* The issue of pixel aggregation at the district level needs to be addressed, it’s a common question (came up during the workshop and we’ve had some email exchanges on this before).
  + Are you simply calculating the average of the pixel values at district level? Or some other measure?
  + Pixels with in area of interest (AOI), district boundaries in this case, can be evaluated on a pixel by pixel basis in utilizing parrallel processing or summarized by the AOI's mean value for each image. In this study, district level mean values of EVI are used to represent agricultural productivity for each 8 day period.
  + Is aggregation happening before calculating the summary statistics and integration measures then used as indep vars? Or these measures are first calculated at pixel level and then aggregated? It’s probably the former but these two methods could give quite different results
  + It can be done either way, in this case we find mean values for each 8 day period then run statistics on that.
  + In any case, some robustness analyses could be done to see if and how aggregation affects results?
  + This could take a while, so I will skip unless reviewers give us trouble.
* Related to the above, are the pixels used within a district always the same across years? Say, a district encompasses pixels 1 through 10. Now, first year pixels 2 and 5 have cloud cover. Then again some of them are removed through land use information (which I assume remains constant over time?). Second year, perhaps pixels 3 and 7 have cloud cover. But then the pixels used for the aggregation are different and differences between years 1 and 2 could (at least partially) come from this variation in the sample of pixels. This is important to address.
* Masking of land use changes varies by time, same with cloud cover. So the number of and ID of pixels used will change over time. Not sure there is any good way to get around this. However considering the use of 8 day composite images, most images will be ‘cloud-free’.
* You mention spline smoothing of EVI values + outlier removal. Are these two different procedures you are conducting? Or just stating that spline smoothing is implicitly removing outliers? If the former, some explanation would be good.
* To minimize the effects of the artifacts described above we test the use of temporal smoothing splines and outlier removal . Where outliers above three standard deviations are removed before applying a cubic smoothing spline.
* Overall, I think the introductory section on setting up all the variables to be used later in the model needs more explanation and a more clear description of the steps followed.
* Ok will try to improve
* Now, to the variables and the econometrics. The discussion of percentiles wasn’t very clear to me. You mention that you use historical percentiles in order to see variation with respect to some historical benchmark. But these are then constant across all years for a given district, right? Cause in the table it somehow seems that percentiles are calculated within a single year. Perhaps these percentiles could be used differently, for example rescaling variables in different districts as a percentage of their 95th percentile before the PCA? I think this would capture this historical benchmark a bit better.
* Explanatory variables. Some of them are clear but some are not, I think a short note is needed in the table briefly explaining how these are constructed or what they represent. In particular:
  + VEG\_growing\_max\_date. Is this the date at which the max EVI is reached?
  + VEG\_growing\_v2, there is no explanation of this v2 and how it differs from “v1”.
  + “Leading”, “trailing”, “diff”, what do these mean exactly?
  + There’s 3 yield variables in the table, you mention you use the first one as dependent, but then what are the others there for? I don’t think they are being used as independent, right?
  + R\_mx\_dates, what is this?
  + Some summary statistic variables could make sense, but some others make me wonder, like for example min or mean. In a way MEAN is = to AUC/N (with N=number of 8-day periods), right? Aren’t these variables co-moving almost identically? Also, min perhaps should be used in the same way as percentile, that is, as a benchmark for that district/pixel? For example, two districts, one of them has more/less water than the other, so its EVI value is consistently higher/lower. Then shouldn’t EVI be somehow rescaled to make them comparable across districts? One way of doing this would be to substract the minimum EVI value from all EVI values and then calculate all measures using this rescaled series…
  + Finally, around 40+ variables are listed in the table, but you mention you use 28 for the PC analysis, which ones are these?
* Regressions.
  + You use the 22 principal components in the regressions. I would probably use a lot less and see if the main PCs can explain yields reasonable well. What you are getting from the last ones is probably mostly noise, as they explain very very little variation. I am not super familiar but I believe there is some sort of rule of thumb to decide up to which PC to use.
  + It would be nice to see how the first few PCs are constructed. Perhaps there is a nice story behind them, with each representing a few main variables. This could help in teasing out what works best at explaining yield variation and it would be a nice addition to the discussion section.
  + What is the rationale for having the 1st PC lagged? That wasn’t clear to me and I wonder if doing that does not introduce some econometric issue… Also, in equation 2 you show PC#1 as lagged, and then PCs #2 to #22 as contemporaneous, but then in the R output it looks like PC#1 is both lagged and contemporaneous…
* The spatial autocorrelation is interesting and probably important in our case. However, I find it odd that while Moran I’s test shows spatial autocorrelation the spatial RE regression output you have at the end shows an insignificant spatial coefficient (lambda), which would imply that there does not seem to be spatial autocorrelation! Perhaps it’s an issue of the exact spatial matrix you were using or the method itself (I understand one can introduce spatial lags at the independent variable, dependent variables, or error term levels, though I am not sure if all of them have an RE estimator version…)
* Out-of-sample prediction would be a really nice addition, as I think many of the issues I’ve raised ultimately link to this. Two alternative procedures could be (1) a temporal out-of-sample (calibrating the model with the first, say 8, years, and forecasting the last 2 years for every district; and (2) a “spatial” out-of-sample, calibrating the model with N-5 districts and then predicting in those 5 districts for the full 10-year period. On specific metrics to assess the goodness of this forecasting I am not sure to be honest.