US Corn Belt: a satellite view

Matthieu Stigler

August 2019

The US Corn Belt is a major producer of corn and soybeans, which are often cultivated in rotation. In this chapter, I show how satellite methods can be used to gain several insights into this production system. Taking advantage of recent developments in yield estimates from satellite data, I build a dataset of field-level yields for ten years and close to two million fields in nine states. The dataset is then complemented with weather information, as well as prices.

This field-level dataset helps document several stylised facts that have received less attention. First, I show how a large proportion of fields always follows a corn-soybean rotation, independent of any price or weather changes. Second, I document a positive correlation between a field's potential for corn and for soybean yields. This fact goes well in hand with the previous fact that many fields follow rotation patterns. Finally, I show that there is a tendency to plant more often corn on fields with higher fertility, while low fertility soils tend to be planted to soybeans.

1 Introduction

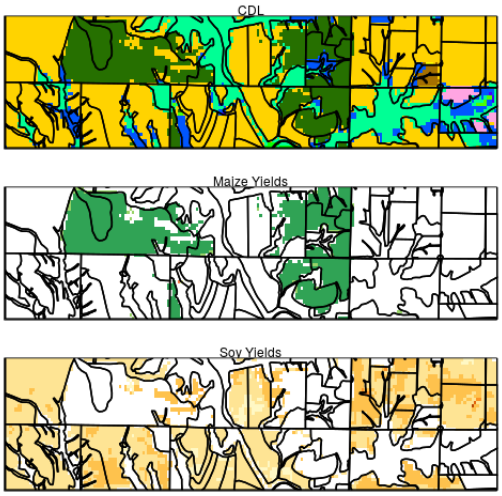
The aim of this chapter is to describe the dataset assembled, and to conduct an exploratory analysis, highlighting several stylised facts. In Section 2, I describe the different data sources, crop maps, yield data, field boundaries as well as weather and price data. Both the crop maps and the yield data are at the pixel scale, but we are ultimately interested in an analysis at the field level. To do so, I use a field boundary dataset, the Common Land Unit (CLU). Several important data processing decisions are made at that step. First of all, only a subset of the fields needs to be kept, as the CLU contains a very large number of units, sometimes delineating roads or forests. Second, the crop classification for these selected fields might be imprecise, with multiples crops predicted in the same field. This calls for a second filtering, where only fields with a high *classification agreement* are kept. Once the field selection step is done, the yield data is averaged over each field, removing boundary pixels which might be contaminated by extraneous signal from roads, farms etc. I describe then how the price data is constructed, as well as how weather variable are assembled.

In Section 3, I proceed to an exploratory analysis of the dataset, which serves as a starting basis for the next chapters. I show several interesting stylised facts, which, to the best of my knowledge, have received less attention in the literature.

2 Data description

To conduct an analysis at the field-level, I assemble data from three main sources: a crop classification at the pixel level, a yield map for the corresponding corn and soybeans pixels, and a field boundary dataset. Figure 1 illustrates the three datasets combined. The first panel shows the crop classification, together with the field boundaries. The second panel shows the yield predictions for the pixels for which the CDL predicts maize. The third panel shows the soybeans yield predictions.

Figure 1: Illustration of the CDL, yield and boundary data



2.1 Crop classification

The crop data comes from the USDA Crop Data Layer (CDL) dataset (Boryan et al., 2011). The CDL classifies Landsat pixels of 30m$\times$ 30m into a large number of classes. Corn and soybeans appear in multiple distinct classes beyond the simple corn and soybeans classes, such as double-crop categories like "Winter Wheat and Corn" or "Soybeans and Cotton". Due to the small share of the alternative classes, I focus here only on the corn and soybeans only categories. The accuracy of the classification for maize and soybeans in the Corn Belt is very high, in general above 95%[1]. It is not clear how this accuracy measure is validated, and practical experience shows that a fair number of pixels appear to miss-classified. A typical miss-classification, known in the remote sensing literature under the *salt and pepper* effect, arises when say a single corn pixel is found in a soy field. This *salt and pepper* effect can be seen in the first panel of Figure 1, where a few corn pixels are found in a field otherwise planted to soybeans. I deal with this issue choosing the mode of the pixel classification for each field, removing fields where the percentage of the mode is too low. See Section 2.3 for more details.

2.2 Yield data

The yield predictions are based on the scalable satellite-based crop yield (SCYM) method of Lobell et al. (2015) and Jin et al. (2017). The method predict yields based on the Green Chlorophyll Vegetation index (GCVI). This index, similar in spirit to the more widely known normalised difference index (NDVI), is based on the near-infrared (NIR) and green (G) bands, $GCVI = NIR/G - 1$. The model is a simple linear regression, including weather covariates W.

$$yield_{it} = \alpha + \beta GCVI_{it} + W\Gamma + \varepsilon_{it} \tag{1}$$

The parameters are obtained using a pseudo training sample obtained from a crop growth model. In a nutshell, the APSIM crop growth model is calibrated using values representing cultivars and farming practices typical in the US Midwest. Yields and leaf area index (LAI) variables are simulated from the model, and the LAI values are converted to GCVI values. The resulting output is a set of yields, GCVI and weather simulations. This pseudo training sample is used to estimate Equation (1). Once estimated, the equation is used to predict yield, using the satellite-based GCVI values obtained from the Landsat satellite, which has a 30x30 m resolution. Further refinements were suggested in Jin

[1]See `https://www.nass.usda.gov/Research_and_Science/Cropland/metadata/meta.php`

et al. (2017), which increase the accuracy of the estimates. This newer version, which I refer henceforth to as SCYM v2, is however only for corn, yet extends the sample to six more states compared to the 3I states (Iowa, Indiana and Illinois) in the version 1 (SCYM v1).

Both the soybeans and corn models were rerun to add two more years, 2016 and 2017. The soy estimates from SCYM v1 and were initially for the 3I states only but were extended to the six more states used in the Jin et al. (2017) version, i.e. Michigan, Minnesota, Ohio, South Dakota and Wisconsin.

2.3 Field boundaries

An issue with the CDL data is that the analysis is done at the pixel-level, while we are interested in field-level analysis. There exists however a dataset of fields boundary, the USDA Common Land Unit (CLU).[2] Unfortunately, the actual dataset is not publicly available, so that only a copy of the 2009 version can be used.

Two issues arise when using this dataset. Firstly, as the data is from 2009, fields boundaries may have changed. Drastic changes are unlikely, but cultivation of two different crops in the same field is possible. The second issue is that given that the CDL analysis is at the pixel level, instead of being at the field level, pixels in a field can contain multiple crop classes. Preliminary investigations showed clear cases of border contamination, where pixels at the edge of the field were attributed other classes (in particular classes corresponding to bush/forest elements).

These two issues call for specific rules for the attribution of a crop to a given field. Hendricks et al. (2014) used a *centroid-offset* rule, where the field's class is attributed according to the class of the pixel that lies at a certain distance of the field's centroid. This procedure suffers from two issues: firstly, it is not guaranteed that the centroid offset falls within the field itself. Second, if there are really two crops cultivated in one field, the method will attribute only one class. Arguably, the arbitrariness of the offset rule should guarantee that there is no bias in which class will be chosen.

Another method is followed by Stevens (2015), who selects the mode of the classes found within a field. This method does not suffer from the issue of the centroid falling outside the field, yet also can attribute one class whenever there are actually two. To avoid this issue, I focus on fields with a relatively high classification agreement, i.e. I set a minimum threshold on the frequency of the mode. Further, I only take into account for this calculation interior pixels, i.e. pixels that do not touch the border of the field. This avoids to consider mixed pixels, that are potentially contaminated by elements outside of the field.

Figure 2 shows the frequency of the mode, with either all pixels taken into account, or only the interior ones. This is made for all plots in the nine states, using CDL classification for 2015. It is interesting to see that although the field boundaries were made in 2009, there is still a relatively good agreement for the year 2015. One can see that taking only interior pixels instead of all pixels leads to a much better result: a much larger proportion of the fields have 90% or more of the pixels showing the same value. One see furthermore a few bumps around the value of 50%, 66% and 75%. This suggests that the field was planted to two distinct crops (or more), using either a 1/2, 1/3 or 1/4 proportion.

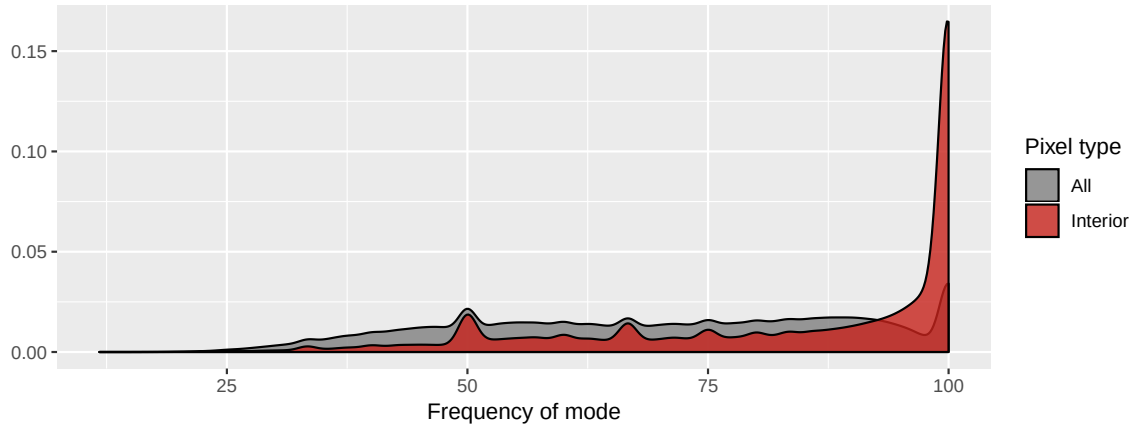
To retain only fields with a good classification accuracy, the threshold was set at a minimum of 85% over all years considered (2008-2015). This is arguably a rather strict value, but it ensures that the data considered is how high quality. This is particularly important for the yield data, for which we want to make sure that we are not averaging over contaminated pixels, which can have a drastic effect on the final yield estimate. In fact, using the high-quality classification fields only, the correlation between NASS county yields and averages from the SCYM dataset is improved.

2.4 Weather data

Weather variables are introduced as control variables to avoid omitted variable bias. While it is reasonable to think that weather is not influenced by prices, it is still the case that prices might anticipate

[2] `https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-products/common-land-unit-clu/index`

Figure 2: Agreement of pixel classification, over all fields, 2015 data



weather events later in the season. To prevent this, I include a large set of weather controls from the DAYMET dataset (Thornton et al., 2017), which is at a resolution of 1000m $\times$ 1000m. The dataset includes precipitation, minimum and maximum temperatures, as well as partial pressure of water vapour. These daily measures are averaged per month, and squared terms are included. Growing degree days (GDD) will be included later on, following the work of Schlenker and Roberts (2006, 2009).

2.5 Price variables

Price variable $\boldsymbol{p}_{it}^{M}$ and $\boldsymbol{p}_{it}^{S}$ are futures quotations for post-harvest delivery (December for maize and November for soybeans), quoted pre- and in-season. The pre-planting period is defined to be the month of February and March. This is chosen earlier than actual planting times which are Mid April to May for maize, and May to June for soybeans. Given that the choice of crop is almost only between maize and soybeans, the planting period relevant to maize is also the one relevant for soybeans. Finally, this is also the period chosen by Hendricks et al. (2014). The pre-planting price is also relevant for the yield equation, as farmers can influence yields by choosing specific types of hybrids or the sowing densities. Later on, I shall include as well a post-planting price, which shall be defined as the May–June period. This is intended to reflect within season adjustments, such as fertiliser application. Given the sunk costs already supported, it is expected that post-planting price changes will have a smaller effect compared to pre-planting ones in the yield equations.

Futures prices are adjusted for the local basis, which is taken as the difference between the closest delivery futures price and the local spot price at neighbouring elevator. The basis is measured at the same period that the price is defined, i.e. for pre-planting prices, I use an average of February–March futures (for the December maturity) and an average of the basis at the same period.

The cash prices were obtained from elevator data found in Bloomberg[3]. I end up with a dataset of close to 2000 elevators points. Data at the field level is obtained by spatial interpolation from neighbouring elevators. I use inverse distance weighting; interpolation parameters are obtained by cross-validation. It might be objected that possible transportation costs should be considered, taking for example at distance to the elevator. However, given that I use a fixed-effect strategy at the field-level, there is no need for such an adjustment, as it will get absorbed by the fixed effects.

Figure 3 shows the location of the grain elevators and a smooth representation of the local basis. The location of the elevators follows closely where corn and soybeans are planted, compare with

[3]Bloomberg disseminates data originally collected by Data Transmission Network and Geograin. Data was geolocated, and databases were consolidated, averaging quotations over close vicinities.

Figure 3: Location of elevators and basis interpolation (corn, March 2014)

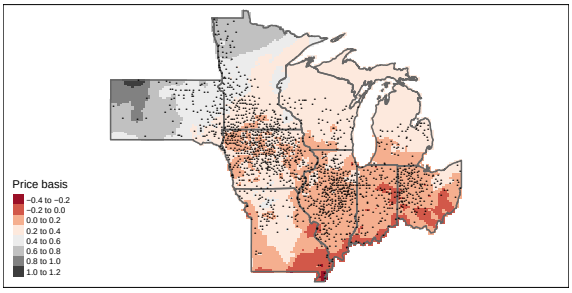


Figure **??** on Page **??**.

On ethanol refineries There is an extensive literature (see Motamed et al., 2016 for references) finding that ethanol refineries have an impact on local maize acreage response. Motamed et al. (2016) for example find that the elasticity of maize acreage with respect to local refining capacity is about 1.5. As local refineries are likely related to the price variable, this suggests that one should add a refinery vicinity variable to avoid omitted variable bias. This however raises the concern that we are adding a so-called *bad control* (see Angrist and Pischke, 2008 section 3.2.3). Bad control happens when the control variable is itself endogenous to the outcome variable. This is unfortunately likely to be the case here, where location of refineries itself depends on acreage response. This is at least the argument made by Motamed et al. (2016), motivating their search for IV variables. Besides this, effects of the refinery location are likely to translate into changes in the local basis (as found by McNew and Griffith, 2005). This implies that the yield response I am measuring is also including the effect of refineries. This only changes the interpretation of the response coefficients: they include not only year-to-year variations, but also more longer term variations.

2.6 Yields

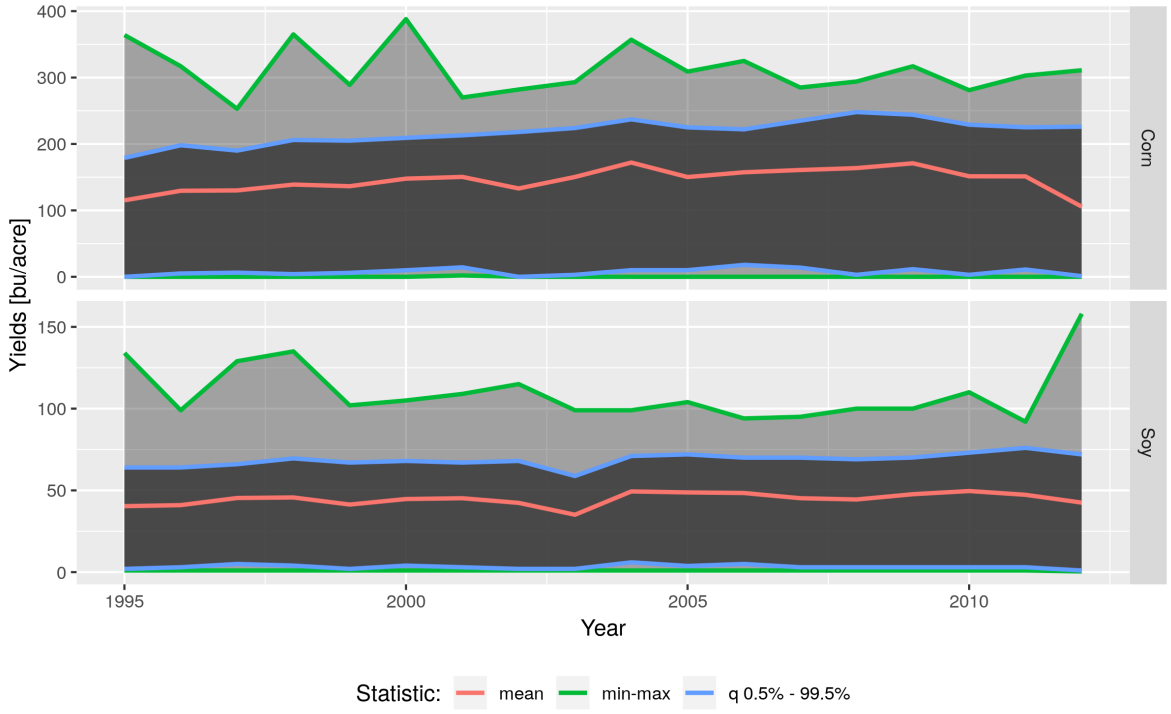
2.6.1 Data cleaning

The yield data requires a few cleaning procedures for possible outliers. The SCYM model, in particular the version 1 for soy, predicts a few negative values. This happens in six of the nine states, with the most values concentrated in IL and MO. Very high values are also found in a few cases. Setting an upper threshold is complicated. We do this looking at the USDA risk management agency (RMA) data publicly available from Lobell et al. (2014). This contains a random sample of 100 fields per county in the 3I states over 1995 to 2012. The states have among the highest yield in our subsample, so we assume that the maxima found in these states are upper bound for the other states. Figure 4 shows the mean, range and 0.5-99.5% quantile for this dataset. Based on this, we set a maximum threshold of 350 for corn and 100 for soybeans. This is a very conservative number, which is well higher than the 99.5% quantile. In fact, doing so we remove 160 corn field-year observations, and 60 soy ones.

Setting a lower limit would be also worthwhile. As the data from RMA however suggests, minimum values can be very low, often close to 5 or 10 [bu/acres] for each crop. The fact that the dataset is from the risk management agency suggests that these yields are reported to RMA for insurance purpose. It is possible that with such low yields, farmers would not even harvest the field.[4] As we observe only final yields and do not observe whether a field is harvested or not, these very low minimum values are relevant for our purpose. In consequence, I do not choose a lower minimum threshold, and only discard field-year observations with negative yields.

[4] As a consequence, these low numbers would possibly not enter the NASS county estimates.

Figure 4: Yields in 3I states, based on RMA



A few fields have also more than one year with negative values. For some of these fields (particularly in MO), the other years with positive yields are yet very close to zero, as shown in Figure 5. This suggests that there is something specific about these fields that is not well captured by the SCYM model. In consequence, I remove the entire fields that have two more more negative values, while for those with only one negative value, I only remove the specify year.

2.7 Accuracy of the yield estimates

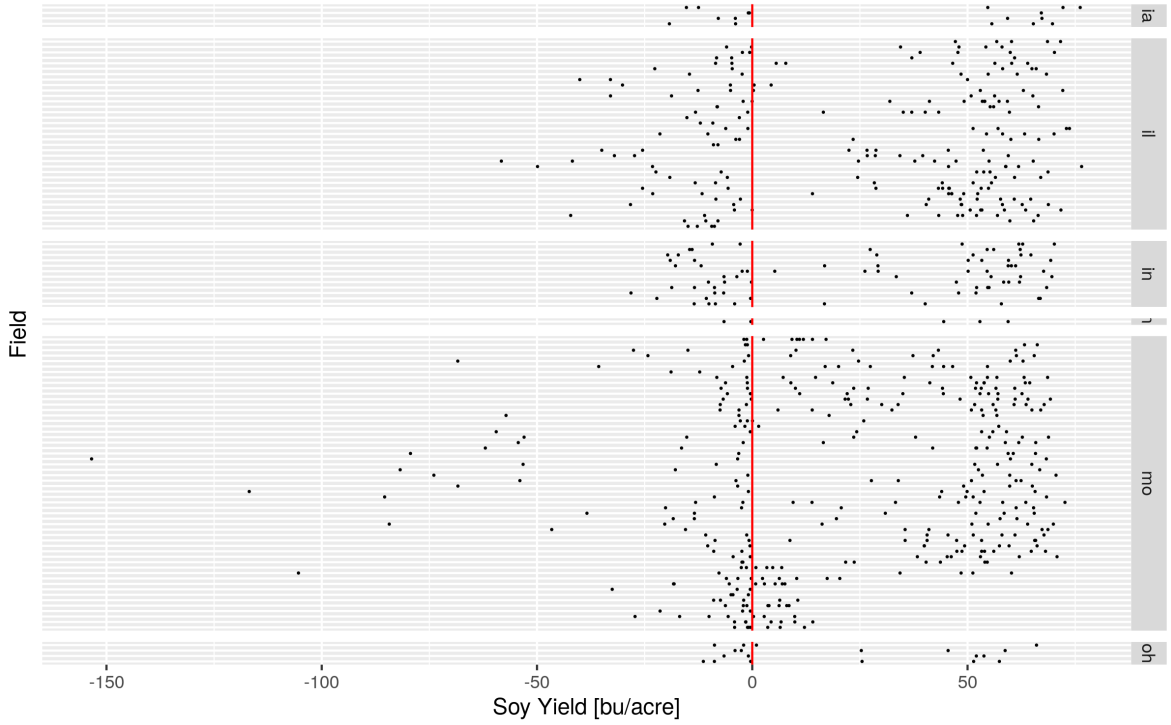
In this section, I discuss first previous estimates of the SCYM accuracy found in the literature. I analyse then the accuracy of the new SCYM estimates that were rerun for this study, extending it to 2016 and 2017, and adding more states for soybeans.

2.7.1 Previous results from the literature

The accuracy of the SCYM v1 and v2 dataset has been assessed in several previous studies, either using field-level or county aggregates yields. Lobell et al. (2015) test the accuracy of SCYM v1 in Illinois, Iowa and Indiana using data for ~10'000 fields obtained from the USDA Risk Management Agency. They find prediction R^2 between 0.14 and 0.58 for the state-year maize pairs, while the R^2 on the full sample is 0.35. Predictions for soybeans are less accurate, ranging between 0.03 and 0.5. Prediction bias in $Y^{True} = \alpha + \beta\hat{Y} + \epsilon$ arises both in the intercept and slope, although there is no clear tendency in over- or under-estimation of the values. Disaggregation of the bias suggests that it is commodity, year and state specific. Farmaha et al. (2016) use the SCYM v1 in Nebraska in a study on yield gap with 3'000 fields, and obtain predictions R^2 ranging from 0.12 to 0.34, with a tendency to over-estimate yield.

Without field-level data, accuracy of the estimates can be assessed comparing fields aggregated at

Figure 5: Soy yields of fields with at least two negative yields



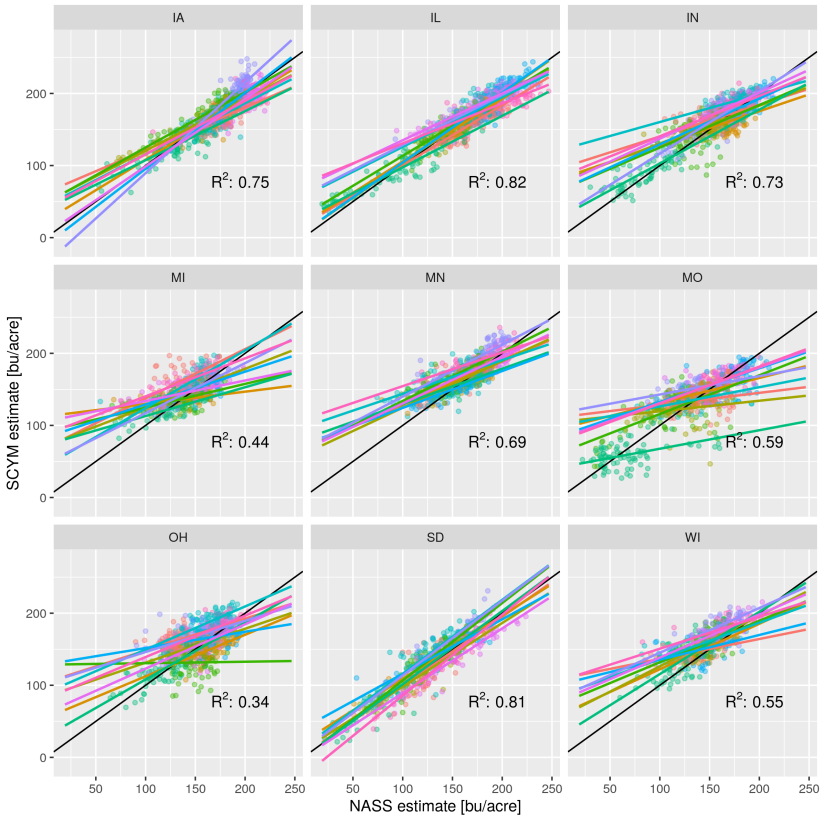
the county level to the official USDA NASS county statistics. Lobell and Azzari (2017) use the SCYM method to study yield heterogeneity at the county level. They predict field-level yield for maize and soybeans in Illinois, Iowa and Indiana, from 2000 to 2015. Averaging their yield prediction at the county level, they find R^2 of 0.67 and 0.74 for maize and soybeans respectively when comparing these with USDA county estimates. The SCYM is found to overestimate yields, with larger bias at higher yields. The authors use then the USDA county averages to calibrate their data. For SCYM v2, Jin et al. (2017) report R^2 of 0.23 and 0.88 against county NASS estimates (with median 0.72) for the 3I states, with Indiana performing worse than Illinois and Iowa. When extended to the six more states (Minnesota, Missouri, Wisconsin, Ohio and Michigan) they obtain relatively close R^2 compared to the 3I states. States with lower prediction accuracy were Michigan, Wisconsin and Missouri, which the authors explain by pointing out that these states have smaller fields, as well as more sporadic distribution of field location.

2.7.2 Assessment of the data over the extended period

The SCYM dataset was rerun for this specific study to either extend it in time, adding 2016 and 2017, or extending in space as well, adding more states for soybean. For corn, we rerun the SCYM v2, adding years 2016 and 2017. Some of the satellite products used by SCYM were in the meanwhile updated, so even on the same 2000 to 2015 period, results here are not exactly as those in the Jin et al. (2017) paper. For soybeans, I used the SCYM v1, that was run initially for the 3I states only. I extended this to 2016 and 2017, and added six more states.

Corn Figure 6 shows the SCYM v2 versus NASS county means for each state. Points are shown in different colours for each year, and lines represent the simple regression of SCYM against NASS yields. The R^2 indicated for each state corresponds to the R^2 of the general regression of all county-

Figure 6: Corn SCYM estimates, by state, 2008 to 2017



Note: colours represent different years, and the lines show the regression of SCYM against NASS yields. The line in black represent the perfect prediction line with slope 1 and intercept 0.

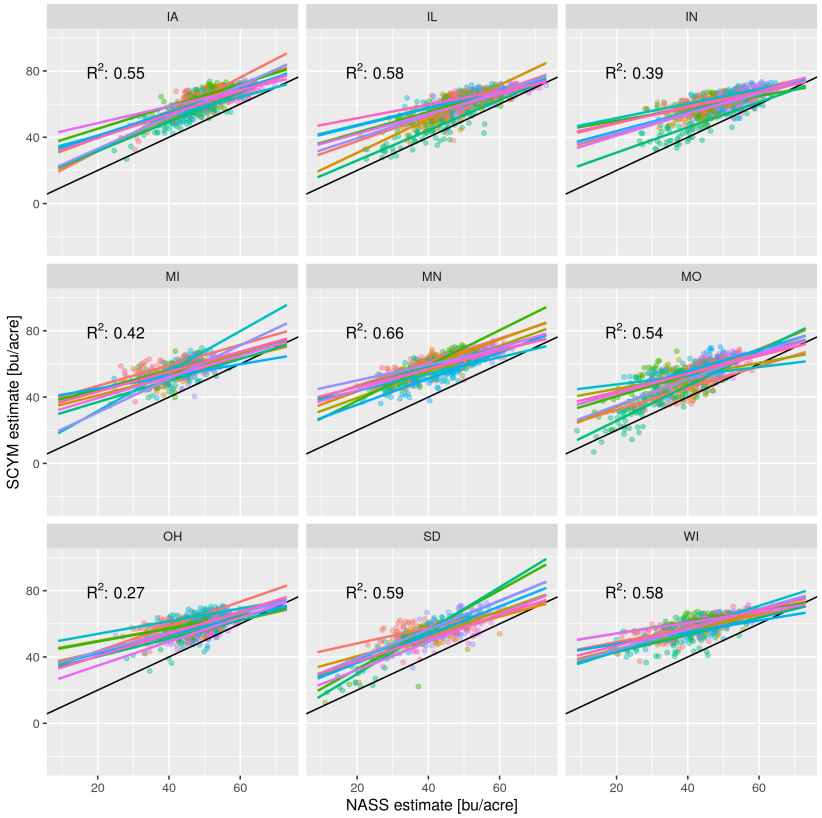
year pairs within a given state. In general, accuracy numbers are lower than those found in Jin et al. (2017). Reasons for this are not clear yet. Our data has two more years, but in general adding years would rather increase the R^2 rather than decrease it.[5]

We see that there are some important state differences. Among the 3I states, Illinois (IL) has a particularly high overall R^2, at 0.82 (correlation of 90%). The regression lines seem well aligned and relatively parallel to the 0-1 line (in black). Interestingly, the second most accurate state is South Dakota (SD), which has likewise well-behaved prediction lines. For some states, predictions are much poorer. Ohio (OH) in particular shows a low overall R^2 of 0.34 and some regression lines very distinct from the 0-1 perfect prediction line.

Soybeans Turning to soybeans from the SCYM v1 model, shown in Figure 7, results are less good. R^2 values are lower and there's a clear over-estimation bias. There is ongoing work at David Lobell's lab to improve the model for soy, which shall be integrated later on in the analysis.

[5]Indeed, if we have two subsets of data with similar R^2, a R^2 based on the two datasets assembled would increase if the two dataset have a different mean-location.

Figure 7: Soy SCYM estimates, by state, 2008 to 2017



Note: colours represent different years, and the lines show the regression of SCYM against NASS yields. The line in black represent the perfect prediction line with slope 1 and intercept 0.

3 Stylised facts

3.1 Crop shares and cropland expansion

How did cropland evolve over the period of interest? There is a large literature documenting the recent expansion of cropland in the US Midwest (Lark et al., 2017; Ren et al., 2016; Shao et al., 2016). Lark et al. (2015) analyze the cropland gains and losses during the 2008-2012 period in the whole US. They observe an overall large increase in cropland, of about three million acres. In the Corn Belt, South and North Dakota are the states that experienced the largest increase, whereas Southern Iowa and Northern Missouri also had a substantial amount of new land. The new cropland rose primarily at the expense of grassland. Based on the Natural Resource Conservation Service's (NRCS) land capability classification (LCC) system, the authors establish that this was mainly marginal land classified as suffering "severe limitations". New land was found to be principally cultivated to corn or soybeans in our area of interest, although some new land was used for wheat in the Western part of South and North Dakota, and for alfalfa on the opposite side in Wisconsin and Michigan.

Our dataset allows for some limited comparison of this phenomenon. The principal issue we face is that the data is based on the fields boundaries in 2008. If a land was previously uncultivated and used after 2008 for agricultural purpose, it might not appear in the dataset. Nevertheless, we can observe a qualitatively similar phenomenon looking at the shares of crops cultivated during the period. Figure (8) shows the percentage of the major crops in the two groups of states, 3I and 6+.[6] We see that corn and soybeans dominate the total cultivated area by a large amount. The value is close to 90% in the 3I states, and around 70%-80% in the 6+. This share also increases over time, particularly in the 6+ states. Figure 9 shows the state-by-state evolution of the six most important other crops. Interestingly, we see that in the 3I states, there is nearly no other competing crop, the remaining land being mostly uncultivated (fallow/grass categories). This is in contrast to the states in the 6+ group, that saw a decline in wheat (MN, OH and SD) and cover/forage crops (MI and WI).

3.2 Rotation patterns

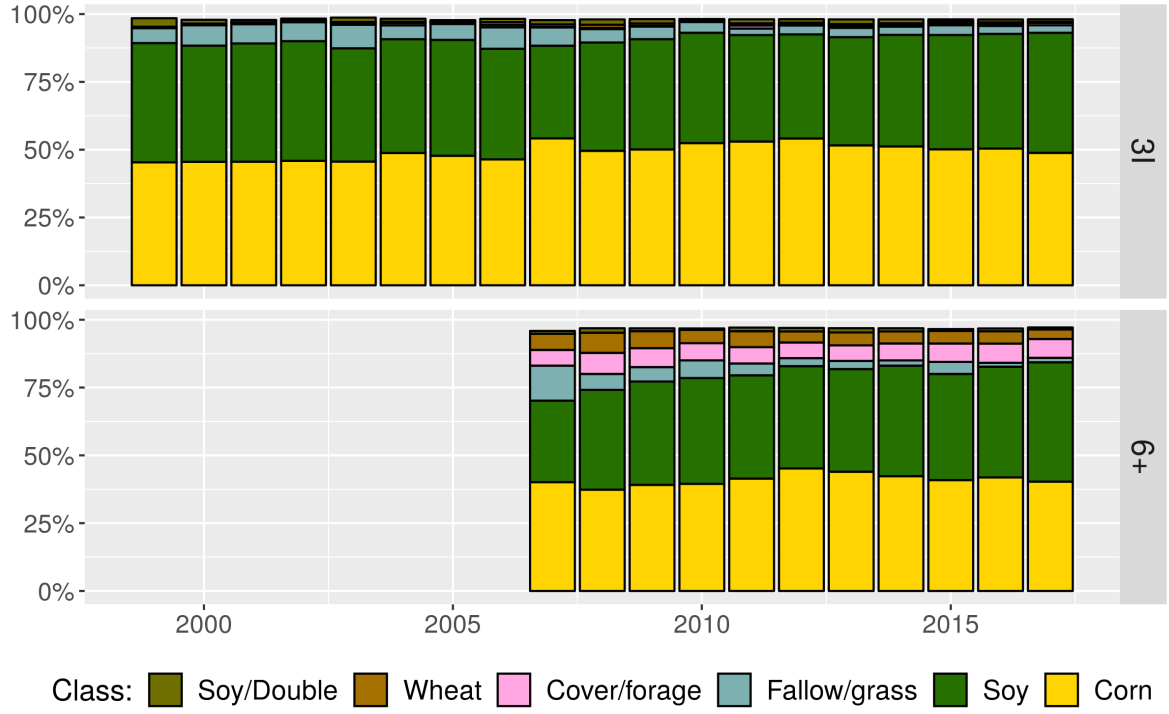
The previous Section (3.1) discussed general patterns of land use and its evolution over time. We turn now to a discussion of the rotation patterns, investigating how corn and soybeans are cultivated together. The CDL is particularly suited for this analysis, as it allows to trace back the entire history of a single field. Several studies have used the CDL for this purpose. Plourde et al. (2013) computes rotation patterns in the Corn Belt over two periods, 2003-2006 and 2007-2010. Comparing the changes over time, they find only a slight increase in the total area cultivated to corn and soybeans, but a large increase in mono-cropped corn. Ren et al. (2016) apply similar techniques in Iowa and find also an increase in continuous from 2001 to 2012. Stern et al. (2012) find likewise that most of the increase in corn area in Iowa during the 2007 price increase was achieved by corn-soybeans rotation switching to a corn-corn one. More generally, Sahajpal et al. (2014) provide an algorithm to identify relevant rotation schemes, addressing the issue that the potential number of different rotations becomes quickly very large.[7] They restrict however their analysis to a relatively short period of three years, and find that they need up to as 82 rotation sequences to cover 90% of the 13'000 actual rotations.

With the dataset at hand, I revisit several of the findings discussed in the literature. I show first that the increase in corn monocropping somehow reverted. I show also that there is a large part of fields that always rotate over the full sample 2008 to 2017, and even over the longer sample 2010 to 2017. To start with, Figure 10 shows the average rotation pattern for corn and soybeans over the 2008-2017 period. There is in general slightly more corn than soybean on a typical year. The following year, almost 70% of the fields will rotate. For fields planted to the same crop, a large part will be corn-corn, while a very small part will be soy-soy.

[6] The 6+ states correspond to Ohio, South Dakota, Minnesota, Missouri, Michigan and Wisconsin.

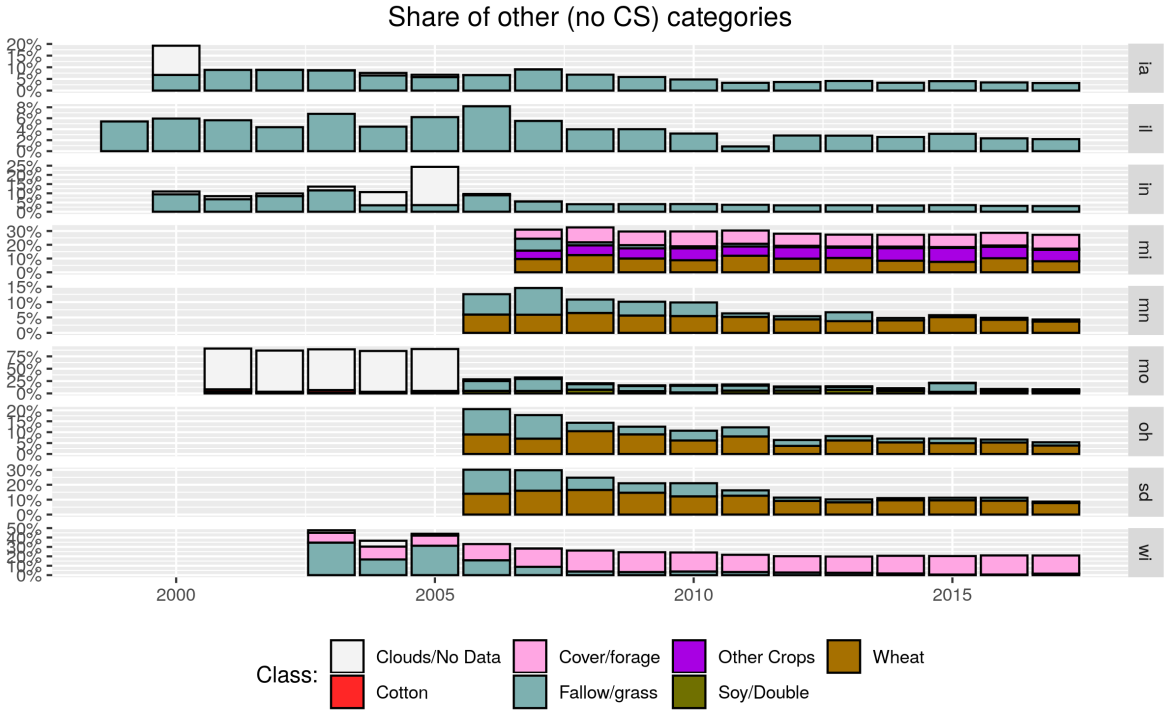
[7] The CDL has 132 distinct classes. So after N years, there are 132^N potential rotation schemes. For $N = 3$, this is 2 millions, and for $N = 4$, it is 300 millions.

Figure 8: Shares of crops over the 2000–2017 period



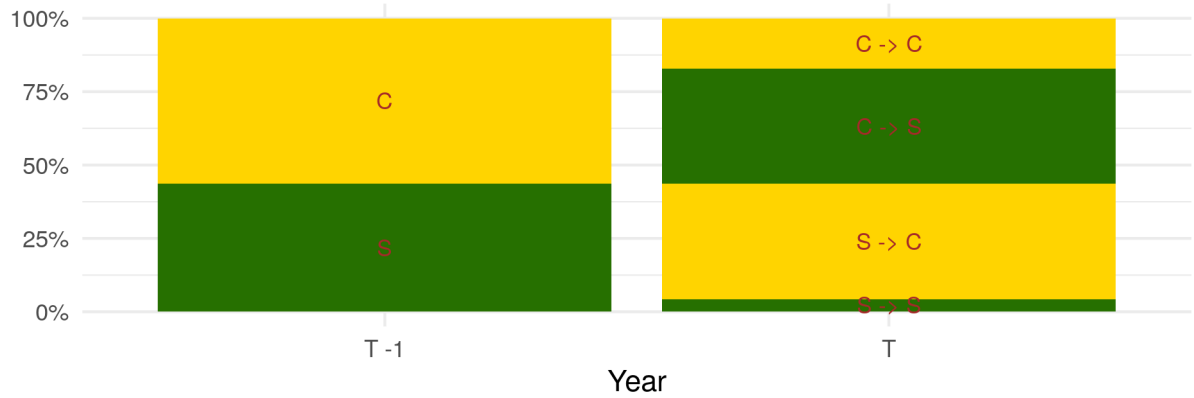
Note: data shows the percentage shares of the most important crops, based on the CDL.

Figure 9: Shares of other crops, by state and year



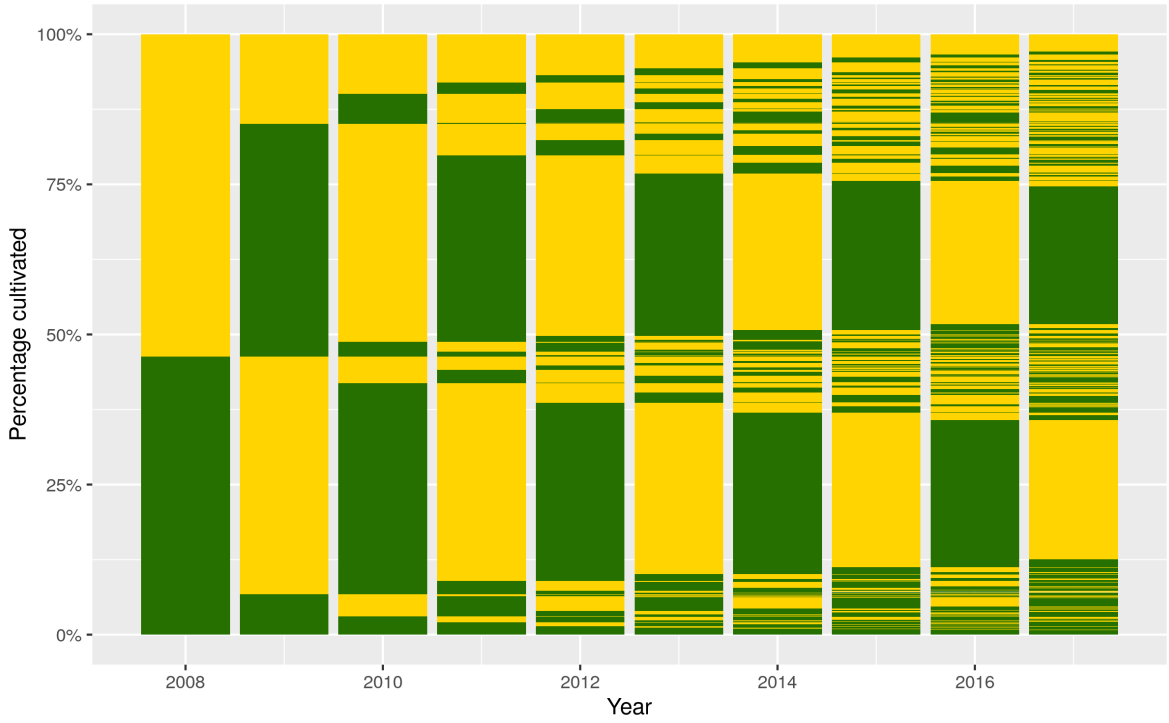
Note: shares based on the fields available retained in the dataset.

Figure 10: Average rotation pattern for corn and soybeans



1 Note: data shows the average shares at time $T-1$, and conditional on those, the shares in time T.

Figure 11: Conditional rotation patterns for fields doing always either corn or soy, starting in 2008



1 Note: data shows the planted share, conditional on previous years, recursively, starting in 2008.

Figure 11 extend the previous figure to the whole period. Starting from the left, we see the share planted to corn or soybeans in 2008. In 2009, we see how many were planted to corn and soybean, very similar to the previous figure (which was averaged over all years). The next period, 2010, is again conditional on the choice of 2009 and 2008. We can see fro example that the share of C-C-C is the third most frequent sequence, followed by C-C-S. The last column in 2017 shows the resulting sequences, after ten years. Theoretically, we have $2^{10} = 1024$ different possibilities. It is striking however to realize that there is a large part of fields that always did rotate, at 46%. The share of fields practising monocropping over the whole period shrinks to 3% for corn, and to a very small 0.7% for soybeans. These numbers were for the restricted subsample of fields that do always either corn or soybeans over 2008-2017. Figure 17 in the appendix shows the conditional rotation patterns for all fields, adding the *Other* category. While we have now $3^{10} \sim 60'000$ possible sequences, the always-rotating sequence is still strikingly visible.

While the conditional rotation figure indicates clearly a predominant category of always-rotaters and always-corn, it it less clear which other categories are often represented. The issue is that a sequence with nine rotations can be represented in many different ways, depending on the time when the corn or soybean value is repeated. Table 1 shows the percentage of fields doing ten, nine up to zero rotations, in the subset of fields doing always either C or S, or sample of fields doing at least once another crop. We see that in the CS-only subset, there is still a large proportion of fields almost rotating, i.e. doing nine rotations over ten years.

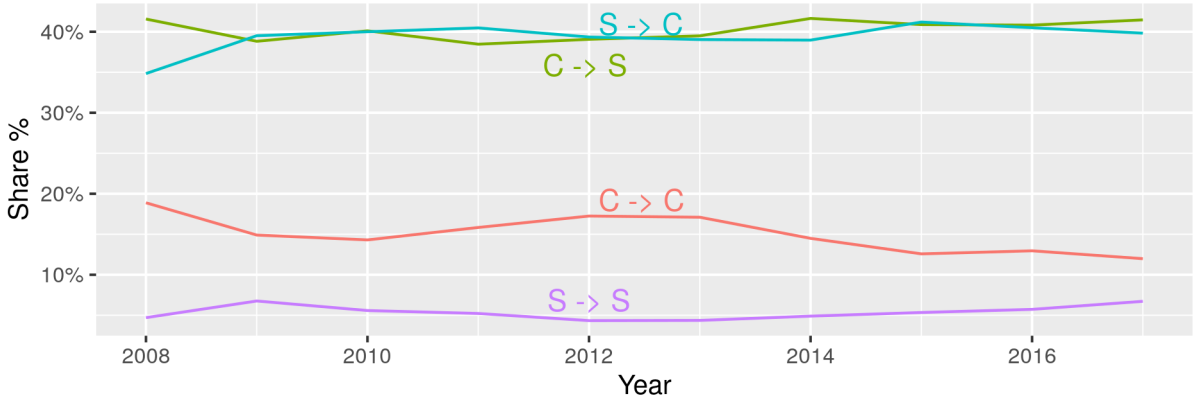
Turning finally to the evolution of the rotating patterns over time, Figure 12 shows the evolution of the year-to-year rotation patterns. In this subset, the increase in C-C monocropping described elsewhere in the literature seems to slow down, or even decrease.

Next Figure 13 shows similarly the conditional transition probabilities, rescaled each year to 100% for each crop. Looking at the first panel for corn, we see what is the probability to stay in corn in 2008, 2009 etc, as well as the probabilities to switch out of corn to the *Other* category. We see that

Table 1: Number of rotations per field, percentage

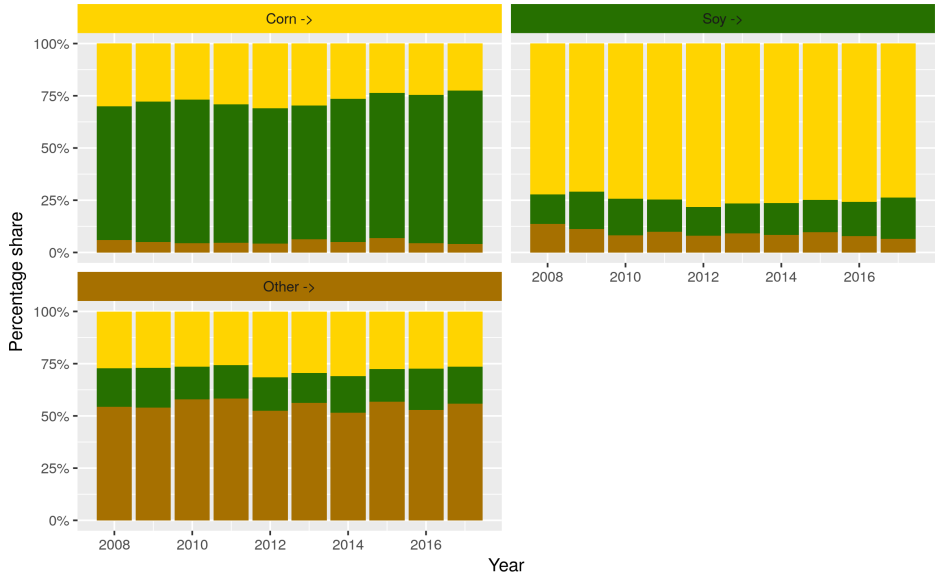
	Only CS		All	
N rotations	N	Perc	N	Perc
0.00	41’241	1.9%	245’648	11.2%
1.00	10’454	0.5%	71’136	3.2%
2.00	35’590	1.6%	89’962	4.1%
3.00	23’920	1.1%	85’250	3.9%
4.00	46’783	2.1%	94’674	4.3%
5.00	46’560	2.1%	87’044	4%
6.00	72’769	3.3%	94’661	4.3%
7.00	94’043	4.3%	80’006	3.6%
8.00	133’081	6.1%	85’982	3.9%
9.00	195’883	8.9%	29’892	1.4%
10.00	529’085	24.1%	0	0

Figure 12: Evolution of the year-to-year rotation patterns



1 Note: data shows the share of year-to-year conditional probabilities, on the subset of fields that do always either corn and/or soy.

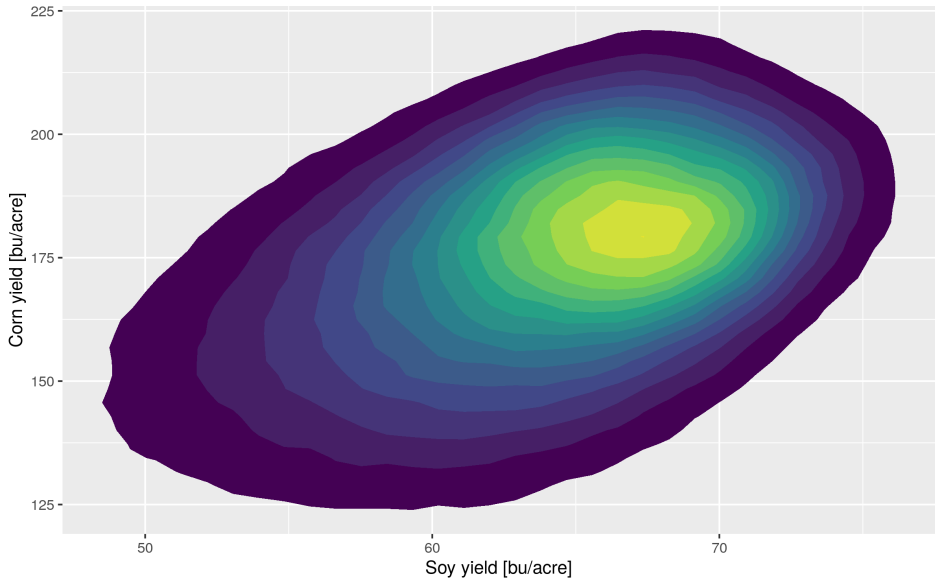
Figure 13: Year to year transition probabilities for each category



1 Note: data shows the share of year-to-year transition probabilities, conditional on the previous crop.

the probability of switching to *Other* is very low for corn, and low as well for soybeans. When new land (i.e. from the *Other* category) enters corn or soybean, it will mainly do so starting with corn. It should be emphasised here that the percentage shares are rescaled each year for each crop. Observing a higher rate of exit of *Other* (high $O \to C$ and $O \to S$) compared to rate of exit of *Corn* or *Soy* to *Other* ($C \to O$ and $S \to O$) implies that the total share of *Other* decreases over time. This reinforces the point made earlier of a decrease in *Other* crops in the total shares (see Figure 9 and 8).

Figure 14: Distribution of corn and soybean yields



1 Note: data shows the 2D density estimation of the distribution of corn and soy means per field.

3.3 Land quality model

The crop distribution maps suggest that there are specific patterns of crop location, with some areas in the 3I planting predominantly corn, and some further South planting more soy. Does this mean that some areas are good for corn but bad for other? Or is a land that is good for corn also good for soy? This question is difficult to answer, since we never observe two different yields on the same field. Luckily enough, with ten years of data, we observe frequently at least one corn and one soy yield on the same field. To verify the correlation between corn and soybeans yields on the same field, I average corn and soybean yields for all fields that do at least once corn and once soy. Figure **??** shows the distribution of the plot average yields, over the subsample of fields that do always corn or soybeans. There is a clear positive correlation between the yields of the two crops: a field that is good for corn is likely to be also good for soybeans.

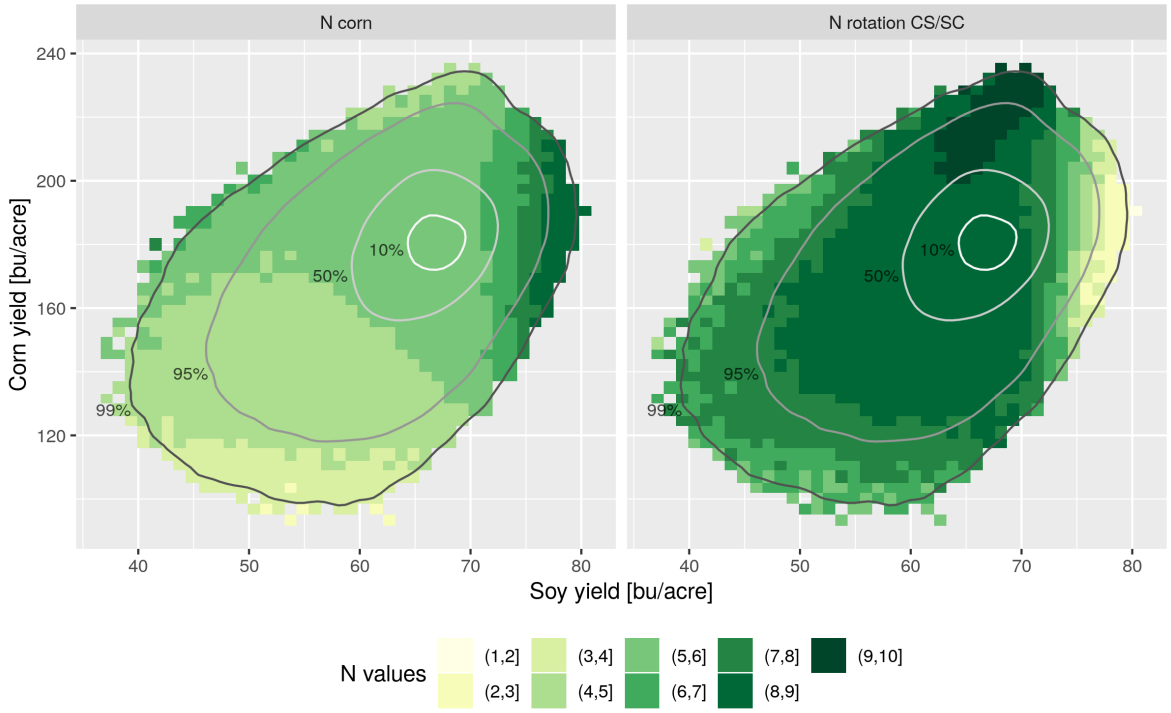
Figure 15 shows the average number of years planted to corn, and average number of $\langle CS \rangle$ rotations for various values of yields. The left panel shows how many years the fields were planted to corn. Fields with low average productivity, be it for corn or soybeans, are rarely planted to corn (and hence often to soybeans). On the other side, higher fertility fields are most likely to be planted to corn. Turning to the right panel, showing the number of $\langle CS \rangle$ rotations, we can see that in general, the probability of doing many rotations is particularly high for fields that have a rather similar value of corn and soy yields.

3.4 Yield variations at the county and field level

How much variation in yields is uncovered when moving from a county analysis to a field-level analysis? Or put differently, how does the between-county variation (observable with official statistics) compare to the within-county variation (observed with satellite data). This

This questions is important in many aspects. First, it has implications in terms of sampling rates required to obtain accurate measures of the county average yield. The higher the correlation (the lower the variation), the lower the number of observations are required to estimate the mean. Second, it can inform the literature on crop insurance, where the concept of risk pooling is intimately linked

Figure 15: Average number of corn and rotation per cell



1 Note: data shows the 2D density estimation of the distribution of corn and soy means per field.

to the (spatial) correlation among among observations. Finally, and more generally, it gives an idea of what we gain by using satellite data instead of county data.

To measure the amount of variation coming from different spatial scales, I run an analysis of variance (ANOVA). An ANOVA informs on how much each set of variables (here the fixed effects) contributes to the total variance. Figure 16 shows the percentage of the total variance for different levels.

4 Conclusion

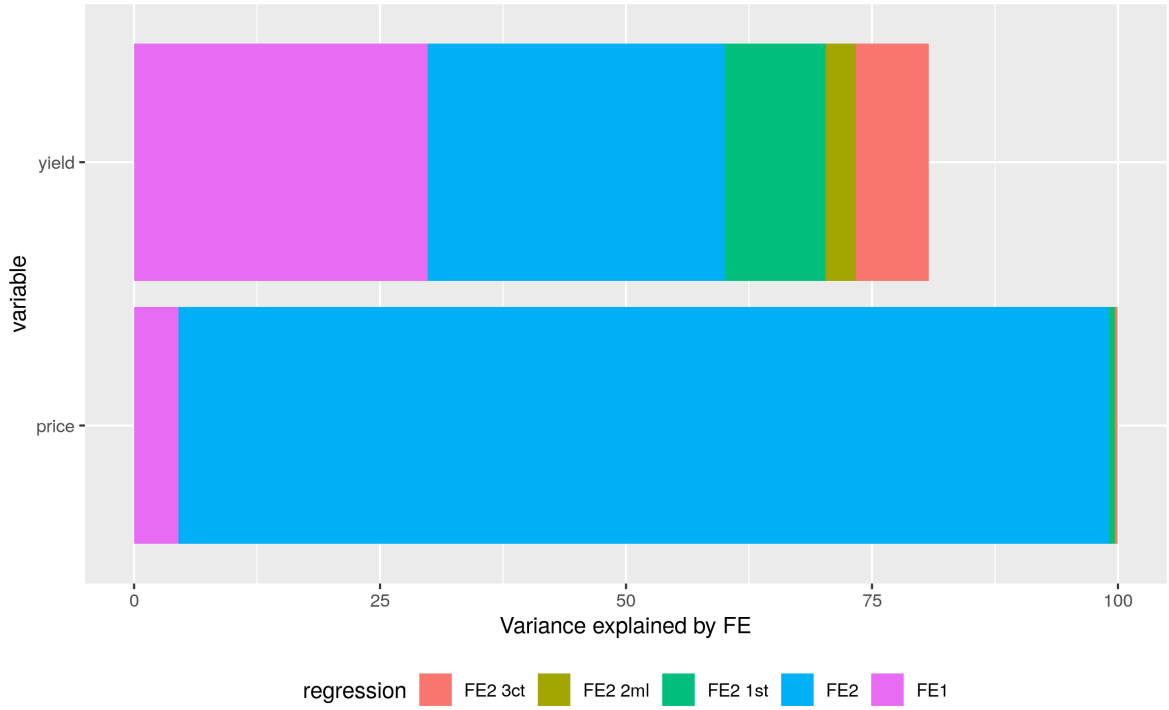
This chapter gave an overview of the dataset, its different sources, and how it was assembled. Building on this, an exploratory analysis was carried on, that uncovered several stylised facts that to my knowledge have not received much attention in the literature. Particularly interesting is the finding that corn tends to be planted more often on fields that have high apparent yields, and soybean more often on the lower quality fields. The implications of this finding are discussed in more details in the next chapter that seeks to estimates the yield rotation effect.

5 Appendix

References

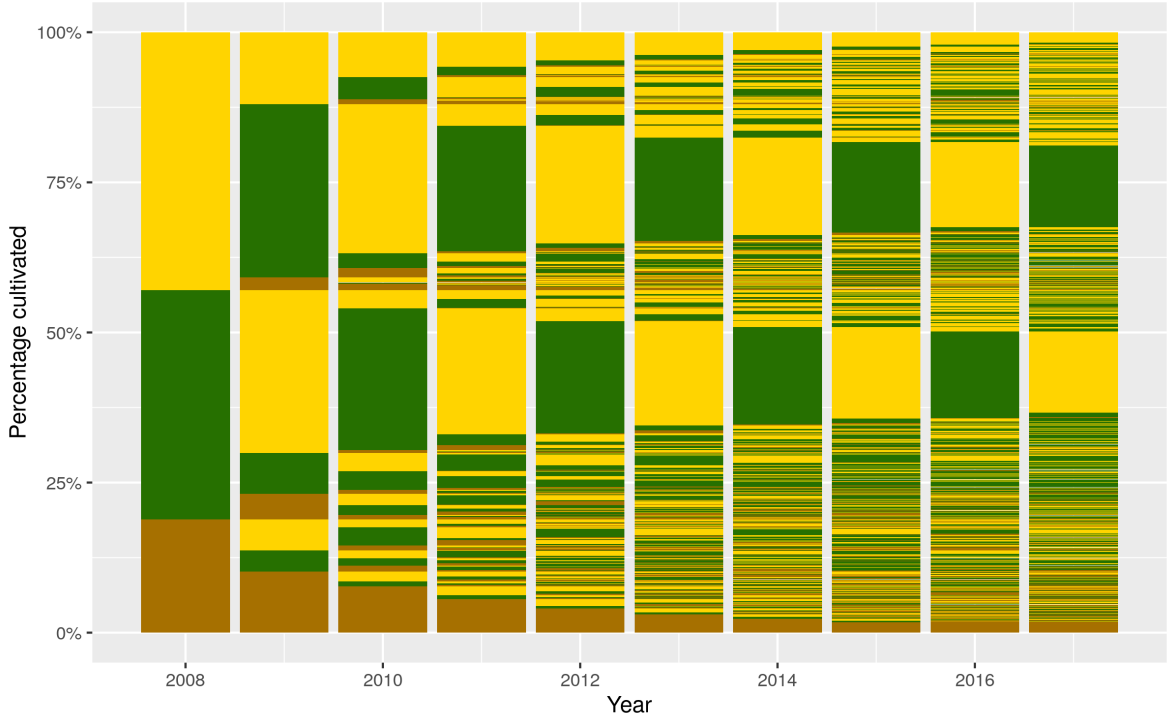
ANGRIST, J. D. AND J.-S. PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Figure 16: Analysis of variance of yields and prices



1 Note: data shows shares of total variance explained by each set of fixed effects.

Figure 17: Conditional rotation patterns for all fields, starting in 2008



1 Note: data shows the planted share, conditional on previous years, recursively, starting in 2008.

BORYAN, C., Z. YANG, R. MUELLER, AND M. CRAIG (2011): "Monitoring US agriculture: the US Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program," *Geocarto International*, 26, 341–358.

FARMAHA, B. S., D. B. LOBELL, K. E. BOONE, K. G. CASSMAN, H. S. YANG, AND P. GRASSINI (2016): "Contribution of persistent factors to yield gaps in high-yield irrigated maize," *Field Crops Research*, 186, 124 – 132.

HENDRICKS, N. P., A. SMITH, AND D. A. SUMNER (2014): "Crop Supply Dynamics and the Illusion of Partial Adjustment," *American Journal of Agricultural Economics*, 96, 1469–1491.

JIN, Z., G. AZZARI, AND D. B. LOBELL (2017): "Improving the accuracy of satellite-based high-resolution yield estimation: A test of multiple scalable approaches," *Agricultural and Forest Meteorology*, 247, 207 – 220.

LARK, T. J., R. M. MUELLER, D. M. JOHNSON, AND H. K. GIBBS (2017): "Measuring land-use and land-cover change using the U.S. department of agriculture's cropland data layer: Cautions and recommendations," *International Journal of Applied Earth Observation and Geoinformation*, 62, 224 – 235.

LARK, T. J., J. M. SALMON, AND H. K. GIBBS (2015): "Cropland expansion outpaces agricultural and biofuel policies in the United States," *Environmental Research Letters*, 10, 044003.

LOBELL, D. B. AND G. AZZARI (2017): "Satellite detection of rising maize yield heterogeneity in the U.S. Midwest," *Environmental Research Letters*, 12, 014014.

LOBELL, D. B., M. J. ROBERTS, W. SCHLENKER, N. BRAUN, B. B. LITTLE, R. M. REJESUS, AND G. L. HAMMER (2014): "Greater Sensitivity to Drought Accompanies Maize Yield Increase in the U.S. Midwest," *Science*, 344, 516–519.

LOBELL, D. B., D. THAU, C. SEIFERT, E. ENGLE, AND B. LITTLE (2015): "A scalable satellite-based crop yield mapper," *Remote Sensing of Environment*, 164, 324 – 333.

MCNEW, K. AND D. GRIFFITH (2005): "Measuring the Impact of Ethanol Plants on Local Grain Prices," *Review of Agricultural Economics*, 27, 164–180.

MOTAMED, M., L. MCPHAIL, AND R. WILLIAMS (2016): "Corn Area Response to Local Ethanol Markets in the United States: A Grid Cell Level Analysis," *American Journal of Agricultural Economics*, 98, 726.

PLOURDE, J. D., B. C. PIJANOWSKI, AND B. K. PEKIN (2013): "Evidence for increased monoculture cropping in the Central United States," *Agriculture, Ecosystems & Environment*, 165, 50 – 59.

REN, J., J. B. CAMPBELL, AND Y. SHAO (2016): "Spatial and temporal dimensions of agricultural land use changes, 2001-2012, East-Central Iowa," *Agricultural Systems*, 148, 149 – 158.

SAHAJPAL, R., X. ZHANG, R. C. IZAURRALDE, I. GELFAND, AND G. C. HURTT (2014): "Identifying representative crop rotation patterns and grassland loss in the US Western Corn Belt," *Computers and Electronics in Agriculture*, 108, 173 – 182.

SCHLENKER, W. AND M. J. ROBERTS (2006): "Nonlinear Effects of Weather on Corn Yields*," *Applied Economic Perspectives and Policy*, 28, 391.

——— (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.

SHAO, Y., G. N. TAFF, J. REN, AND J. B. CAMPBELL (2016): "Characterizing major agricultural land change trends in the Western Corn Belt," *ISPRS Journal of Photogrammetry and Remote Sensing*, 122, 116 – 125.

STERN, A., P. DORAISWAMY, AND R. HUNT (2012): "Changes of crop rotation in Iowa determined from the United States Department of Agriculture, National Agricultural Statistics Service cropland data layer product," *Journal of Applied Remote Sensing*, 6, 1 – 16 – 16.

STEVENS, A. (2015): "Fueling Local Water Pollution: Ethanol Refineries, Land Use, and Nitrate Runoff," Tech. rep., University of Berkeley.

THORNTON, P., M. THORNTON, B. MAYER, Y. WEI, R. DEVARAKONDA, R.S.VOSE, AND R. COOK. (2017): "Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version3." ORNL DAAC, Oak Ridge, Tennessee, USA.