**Meeting with Gergana** – 16 January

* focus on habitat loss/biodiversity loss rather than effects on climate
* add non zotero references to zotero by inputting info
* add Background sub-heading and combine 3 short paragraphs under one background
* add more concrete topic sentence at start of document
* my prediction is that you can detect the effects, but it is already a well-recognized fact that land use change and socio-political events can be linked
* can frame it like: can you see the signature of this happening from space?
* This study is important for understanding the rate of change and variation across country
* get the CORINE data set to get more classifications (with more classes) after to get more detailed results
  + then can say - from looking at other data, these intensive land classes used to be X class
* LUCAS data set to ground truth classification - use for validation
  + Is field data with abandonment
  + find points of abandonment and compare with visually and statistically to assess accuracy of classification
* join GEE user group

**Meeting with Gergana** – 23 January

Questions beforehand

* If I’m coming from a satellite perspective i.e. is there a marked signature notable from space, should I be focusing the background on satellites? Or why this is important in terms of land-use change? Or both?
* Should my questions allow for different answers after each event? Or is this implied?
* Need to change to more hypothesis testing framework
  + Was thinking of changing more to “how/what” questions but there are no clear answers to this – may make my study less quantifiable
  + E.g. Through land-use change, how marked are key socio-political events in Latvia as detected by satellite imagery?
  + E.g. What is the strength and direction of land-use change for each land-use type?

Notes

* What the results would mean e.g. if there is a signature, it means this…
  + Doesn’t mean we have to change the question
* Is this driver so strong that you can see it for the whole country using satellite data?
  + Is this homogeneous or heterogeneous – implying that the results are different and having different effects
  + Signature is homogeneous – socio-political is main driver
  + Heterogeneous – fine scale variation, socio-political is not main driver or it’s interacting with other aspects
* Overarching question and then specifically look at how (1), (2) and (3) etc.
* Prediction for each question and give reasoning
  + Hypotheses don’t give reasoning, just a statement of what you think will happen

**Meeting with Isla** – 28 January

* Goals for Friday:
  + Finish dissertation plan
  + Do a few GEE tutorials
  + Load MODIS, CORINE, LUH into GEE as different layers and take a few screenshots – can ask Gergana for starter code here
* Better to fully answer first question, then move on to next etc. so can remove a question if I run out of time rather than answering them all poorly
* Should have a full answer to Q1 by end of Feb (as in the result)
* Should write methods as I go
* Prediction figures are a good idea – make in PowerPoint
* Continually work on Introduction
* Set goals in the beginning of each week – super ambitious ones and then ones that absolutely need to get done

**Feedback from Gergana on full plan to consider in future**

* Use active voice
* Make sure first and last sentences are stand-alone and link together
* Can make a conceptual diagram about objectives
* Even if one reference has everything, use a variety of new and old papers
* Potential to make hypotheses even more specific/directional – something to think about
* Split up methods by sub-question to make clearer
* Add model equations

**Meeting with Gergana** – 6February

* How to load LUH dataset – historical states.nc
* Need to include the different stages of abandonment
* Don’t use fusion tables, use import asset as a shape file
* Draw around a few points of abandoned land at a time and see what happens

**Meeting with Gergana** – 27 February

Questions beforehand

* Is this the best choice for getting the agricultural data points? seems like best choice is to use 2012, 2015 and 2018 – Use U410 and filter for the B, D and E land cover (cropland, shrubland and grassland). Also, can filter for U112, D10 and D20, which are stated to be abandoned agricultural land
* Is it okay that the CRS won’t work/assign properly?
  + Think it is fine because the default projection in GEE is WGS84, which is the same as LUCAS
* How do I set the boundary? In the tutorial, they had another file that they needed to combine with but that’s not the case for me
* Feedback about prediction figures

Notes

* CRS("+init=epsg: 3857”): specific for GEE
* atc\_m <- projectRaster(atc\_mar, crs = "+proj=longlat +datum=WGS84 +no\_defs +ellps=WGS84 +towgs84=0,0,0") 🡪 change projection
* bt\_spatial\_df\_m <- SpatialPointsDataFrame(coords\_m, bt\_coords\_m)
* set extent as boundary of Latvia
* exclude any non-natural for U410
* U420 – spontaneously re-vegetated land (land cover) 🡪 add this
* Do trial about just 2012 data and then if time/can do, add 2015 to see change accurately if so
* Feedback about prediction figures in PowerPoint

**Meeting with Gergana** – 6 March

Questions beforehand

* Something has gone wrong with my classifier – abandoned class covering everything! – potential ideas
  + When troubleshooting – thought it may be because I only have one class – was getting errors saying no training data found
    - Changed class value from 0 to one and now everything is abandoned
  + Think it has to do with when I say for it to get training data – no columns (but this was the case for the tutorial I completed on classification?)
  + Could be that the polygons I created represent all bands? But how would I know if this is the issue?
  + Could it be the way that I’m visualizing it that is an issue? But then still one class would be covering a lot of area
  + **Because I’m creating a classification on one class?** 
    - Create an everything else dataset – how do you do that?
* How to apply classification to each year – especially if need to switch to different satellite imagery 🡪 image iteration? <https://developers.google.com/earth-engine/ic_iterating>
  + Would you create a list of images and then iterate through them?
  + How do deal with different pixel size?
* Extensive and intensive datasets – want to confirm that I will be using CORINE?

Notes

* Make a forest class/natural vegetation for another class
  + Something that’s different but similar to abandoned land
* Resolution as fixed effect in analysis but no way to correct for that as of now
* Can map LUCAS on CORINE to see if what you think in LUCAS is intensive/extensive is correct
  + See papers using CORINE/LUCAS for advice
* Conceptual diagram of project
* Diagram of workflow
* Writing 100 words a day
  + Then get feedback as you go and not so overwhelming
* Can write code for graphs as you go as well and for stats models
  + Will help see what results you need as well
* Five key paragraphs of a paper -- on Team Shrub mentorship page
  + Perfect these!
* Perth is 8 hours ahead of time 🡪 Skype at 9am would be 5pm Perth time, or 8am would be 4pm
* <https://code.earthengine.google.com/309ed9d52a82249a21b106798fea495b> - Gergana’s code about going through many images

**Meeting with Isla** – 14 March

Questions beforehand

* Do I need to correct more for clouds?
  + Could do something like choose the pixels with the least amount of clouds

*There are ways to remove clouds, but the easiest thing is to work with versions of the satellite data that already have the clouds removed.*

Find Water and Remove Clouds with Fmask on Google Earth Engine

<https://medium.com/upstream/find-water-and-remove-clouds-with-fmask-on-google-earth-engine-261c090cb62d>

* Can CORINE be used as a validation set, because it in itself is a classification? Maybe it would be best to use LUCAS (2015) as a validation set? Is this possible even though if it’s for a different year?

*Any land cover dataset is based on some sort of classification. Let’s go back to first principles with this one. What is your question? What datasets will help you answer your question? Do you need to “validate” or test out different land-cover classifications? Or do you want to compare different classifications? What is your “real world” dataset that you would compare to?*

* Can I just assume that my points loaded correctly, because a lot of things ride on the fact that this is true, which is a bit scary? Some of them were in water bodies which is obviously not correct.

*What are your points? What is the precision on the points? There could be error in the points or in the underlying data and whether the water bodies are accurately detected depending on what land cover/imagery you are using.*

* Do I need to divide up Latvia into sections if I’m looking at pixels/area covered? If so, how should I go about doing this?

*This depends on your question. Does your question have a regional element to it? Or maybe are you wanting to ask how land cover changes within broader land cover classes? Or maybe everything is at the country level?*

* Should I make an ‘other’ classification? Not all land will be abandoned, extensive or intensive or forestry (which I use as my other category when classifying abandoned). I think this would improve accuracy but maybe there’s a better way to do this.

*I think you can do this. Up to you. You could even want to try to classify multiple other classes.*

* How to apply classification over time: 2 parts to this question/issue and I think basically it is necessary to find an innovative way to do this to prevent error and promote organization to get data for each year

1. How to extract the pixels for each year ---
   1. When I try to do this, I only figured out how to do it by loading a chart and then exporting the underlying data as a csv – surely there is a better way to do this, but I can’t find it so far. When I do this for more than one year, GEE runs out of memory and can’t display the graph and therefore can’t get the underlying data

*I have found that it is better to export the data as csvs or tifs and not use the graphing functions. They used to work and now they don’t work as well.*

* 1. Would be good to find a way to save it all into one csv to prevent error and save time and memory

*You can either set up a function to do this, or you can explore each year’s data and then combine those files in R. The first might be the better way to do it, but I would focus on what is easiest for you to do first. You can always go back and try to improve the workflow later.*

* 1. Maybe also this is not a good way of doing it because it would be good to have each pixel value to see what pixels are turning into what i.e. is intensive land coming from abandoned, extensive or neither

*You can also be exporting your tifs and then doing the comparisons in R of the land-cover change. I personally find this approach to be easier as R is much easier for me.*

1. How to apply classification for each year –
   1. Prone to error if do this manually and run out of memory very quickly doing this in the same script

*You should be able to subset to the specific data you want to use, run the classification and then output the results. You can do this individually for each year. Or you can set up a function to do this across years, but this is trickier programming.*

* What other datasets should I compare my classification to? To me, it makes more sense to compare it to CORINE/other classifications, rather than using these classifications as validation data. How do I assess accuracy of other classifications though? Is it more that I would see how the areas overlap, but can I really determine which one is more accurate?

*Do you have actual ground data to compare to? I think you need this to do any accuracy tests. Are there in situ data you can use to compare to all of your classified land-cover data?*

* If you look at my accuracy/error spreadsheet, why is my accuracy going down with more points? Could this be normal or does this indicate that I’ve done something wrong?

*I think generally there is an optimum with these things. The classification “success” goes up for a while and then starts to go down as your points start to encompass more variation. The more points the more “accurate” to the real world it probably is, but the classification becomes harder as the groups become more similar and there might be more “noise” in your classification. There is some sort of trade off here.*

* I’ve set the border of Latvia as the bounds of the Landsat image, but it doesn’t actually seem to be applying the classification just to Latvia 🡪 is this affecting the area size?
  + When I downloaded the csv of the pixel area of each class, it did not add up to the total area of Latvia – have I calculated something wrong or perhaps it’s the bounds that caused this?

*Generally, the Earth Engine will apply any classification to any area it can see in the map window unless you have indicated otherwise. For your project, I would just set the window to a consistent zoom with the same centre point and work from there.*

* How do you recommend determining if the land is extensive or intensive?
  + Option 1: Use LUCAS again and make own decisions about whether land is extensive/intensive (see decisions document on GitHub). I so far did this and split it by industrial agriculture versus “kitchen gardens” which is basically allotments. However, there are a lot of industrial agriculture points as compared to the extensive ones. Potentially I need to refine more and filter by irrigation type?
    - I tried to compare my points visually to the CORINE dataset, but when I loaded this, the band will not display in the inspector tab. Do you know how to fix this?

*I could try to trouble shoot with you. It is probably a visualization scaling issue?*

* + Option 2: Split up points by nitrogen input using the CAPRI dataset – Common Agricultural Policy Regionalised Impact dataset (Temme and Verburg, 2011 had methods for this), but this seems a lot more difficult. I tried to download the data, but it is in a GDX file. I spent about an hour trying to open it and it wouldn’t work. The methods seem a lot more complicated, but could be less subjective?

*This seems a bit more indirect to me and if it is “hard” maybe it isn’t worth it? But what ever you think best. Maybe there are other nitrogen input datasets already in the Earth Engine?*

* I filtered the date from the start to the end of the year, but I saw somewhere that you could take a median of this. Would this be better? Is it inaccurate to filter the date for a whole year span, as there are multiple images in this?

*This is the stuff that will really matter to your classification. I think you want to set up for yourself a really clear justification for the imagery that you use and you will want to devote text in your methods to clarifying why you chose what you chose. As long as they are based on good rationales or other existing data, I think you will be good. You could also choose to take a couple or more different classification approaches with different data and ask if they produce similar or different results. This might be an important part of your story.*

* Should I include fallow land at all (land set aside for more than a year as part of field rotation or set aside long term)? If so, how would I determine if this was extensive or intensive, or is there a way to see if that pixel was intensive/extensive the previous year and apply it to that year?
  + There is a temporary grassland category that is already included in my datasets

*Not sure about this one. It will definitely be “extensive” from first principles – but could overlap somewhat with a temporary grassland category.*

Izzy’s notes from meeting

* Do classification first on POINTS rather than polygons
  + If have time, add a buffer – this is still accurate as it will pick up on nearby farms (remember that Landsat itself is not the finest scale dataset and details are still missing from pixels)
  + If do a buffer, construct code to determine what is the most dominant land-use type aka pixel type
* You can subset the dataset to split it for validation points as well, just like how you do for training and testing data
  + Here, I can choose whether to include points from other datasets i.e. CORINE points or make own points by looking at imagery (to me it makes sense not to but need to justify this)
* For intensive, may need to compare two time points of data to see the difference i.e. between ploughed and grown crops
* Try to get data for a lot of years, but it is most important to get it for years around the events
* Get each year classification data as a tif and work on it in R
  + If you change the resolution when exporting this file, GEE does something called pyramiding where it alters the classification based on the new resolution – if needed, look into how this influences your data and what’s going on
* Use pre-cleaned Landsat data, that accounts for clouds as well as the change in resolution between different Landsat data
* Run the script one year at a time and replace the year in the script rather than trying to have GEE execute multiple years at once, just do this manually
* Don’t classify over the whole year – pick a time period sing ecological reasoning i.e. a window of time like summer months and take the median
* I can compare my classification to CORINE to assess accuracy and see similarity – if both my classification and that one portray the same idea, it strengthens my argument
* Revisit prediction figures based on notes in the PowerPoint
  + Figures for each sub question with a hypothesis, null hypothesis and alternate hypothesis
* Do more reading to make decisions i.e. how best to classify agriculture
* ALWAYS GO BACK TO THE QUESTION and think is what you’re doing important and necessary to answer the question – if not, rethink what you are doing!
* Put weekly goals as an issue on GitHub and meet them!

**Meeting with Gergana** – 20 March

Questions beforehand

* Is what I am doing *enough* for validation? I’ve read conflicting things online about how validation differs from testing a classification
  + I’ve seen people do things like cross-validation and *k*-fold validation, but I’ve had trouble figuring out how to implement this to my code and also don’t understand if this is necessary
* What is the benefit and what was the intent behind me using LUH? I remember talking about this now, but I don’t see where it fits in and how I was meant to employ it, to be honest
* Is taking the median enough for cloud cleaning? I’ve seen a lot of people do this, as it removes clouds and shadows
* I was reading about the Landsat 7 collection and how it has a lot of imagery lost. Do I need to do something special to account for this? I have one study year that would not be covered by Landsat 5/8 if I didn’t use Landsat 7, and unfortunately, that year is 2012, which is the year I am using the LUCAS data for
* Isla and I were talking about, when I’m done with the classification, just changing what imagery I put in the start and downloading the data that way. After thinking more about this, I want to check that it is accurate because surely it would be re-doing the classification on imagery that represented a year different to my points? I was wondering what you think about this.
* I added in a bunch of other classes to try to help improve accuracy, but my accuracy went way down (over 90 to ~64%). I’m not sure why this has happened. Do you have any ideas on how to improve it?

Notes by Izzy

* Don’t worry about validation right now
* Compare classification with CORINE first, then LUH if time
  + Unless you think LUH is better – which may be the case because it covers more time
  + Can compare CORINE with one point in time rather than the whole-time scale
  + LUH can do a whole-time scale
* Try to find the pre cloud cleaned Landsat
  + Processed Landsat layer/composite
  + Ask Isla if for link if can’t find
* Check which areas of the world that Landsat 7 messed up over
  + See what data are exactly missing
* Need to Google how to carry a classification over a different year
  + Ask on the user’s form
* Compare LUCAS and CORINE
  + Could ask for specific photos for each type from LUCAS to validate points
* Filter LUCAS points by CORINE points in R
* Use CORINE/LUH more to back up my story then to compare the classifications
* Visually compare maps

**Meeting with Gergana** – 27 March

Questions beforehand

* Is using the Surface Reflectance dataset for Landsat better than the top of atmosphere one? I think so but want to confirm
* When I tried to filter LUCAS by CORINE, I realised that the CORINE projection is very different (EPSG:3035 (ETRS89, LAEA). When I tried to change it to the same coordinate system as LUCAS, it doesn’t change to LAT/LONG. Do you know how to fix this?
* My area calculation export doesn’t load even if I try doing only one class and a coarse scale – how can I combat this? Also, my accuracy and error exports won’t load either, or if they do, they take several hours. Is there any way to speed this up?
* I extracted the data for 2012 and formatted it in the lucas\_to\_shapefile script. My idea is to create an index that numbers each point. I would like to make a function that, when I include another year’s data, would put the same index number if the lat and long matched. Do you know how to do this? Do you think how I’m formatting it is a good idea?
* General feedback on what I said in the email before

Notes

* Pixel scale or overlay Latvia with hexagons of the same area in R
  + Piece of code that groups by cell
  + Change in dominant land-use
* Workflow diagram for each question
  + With syntax for models
* 1985average class, 1991average class and test column
  + In test, see if they are the same
  + df$test <- df$class1985 == df$class1990
  + Then filter for records that say false
  + Then look for the transition 🡪 paste two columns together
  + df2$transition <- paste)df2$class1985. df2$class1990, sep = "-")
  + Then group by transition type
  + group\_by(transition) %>% tally()
* Area of classes in R with cells
* Could transform the LUCAS points to CORINE
* Extract function in R
  + classes <- as.data.frame(raster::extract(corine\_raster, your\_points)
* Surface reflectance vs top of reflectance
* Time lag
  + For each year, how much area underwent transition 🡪 see when peak occurs
  + Year with largest area for each transition
  + How many years before abandoned to intensive e.g.
    - time (years) ~ 1 – number of years
  + time lag, don’t need what’s before – always relative to one point
  + time lag ~ event (Soviet vs EU)
* Screenshots of classification for key years
* Put images right in email to Isla

**Meeting with Sam** – 29 March

* use Landsat 5 for 2011 for the training points because unlikely to have changed
  + miss out 2012
* so just add a new image at the end, don’t do new training points
  + var classified = clipped.classify(trainedClassifier, 'classification');
* just export
  + may need to lower resolution to save on space
* validation accuracy would change by year so export
* change resolution if want to speed it up
  + depends on accuracy needed
* bigger error if more cloud
  + can look through all images of a collection
* Landsat 5/8 have different band widths
  + have to figure out how the bands correspond
  + rename image for list of bands []
* don’t want the atmosphere bands
  + look at the land ones - red, blue, green, near IR, longwave IR
* TOA or SR is fine just keep uniform
* Error may be that agriculture looks different from now and in the past so need to say that this could be an issue but not really a point to account for it

**Meeting with Isla** – 12 April

Questions beforehand

* I feel like I may have an error in my code
  + I have a buffer around each point of 90-metres, but I have not specified in my code to pick the most dominant type 🡪 is this an issue? I feel like it is, as some extensive/abandoned land may be under 90 m2 and this may explain the extreme amount of intensive land 🡪 should I take out the buffer or make it smaller?
  + I don’t think I am getting training data proportionally by class – could have an impact as there are only 25 extensive points
  + Running everything again takes over 15 hours – are these things for the limitations or is it worth running it again?
* Normality:
  + A lot of zeros in transitions i.e. when a transition doesn’t occur – when I set these as NAs it is still very left-skewed as lower amount of areas are going through transitions more often than higher
  + No zeros for total class area but lots of low values with few high – potential to remove some with high cook’s distance?
  + How to assess normality of hierarchical data? Filter per year and look at year data?
* Slope is so high because in m2 working in area units. Indicates the strength of the transition
  + So, if I want that to be smaller can change to bigger units (kilometres?)! 🡪 can scale log? – this may help with normality
    - Scale() to standardize explanatory variables – explanatory shouldn’t be the issue though
* Cell = cell ID – for each cell there is the amount of pixels per each land use type and per each transition so each cell is composed of pixels – over 150 cells so fine for random effect BUT I don’t have the pixel scale data so I just have cells across years – to my knowledge this is okay for a random effect?
* Go through models/script
* Go through figures

General notes

\*\* other notes in script!

* Acknowledge buffer could do mini study for appendix
  + Same for proportions of classes if time
* Can have different levels of confidence when explaining results
* Need to clarify all questions so that its clear what statistical test directly follows – should be clear what data is needed and what kind of result there will be
  + For each main question have a general question then have sub questions that are hypothesis questions – basically framing, is my hypothesis correct?
* Think about using proportions e.g. % of each land use type per grid – will this help the reader?
* No need for model selection
* Can get pseudo p-values for LMM – package for that
* Report sample size, model structure, slope, error around slope, t statistic, p-value estimate, R2 estimate
  + Look at error in relation to slope
  + Make sure that you take p-value and r2 with pinch of salt
* Create dataframe of outputs to make an effect size graph
  + Tidy package to apply to outputs
  + Height of the bar = slope
  + Error on the bar is the error + slope and error minus
* Note that geom\_smooth doesn’t work for hierarchy!
  + Need to be careful when visualizing hierarchical data
* Assumptions
  + Look at the normality/heteroscedasticity on the biggest grid cell level
  + Visualize by facet with grid cell and land use
  + If don’t meet assumptions still after changing models, talk to Isla
  + Examine visually and see how much of a problem it is for each sub-category
  + If log transform to fit normality or assumption, can back transform the slope/error so that it can be visualized effectively
* Can still plot over time to visualize but this isn’t visualizing the model – it’s just visualizing the data
* <https://stackoverflow.com/questions/36420367/difference-between-using-bquote-and-expressionpaste-for-labels-in-ggplot2?rq=1>
* If time, can think about doing a break point analysis (see coding club time series tutorial) – for time lags
* Consider scaling area to make everything the same scale/comparable – can play around with this for graphs – may help visualize

TOMORROW

* Bigger grid
* Aggregate by that
* New stats
* New vis