Sept 4

**Desk**

Started new GEE file ‘Crop Timing Map’ to look at general temporal and spatial trends in soy timing – made video of median plant, harvest dates over Dave’s cells over 2003 to 2016, and also calculated time trends

Sept 5

**Desk**

In GEE file ‘Crop Timing Map’, continue looking at spatial and temporal relationships. Tried Geary’s C and GLCM contrast for spatial correlations; autocorrelation and autocovariance for temporal trend; and Mann Kendall for trend detection, but it’s not very illuminating for pixel level estimates… perhaps need to do cell level.

For autocorrelation stuff, perhaps the “newness” of a cell in terms of soy planted will decrease its autocorrelation level. So perhaps get rid of cells that don’t have all 14 years of data to make each cell comparable.

Also tried to filter out “bad estimates”, but once we do that, there are very few pixels left to work with for some reason. So, don’t filter out bad estimates; instead, hope that taking the median will help with that. Looks like median is reasonable.

Sept 6

**Desk**

Outline correlation/causal inference studies

**Meeting with Mike Sohn**

Book by Noel Cressie for geostatistics

Is there a reason we’re seeing the spatial pattern in estimated plant/harvest at the cell level? Explain spatial variation.

Need a measure of variability of plant/harvest estimate within each cell (like variogram) to determine appropriate size of the cell.

Also need to compare variability of plant/harvest estimate to variability of weather at the pixel level.

Exploratory data analysis: fit pdfs to different years, different regions (different temporal and spatial scales)

Effect of deforestation may have a lag – consider autocorrelation

Look at map of soy vs no soy – perhaps the effect of deforestation will impact whether soy is planted at all. Look at autocorrelation and spatial correlation of soy map.

Granger causality from a probabilistic view: Given that A causes B, what’s the likelihood of B causes C?

Do causality stuff probabilistically, not classically because things are so variable and “happenstance”

Sept 7

**Desk**

In GEE file “Crop Timing Map v2”, updated quality mask to look at plant/harvest estimates to mask quality, not masked by quarter period or by peak timing.

Sept 10

**Desk**

Wrote Research Plan

Sept 11

**Desk**

New GEE file: “Crop Masking” to look at masking out soy pixels by EVI (EVI must have fitted amplitude above 0.2 and max EVI must be greater than 0.8); by very bad estimates (i.e. planting/harvest estimates are outside of 0 to 365); by crop cycle (must be between 60 days and 150 days) and by median +/- stdev within a 50 km cell.

New GEE file: “Timeseries Analysis v9” in order to export max EVI and EVI amplitude; the assets are named as ‘singleCropPeakTiming\_landMask\_3\_year’, ‘singleCropPeakTiming\_full\_2\_year’, etc. Then, to combine all years including the amplitude and the peak EVI value, new GEE file: Timeseries Produce Final Asset v2. The new assets (with peak EVI and amplitude) are called singleCrop\_timing\_3 and singleCrop\_timing\_full\_2.

**Meeting with Gabriel**

Patterns of planting date:

* Climate is less seasonal in the south; in the south, pretty much no seasonality in rainfall.
* Mato Grosso plants earlier than in the south because onset of rainy season starts in the north, and moves diagonally southward. The sanitary break matters for planting but it doesn’t clash much.
* In NE Brazil, like Para, plant date should be very late.
* Second crop planting is constrained by excess rain during the middle of the wet season. First crop harvest is very sensitive to rain, because harvest happens in the middle of the wet season. In Mato Grosso, there’s a dry spell in Jan/Feb that makes harvesting of first crop, and therefore DC, very favorable
* There is a lot of irrigation on the east of Matopiba and in Goias (east Goias) so may see some unusual behavior there

Gabriel will give me: irrigation map; also look in AgroServe state model database on GitHub; onset map from Xavier

Causal inference:

* At the 50km cell, we won’t get much rainfall data within the cell
* What should I control for in DID analysis? Latitude, radiation (i.e. people might not plant directly following onset; might wait for there to be more radiation)
* Newly deforested areas will be experimental; may see lots of different practices (non-climate related) that may be difficult to do causal inference on.
* Deforestation -> onset is easier to determine causally because it’s all physical; onset -> planting has a lot of economic/human factors so maybe just study correlation
* There is very little research on nonclimate determinants on planting date. People will plant on weird dates due to prices, but it should be a very small signal. Perception of weather can lead to wide range of plant dates, but these decisions do depend on neighbors so there should be a relatively smooth response.
* Probably the main nonclimate determinant of planting date is credit access (at least, it’s perceived as such); ask Avery about how credit access explains planting. Gabriel never saw an example of when late credit access caused late planting; credit access simply gives a recommended window.

Sept 12

**Desk**

Worked on GEE file ‘Crop Masking’, tested two ways to combine masks: (1) feasibility mask x cycle length mask, then use these two crossed masks to do a neighborhood mask (median +/- stdev), then crossed the neighborhood mask with EVI mask. This mask is called the full mask. (2) feasibility x cycle length x EVI. This is called the partial mask, and this is probably better to use.

**Meeting with Sally**

Research related stuff

* Look at recommended planting windows vs what farmers are doing – do they actually follow AgriTempo recommendations?

The overall plan/prospectus:

1. Introduction
   1. Connect my questions to climate change, land use change before going directly to agri. Sell it as environmental engineering. Mention ‘global change’, how agriculture contributes to global change, mention that ‘farmer behavior drives impacts’, ‘LUC drives impacts’, hype the importance of Brazil in driving climate change
   2. Techniques: RS and data science, and how to use these tools to find the answer
   3. Treat Brazil as an example
   4. Uncertainty and lack of information can be helped by using RS and data
2. develop dataset of cropping dates as accurately as possible
   1. Create a diagram describing the algorithm I use to justify keeping various pixels; confirm this with Avery and Gabriel
   2. Lit review of trends in farmer behavior
3. describe temporal and spatial patterns
   1. study time trend in the stdev of predictions
   2. look for autocorrelation in timeseries, and do we need to remove them before fitting? (e.g. if there’s a 7 yr ENSO signal and we have short timeseries, the trend we see really depends on where you truncate the timeseries.
   3. Scales: try different scales and see if they’re different. In general don’t aggregate too soon because we don’t know if the average of the trend is the same as the trend of the average
   4. Lit review of timeseries analysis to pull out trends in means and variances
   5. Ask Avery about nonclimate controls
   6. We don’t have to do triple collocation; maybe weight two things based on the variance? Look up methods for fusing
   7. Can try Landsat alone, Modis alone, and SAR
   8. Look at Bayes or Kalman filtering for fusing
   9. Look up methods for fusing uncertain datasets
4. Think about the nonclimate controls on plant date and how we can tease them out

Sept 13-20

**Desk**

Write Qual Prospectus v2, introduction, and research plan diagrams

Sept 21

**Sally meeting**

Need to get more validation information for plant/harvest dates. Check with Agritempo? Need to know the hard bounds of ground truth information

For land misclassification, read Gopal’s paper on mixed pixels – but really we’re just doing this to get a sense of how much improvement we have in using Landsat over Modis alone

For Sentinel – fit Sentinel to the same scale as EVI (over a time window defined by Modis); for temporal resolution error, compare Sentinel and Modis to Matopiba data

Error propagation: it’s easy to deal with additive and multiplicative error. Think about scale of how errors will propagate. At the scale of the (known) soy pixel, how will errors propagate? On top of the pixel error, across the map, what are the errors due to misclassification? (what would this land use map related error mean for aggregated statistics?)

Sept 26

**Tina meeting**

How do they attribute deforestation to climate change?

How do they know or suspect that climate change leads to productivity declines?

What is the crop model we’re using? How do we attribute climate change to yield declines?

**Gabriel email**

- About the center pivots:

I believe it's highly unlikely that someone would dismantle one for a long time after they are implemented. They are very expensive and take some work to mount. In many regions they even have to dig a well in its center for water. And as they are expensive they're generally implemented after some planning. It is possible that a farmer would turn it off for a year, but very unlikely as they are highly profitable and I don't think it's hard for farmers that already have one to get credit for its operating costs in a bad year.

- About the irrigated area in general:

I never heard of it shrinking anywhere. The above is valid for most irrigation systems (with the likely exception of inundated rice), they are costly to implement and profitable to maintain. What may happen is that in some dry regions there may not be enough water for irrigating in bad years. However, those situations don't happen very often, as they normally rush to dig wells everywhere. What I've seen happening in those situations is the population of nearby towns rioting about not having enough water for their personal needs. Either way, I don't this farmers would just dismantle their irrigation systems, they'll just not irrigate in bad years. So I'd say it's monotonic in a region.

There might be a way to check, as the 2006 census also surveyed irrigated are by municipality, but I don't think it's worth the hassle. You can have an idea from Tabela 2 in the irrigation atlas I sent you, it has the aggregated estimates by state for census years. The late 80s and early 90s were a bizarre time for Brazil, when things like credit and even profit were too unstable due to hyperinflation and fast changing policies. You see some states, mostly on the northeast, reducing their irrigated area between the 2006 and 2015 census years, and the atlas itself says those were probably just idle on that specific year as it was a dry one, you can quote it on that.

- About the regions:

There is more irrigated area in the southeast and central parts of Brazil, yes. I normally think about the northeast first because irrigation is a more important part of agriculture there as they have drier climates, so sorry to mislead you. It is growing like hell in western Bahia though.

- About the precipitation data, that was from Xavier.

- About the agritempo planting recommendations, they are made on a municipality basis. They do they own interpolations of weather data and do some simple water balance and frost index calculations to come up with those planting windows. The data comes in weird tables per municipality on their website, and there are even some missing ones, so the one I sent you was gap-filled with the nearest neighbor municipality. If you want the not gap-filled ones, there should be a variable on that shapefile that still have the holes.

**Sally meeting**

For planting/harvest date error, get a high resolution image where I can see tilled field, versus just planted field; and pre and post harvest, and compare what I would estimate with my method. Look at SPOT, IKONOS and GeoEye. This is another set of checks. To combine different estimates of the same error, take the worst case.

To calibrate: perhaps check the influence of temperature and rainfall between planting and greenup. To condition the calibration on precipitation, see if the rainfall amount between plant and greenup and onset are correlated. If they are correlated, don’t use it.

Incorporating many errors into the regression: do bootstrapping. Have a pdf of onset date representing error of onset (assume we know error due to Xavier precip and assume that it’s normally distributed); have another pdf of plant estimates, maybe we will have a non-normal distribution. Take a sample for each pixel from the pdf to create a bunch of maps of planting and onset, and then do a regression on each map. Summarize the regression coefficients by a distribution of regression coefficients.

For fixed effects: perhaps do fifty ish boxes across Brazil, don’t do too many.

For the memory, do memory of the plant date or memory of the onset from previous years?

Make an onset minus planting map to look at the spatial correlation to decide on an appropriate spatial scale. If there’s not much spatial correlation, may need to go for smaller spatial units.

Questions: (1) how well do farmers match onset? (Do a correlation, maybe with bootstrapping) (2) model the behavior of farmers based on which climate variables? Over what period of time? i.e. 3 months before planting date?

Instead of using actual (omniscient) onset, perhaps use some antecedent conditions as the climate variable.

Read about farmer behavior in Brazil and how they decide on plant date (i.e. what is the climate variable to use?). Ask Avery for suggestions. Frame these as hypothesis tests.

Hypothesize ways farmers make decisions. H1: that decisions vary over space. H2: that decisions depend on antecedent conditions. H3: that there’s memory. H4: that decisions are trending. H5: forecast info from Agritempo impacts decisions.

Think about exploratory analysis that would help inform these hypotheses. Design fixed effects specifications that will result from them.

Think about another model: diff bet plant and onset = fcn(time) to see if the amount of information about onset improves over time or differs over space.

Sept 28

**Meeting with Sally**

For propagating standard error from individual regressions as well as the confidence interval from bootstrapping, do a linear combination of random variables. This is ok because the individual coefficients from individual bootstrapped samples are random variables themselves, and we’re averaging them to get an averaged coefficient. The average coefficient will come with an error term that’s the sum of: [(a\_i)^2][error\_i^2] where a\_i is 1/(number of bootstrapped regressions) and error\_i is the standard error from the ith bootstrapped regression. Also report the percent of regressions that have statistically significant results.

For qual, create maps of onset, onset variance, etc.

For stats: preliminary tests at the level of all Brazil:

1. create a table where each CAR polygon has the following attributes:
   1. size of property
   2. plant date
   3. mean onset
   4. variance in onset
2. do regressions of the form: plant = b\*(one variable) to see if that one variable is at all correlated
3. do regressions including all variables: plant = fcn(all variables) to see if any of the variables are significant
4. do regressions including all except one of the variables to see if the fit changes as a result of taking it out
5. do regressions 2 to 5 for large chunks of Brazil
6. do regressions 2 to 5 for smaller chunks of Brazil
7. move to fixed effects, using the above to inform. Perhaps use a state as the fixed effect.