Oct 1

**Sally Meeting**

* Create a diagram of a data generating process: a conceptual model for how farmers decide to plant. Look at endogenous and exogenous variables.
* Potential methods to use: OLS, fixed effects regression, machine learning (decision tree)

For prospectus:

* The problem statement
* Conceptual diagram of process
* Research questions (description of the past: (plant – onset) and what that depends on; and prediction equation for the future (so this can only be based on what farmers can observe – because it’s explaining farmer behavior))
* Exploratory data analysis (correlations, whether anything is linear; hierarchies through space and time)
* Methodological details to work out
  + Describe climate using relative or fixed time?
  + Which of the potential methods (OLS, fixed effects regression, machine learning)

Read about:

* How farmers make planting decisions anywhere in the world
* Agent-based models, farmer adaptation to climate change

**Octavia Meeting**

* Use a random forest regressor for continuous data. Random forests avoid overfitting because they add noise into data, and fit many decision trees, and each decision tree ‘votes’ on the final decision tree. The output is a decision tree; the inputs are matrices of explanatory and dependent variables.
* To get climate sensitivity, don’t look at decision tree itself; perturb the inputs with simulated data (i.e. unit change in climate causes what change in planting?)

Oct 2

**Avery Meeting**

* For what might control crop date: agroecological risk zoning controls credit access in a certain planting range, so that’s a legal constraint. Gabriel has data; it varies by state and year. Phytosanitary law also changes by state and by year; and law will change to accommodate planting.
* The statistical model must account for whether growers are old or young, whether the farm is big or small. Assume it’s the same grower over a short time period. What happens when you control for farmer characteristics (fixed effects)?
* The map of dates that I produce shows when soy is visible to satellites, not exactly soy planting. Could have crop failure or replanting, so it’s not a perfect measure of decision to plant; it represents when there’s enough greenness to show up in the satellite – it represents planting date decision + something else. So the stat models should focus on variables that will change and that we know how they will change.
* Climate impacts RS indicators of vegetation (i.e. LAI), which impacts RS indicator of planting date, which leads to plant decisions. We are trying to get plant decisions out of climate, so we need to look at how errors/changes in the middle stuff. The errors introduced from RS indicators of LAI to RS indicator of plant is essentially the errors we’re looking at with the high res image of plant date, and the error from RS indicator of plant to plant decision can be deciphered from Matopiba. It can be crop failure rate, but also the impact of the (calibration) relationship between peak/greenup and plant/harvest due to whether the year was wetter, or due to temperature.
* The (plant – onset) model: what is it used for? It tells us the marginal effect of it getting hotter on the delay between plant and onset. It can also be used to estimate plant directly from plant = (plant – onset) + onset.
  + The onset helps decide whether the crop fails. So there are two climate factors that matter: the climate that matters to farmers versus the climate that matters to seeing green crop on the ground.
  + Farmer decisions and climate both impact to planting date, but parts of the climate impact farmer decisions. The planting date then impacts the delay between onset and plant. The climate also impacts the delay.
* For irrigation, use GEE code to map center pivot. Can train on 2014 data.
* For EDA, do it for local anomaly instead of absolute magnitude. This is because fixed effects will use anomalies, and want to do EDA as close to the way I’m estimating the models as possible
* To save time, do analysis of different models and see how they’re different instead of just doing a bunch of EDA. Pick model types and uncertainties that will influence them; pick tests to use to test each model specification
* Add a slide to decide within a model structure and also among model structures, for example compare OLS to random forest. Try model with subregions; model without subregions.
* Do moving window analysis to look at spatial differences – it’s a variant of fixed effects regression.
* Most importantly, update the model for plant and model for (plant – onset) into a coherent process
* Need to propose validation scheme
* Think about SIF as well in addition to Sentinel
* ALOS PALSAR
* Optical satellites have atmospheric distortion because even clear sky days have water vapor, and there are lots of aerosols from forest. So use SIF and PALSAR.

Oct 5

**Meeting with Sally**

It’s okay to do plant = fcn (onset). Perhaps test other definitions of onset, e.g. the date at which some threshold of total rain since Aug 1, or the date at which a storm of a certain depth is reached. Test different ‘other onset definitions’ against anomalous accumulation onset by doing a regression or scatterplot. We know that onset matters – but maybe test out another date.

For the fact that some crops fail: report the percent of farmers who say their crop failed in Matopiba survey. Look at pixels in a polygon that were bare soil or weeds for the entire growing season for an estimate of crop failture. Claim that we are only seeing a certain percent of the planting outcome with our RS method; only the successful crop.

Random forests: can’t impose a subset for random forest, so define subset in a table for the random forest input. Don’t do random forest; use it as a check for OLS (random forest may overfit the data)

To cope with the fact that response to onset is nonlinear: first do plant = fcn(T, size, etc) and get residuals. Then plot residuals on y axis and onset on x axis and see if onset is linear or not. Can also use separate models for early vs late onset but need to define when the threshold is for late onset (sensitivity test)

Subsetting: how do you know what spatial scale creates ‘stability’? look at different scales, plot the coefficient on the y axis and scale on x axis, look at whether there’s a flat part in the curve. If pick a scale, e.g. 50k, do regression for 40km and 60km as well to see whether the coefficients are the same, as a robustness check. Do subsetting hierarchically – within each state, divide into 200km cells, then divide the 200km cells into 50km cells.

For decision tree: if include it in prospectus, make sure to tie it to how we account for each of them in the regressions.

To test predictive ability (and therefore compare OLS vs FE), make sure the new year contains climate that was trained. Can try predictive ability in dry, wet, 50th percentile year, etc.

Can include a trend, if trend doesn’t exist it shouldn’t matter; the only issue is multicollinearity. Test for it by plotting individual variables against each other to see if they are collinear.

Oct 10

**Meeting with Avery**

* Surveys: representative sample. Matopiba is a convenient sample, conducted by NGO to get geographically broad region, but within each stratum of region didn’t make sure the interviewees were representative, didn’t talk to enough people. ‘Recall’ survey; didn’t go and measure planting. Prone to recall bias, and not representative.

Weather data:

* Every Xavier var except for wind is the best for Brazil (most accurate) – Avery will send paper. Xavier is missing a lot of high resolution temperature stuff that matters. Avery downscaled temperature to be 1km dataset, would be useful far from weather stations.
* Xavier precip only has up to 2015, do something else for 2016. CHIRPS? Look at Morgan’s paper about precip datasets.
* Onset definitions: try Abrahao approach (modified AA with threshold) – good approach for cropping. Screens out days with too little water depth for germination. Don’t explore a bunch of other onset descriptions.
* Delay in planting due to too much rain is probably not important; it’s more important for harvest and second crop planting.
* No temperature explanatory variable for temperatures too high
* Not all nonclimate variables can exist together (e.g. not latitude and unit fixed effects together)

Overall structure of model:

* Have a simpler theoretical framework to show that the relationship is true. Causal inference.
* Plant = a\*onset + b\*(planting window) + fixed effects
* Planting window:
  + Historical (past 5 years) of wet season length, temperature, photoperiod; current irrigation
  + How to define expectation of climate – moving window of years
* Planting window – include capital constraints too?
* Onset date, expectation of climate (a single index for how long a window you have to plant), fixed effects – do this for each combo of nonclimate vars
* Quantile regression – slope is different depending on onset, not as good for predictions
* Qualitative rule for when onset is the determinant of planting date in the tropics – so constraint of latitude.

General paper

* Want to know something about climate during growing year to see whether planting date must change due to onset; some people aren’t sensitive due to their growing season climate
* The nonclimate vars: include as sensitivity analysis. Use a dummy variable for different regions and put them in a single regression. Don’t try too many interaction terms.
* Onset and planting aren’t hard coupled. Compare agri productivity diffs under hard coupling and soft coupling under climate change. Use climate projections. Planting matters for exposure windows.
* Also say that the way crop modeling community is handling planting dates is inadequate.

Oct 11

**Meeting Paolo**

Trends in planting date and growing season?

Look at GCM outputs in Brazil; look at scenarios of precip timing for example to set as an example climate perturbation

He will ask:

* Climatology of the region
* How land use change affects climate
* Bonan – Ecological Climatology
* How deforestation in Amazon affects climate in Cerrado
* Ways that land use change can impact agri productivity (through climate AND not through climate)

Oct 15

**Meeting with Sally**

* First order: what’s sensitivity to onset?
* Second order: in addition to onset, are there other identifiable factors that change response to onset?
* It’s a breakdown into two questions. Will still do a lot of robustness testing.
* Start with onset question, and go into robustness check.
* Use SIF as measure of photosynthesis or as another vegetation index? – do research on what SIF actually represent.
* Ask Avery for resources that he thinks I should study to give me the basics.
* Ask people to attack my methodology during my practice.
* Respond to the accusation that we’re data mining. Data mining is that we have complex dataset, look for relationships – don’t have hypothesis that guides the work. But we have robust hypotheses.

Oct 17

**Meeting with Tina**

* Landsat and MODIS: why 8 days versus 16 days? Why are spatial resolutions different? Know a about Landsat and MODIS data products!
* What’s new about this work? What’s the broader context of the work?
* Ground truthing?
* How to combine Sentinel data with Landsat-Modis?
* How will I split this work into papers/chapters? Write this in the prospectus.
* Talk to a remote sensing person about the satellite data products

**Meeting with Ashok**

* Think about motivation for the project – for food security
* Think about climate models – how they are made, how they are validated
* Sentinel: fit harmonic function to the peak as well, just get rid of the middle part; fit only LAI = 1 to 3
* Understand ideas surrounding the research core
* Amartya Sen, Poverty and Famine