)
Elevation	-4.4991**	-7.2887***	-1.3311
	(1.6392)	(1.5197)	(2.7430)
Max(pH6.5-pH,0)	-1.9599***	-2.2358 ***	-0.6687*
	(0.3199)	(0.2599)	(0.3482)
Max(pH-pH6.5,0)	-3.4518**	-2.8158**	-0.2465
	(1.1302)	(1.1435)	(0.7782)
Soil Carbon	0.1437	0.0761	0.1245
	(0.1431)	(0.1545)	(0.1322)
Slope	-16.3140*	-23.6936**	-26.5373*
	(8.6467)	(9.8526)	(14.6707)
Latitude	0.3169***	0.3403***	0.0715
	(0.0601)	(0.0413)	(0.0594)
Latitude Squared	-0.0067**	-0.0083***	-0.0031
	(0.0024)	(0.0023)	(0.0043)
Mexico Indicator	-1.2862***	-1.6414***	-0.7098
	(3.1985)	(2.4062)	(3.3592)
United States Indicator	-19.8621***	-21.2622***	-4.6801
	(4.0114)	(2.7395)	(3.7104)

Note: $*p < 0.05$; $**p < 0.01$; $***p < 0.001$

Table 6. Implied marginal effects and odds ratios from the multi-crop model for maize, soybeans, and wheat

Variable	Maize		Soybeans		Wheat	
	Marginal Effect	Odds Ratio	Marginal Effect	Odds Ratio	Marginal Effect	Odds Ratio
Temperature	-0.0119	0.7330	-0.0248	0.7585	-0.0051	1.1831
Precipitation	0.2041	3.9022	-0.2257	32.8382	-0.0134	2.7792
Elevation	0.2219	10.0657	-0.4400	0.0382	-0.0024	70.6299
Max(pH6.5-pH, 0)	0.0066	0.8911	-0.0788	0.5189	-0.0046	3.0646
Max(pH-pH6.5, 0)	-0.0968	0.3468	-0.0061	1.1989	0.0050	9.5469
Soil Carbon	0.0079	1.0947	-0.0049	0.9820	0.0018	1.0646
Slope	0.5607	18.7818	-1.2198	0.0002	-0.3749	0.0001
Latitude	0.0163	1.3702	0.0107	1.3750	0.0021	1.1277

Note: The marginal effects and odds ratios are averages over all included pixels.

Table 7. RMSE values for the maize model and the multi-crop model at Levels 1 and 2

	Level 1	Level 2
Maize Model		
Maize	0.0270	0.1902
Multi-crop Model		
Maize	0.0215	0.2554
Soybeans	0.0261	0.5205
Wheat	0.0242	0.1516

Table 8. Correlations between different model estimated pixel specific maize fractions

	Illinois	Indiana	Iowa	North Dakota
Maize Model vs. CDL	0.6317	0.6492	0.6303	0.1721
Multi-crop Model vs. CDL	0.6051	0.6013	NA	0.3555
Maize Model vs. Monfreda et al.	0.7590	0.6716	0.6101	0.2892
Multi-crop Model vs. Monfreda et al.	0.7524	0.6277	NA	0.5733
CDL vs. Monfreda et al.	0.7792	0.8459	0.7841	0.7277

Note: Iowa is not included in our multi-crop model due to an insufficient land area for wheat.

Appendix

Table A.1 Definition of variable measurement and data source

Name	Description	Source
Temperature	Average monthly temperature in degrees Celsius over the period 1961-1990; For countries in the Northern Hemisphere the growing season is March through August, whereas the growing season for countries in the Southern Hemisphere is September through February.	New et al. (1999) http://www.sage.wisc.edu/atlas/index.php, accessed May 25, 2015
Precipitation	Average annual total precipitation in meters/year over the period 1961-1990.	New et al. (1999) Same web link and access date as above
Elevation	Meters above sea level on a 5-minute resolution.	United States National Geophysical Data Center TerrainBase global model of terrain and bathymetry (1995) Same web link and access date as above
Soil pH	Soil pH (0-14).	SoilData System, Global Soils Data Task, International Geosphere-Biosphere Program (IGBP-DIS) (1998) Same web link and access date as above
Soil Carbon	Soil organic carbon density in kg per square meter, 0 to 1 meter depth.	Same source, web link and access date as above
Slope	Eight categories of median terrain slopes: 0-0.5%, 0.5-2%, 2-5%, 5-8%, 8-16%, 16-30%, 30-45% and > 45%. We use the median slopes of the IIASA/FAO slope categories as our slope variable values.	IIASA/FAO (2012)
Built-up Land	Combination of modeled built-up areas based on nighttime lights and observed built-up area based on IGBP land cover data.	https://nelson.wisc.edu/sage/data-and-models/atlas/maps.php?datasetid=18&includerelatedlinks=1&dataset=18, accessed Mar 04, 2016
Protected Areas	Global raster data layer with a resolution of 5 arc-minutes. Each pixel is classified as protected area where agriculture should not be occurring, protected area where agriculture could be occurring, or non-protected area.	http://www.fao.org/geonetwork/srv/en/main.home, accessed Mar 04, 2015
Total land area from Statoids	Total land area in an administrative unit.	http://www.statoids.com/, accessed Mar 05, 2016
Harvested land area	Total areas of land harvested in maize, soybeans, and wheat at Administrative Unit Level 1.	http://kids.fao.org/agromaps/, retrieved Feb 20, 2015
CIA World Factbook	Provides information on the government, geography, etc. for 267 world entities.	https://www.cia.gov/library/publications/the-world-factbook/, retrieved Feb 23, 2015
GADM database	Spatial database on the location of the world's administrative area.	http://gadm.org/, retrieved Feb 23, 2015
USDA NASS Quick Stats	United States Department of Agriculture National Agricultural Statistics Service census and survey data.	http://quickstats.nass.usda.gov/#528F56BC-9FFB-3942-B141-CA0EBDC414C9, retrieved May 16, 2015
USDA Cropland Data Layer	USDA National Agricultural Statistics Service Cropland Data Layer.	https://nassgeodata.gmu.edu/CropScape/, accessed Jan 19, 2016