1. **Introduction** 
   1. **Objectives and rationale**

This study aims to investigate the importance of SPE events as drivers of agricultural LUC in Latvia through the use of satellite imagery. Although the importance of SPE events on LUC is acknowledged (Prishchepov *et al.*, 2012; Vanwambeke *et al.*, 2012), it remains unclear whether a signature of agricultural LUC is left across a country. By using satellite imagery to determine the strength and direction of agricultural LUC and transitions, patterns of a changing landscape can be unraveled to unveil when the effect of SPE events is most visible. My findings will give insight into the homogeneity, or lack thereof, of the effects of SPE events across a Latvian landscape. Results will reveal the transition patterns between each agricultural land use type, including extensive, intensive and abandoned land, providing insight into the relationship between land uses. Ultimately, my study will uncover the importance of SPE events as drivers of LUC in Latvia, permitting predictions about land use under changing SPE conditions to be made.

* 1. **Research questions and hypotheses**

**RQ1: Across Latvia, is intensive, extensive and abandoned land use change visible within the three years following SPE events?**

**H1**: Directly following SUC, intensive agriculture will decrease, and abandoned land will increase across Latvia. Extensive cover change will be weakly positive. After EUA, intensive agriculture will increase, and abandoned land will decrease across Latvia. Extensive cover change will be weakly negative.

**H10**: Intensive, extensive and abandoned land use will not visibly increase or decrease in Latvia directly following SUC and EUA.

**RQ2**: **Do the strength and direction of land use transitions change within the three years following SPE events?**

**H2**: Following SUC, intensive land will transition to abandoned and extensive land, with the transition to abandoned land being stronger. After EUA, extensive and abandoned land will transition to intensive land.

**H20**: Land use transitions between intensive, extensive and abandoned land will not significantly increase or decrease in Latvia following SUC and EUA.

**RQ3: Is there a time lag between SPE events and the visibility of LUC and transitions?**

**H3**: Following SUC, there will be a lag in extensive land cover increase, but not in intensive cover decrease or abandoned land increase. A lag will be present in the transition from intensive to extensive land, but not in the transition to abandoned land. After EUA, a lag will be present for the decrease of extensive agriculture. There will not be a lag for abandoned land cover decrease, and intensive land increase. A lag will not be apparent for the transition from abandoned land to intensive land, but a lag will be observed for the transition from extensive to intensive land.

**H30**: Time following SPE events will not have an effect on strength of LUC and land use transitions.

* 1. **Predictions**

**1.3.1 SUC**

I predict that abandoned land will increase visibly following SUC, due to the sharp decline of the agricultural sector (Prishchepov *et al.*, 2012). I predict that intensive agriculture will be replaced by extensively farmed land due to a large scale shift from Soviet collective farms to small-scale subsistence farms (Vanwambeke *et al.*, 2012). However, I predict that uncertainty around land access during the post-Soviet transition period will cause a lag in the increase of transition to extensive land, resulting in a large increase in abandoned land within two years following SUC. I predict intensive agriculture decrease will be observed directly following SUC, representing the rapid decrease in government investment in agriculture (Vanwambeke *et al.*, 2012) contributing to the high levels of abandonment (Prishchepov *et al.*, 2012).

**1.3.2 EUA**

I predict an increase in intensive land will be visible directly following EUA, as public support has facilitated the increase of agricultural production and income (Veveris and Kalis, 2016). In turn, I predict that abandoned land decreases within two years of EUA to facilitate forming large, intensive farms. I predict that although extensive land will decrease due to difficulty adapting to working within a single EU market, it will not happen directly following EUA. I predict instead that there will be a time lag on this decrease and transition, representing a desire to maintain traditional farming practices (Nikodemus *et al.*, 2010)

**1.3.3 Implications of results**

If areas of a land use increase or decrease significantly directly following an SPE event, the SPE event can be seen as the main driver of LUC. Significant LUC also would indicate a homogeneous shift across the Latvian landscape. If no significant LUC is seen, it is likely that different regions experienced LUC with differing strengths and directions. Such heterogeneous effects would also signify that the SPE event was not the main driver of LUC or that it was coupled with other key drivers.

**2. Methods**

To answer my three main research questions, I constructed a classification of land use change in Latvia between 1989 to 2011 using GEE (Gorelick *et al.*, 2017). My workflow diagram, depicting the key steps to this process, is shown in Figure 1.

**2.1 Study site**

Latvia (55º40’-58 º05’N and 20º58’-28º14’E) is on the Baltic coast, in North Eastern Europe (Prieditis, 1993). Latvia borders Estonia to the north, Lithuania to the south and both Russia and Belarus to the east. Latvia spans 64.6 thousand km2 and is largely flat, with the majority of terrain between 40-200 metres above sea level (Prieditis, 1993). Land is vegetated for 180-200 days annually (Prieditis, 1993). Latvia was a Soviet state during 1945-1990, where there were predominantly large, homogenously cultivated farms on favourable areas, with the rest of the land largely forested (Vanwambeke *et al.*, 2012). Following the SUC in 1991, land reform aimed to restore pre-Soviet farming culture when extensive, small farms were key to Latvia’s culture. Leading up to EUA, agricultural support schemes played a large role in promoting such extensive farming practices (Vanwambeke *et al.*, 2012).

**2.2 Image processing**

Landsat 5 Thematic Mapper satellite imagery (1985-2011) covers my study period well and is commonly used in similar classification studies (Prishchepov *et al.*, 2012; Fonji and Taff, 2014; Sidhu *et al.*, 2018). I selected Landsat 5 Surface Reflectance imagery (30 metre resolution), which is atmospherically corrected, preventing the occurrence of clouds and shadows in the imagery (Zanter, 2018). For each study year (1989-2011), I employed an additional cloud mask to remove any remaining pixels containing clouds or shadows, as well as any edge pixels that do not contain all bands of interest. I chose summer images, depicting the growing season, to best characterise the spectral signatures of my different classes (Fonji and Taff, 2014). I took the median of each year’s image collection to an annual composite for each study year (Pimple *et al.*, 2018). I cropped each year’s image to the border of Latvia to increase the speed of my classification. Lastly, I selected blue, green, red, near-infrared and shortwave infrared bands, each with corresponding wavelengths, for my classification (Pimple *et al.*, 2018). My chosen bands will act as predictor variables for my classification.

**2.3 Classification background**

Classification approaches create categorical datasets, such as land use (Horning, 2010). Classifications aim to investigate the relationships within a group of objects to determine if the data can be summarised into classes (Gordon, 1999). There are three key steps to a classification: train, validate and test (Suthaharan, 2016).

In a supervised classification, relationships are discovered using data of known values, also called training points (Albalate and Minker, 2013). Training points train the classifier to infer prediction rules that form a decision tree (Albalate and Minker, 2013). Decision trees are composed of binary questions which group data by specific characteristics, such as bandwidth (Strobl *et al.*, 2009). With each step of the tree, the classifier asks whether a criterium of a specific class is met. After the classifier is trained, it can be applied across data of unknown classes. Test data of known values, normally obtained in the training phase, are then passed through the classifier to determine classifier accuracy and error (Suthaharan, 2016).

Random forest classifications are a supervised learning technique which is composed of numerous decision trees, creating a forest of trees (Horning, 2010). As multiple trees decide the class of each object, a voting process begins, where the final class assigned is the one that is predicted most (Horning, 2010). Using multiple trees increases classification accuracy (Suthaharan, 2016), as there is decreased overfitting, where noise is used to predict classes (Horning, 2010). A separate validation set, where overfitting is accounted for and parameters are pruned, is therefore not required (Suthaharan, 2016).

**2.4 Training data**

My training data represent the known locations of each land use type in 2012. Field data results in high certainty in the training set (Millard and Richardson, 2015).

I used the LUCAS dataset (Eurostat, 2013), which contains GPS coordinates of both land use, as defined by the socioeconomic activities and land cover for European countries, for every three years since 2009. I chose to use 2012 data, as it is the earliest dataset that clearly separates fallow and abandoned agricultural land. Within R statistical software, version 3.5.3 (R Core Team, 2019), I filtered data to represent each of my land use classes: abandoned land and extensive and intensive agriculture. I also included four additional classes (artificial land, wetlands, water and forestry), as a larger training dataset improves classification stability and accuracy (Millard and Richardson, 2015). This collection of classes is supported by a similar study conducted by Fonji and Taff (2014).

To select which data would form each of my classes, I used the LUCAS 2012 Technical Reference Document (Eurostat, 2013), which contains descriptions of each land use and land cover category. Table 2.1 displays the criteria that define each of my key classes. The criteria to determine additional class can be found in Appendix X.

Following filtering, I had the coordinates of a total of 4013 data points, 1434 of which are from my three key classes (Table 2.2; Appendix X). I transformed these coordinates into a spatial file by setting the projection to that of GEE: Pseudo-Mercator WGS 84 (EPSG: 3857).

Once imported into GEE, I created a 90-metre buffer around each point, turning point data into circular polygons (Figure 2.3). As each point represents an area or polygon of land, a buffer was necessary to compensate for GPS precision (Wulder *et al.*, 2005) and ensure the surrounding pixels were included. I chose a 90-metre buffer, creating a circular polygon with an area of 25446.9 square metres. This buffer size was used by Sader *et al.* (1995) for their land use classification. It is further supported by the average farm size in Latvia, with more than 91% of farms being this size or larger in 2003 (Zdanovskis and Pilvere, 2015).

To set up my training data for my classification, I divided my data randomly by class into two groups: training (80%) and testing (20%), also known as a hold-out approach (Suthaharan, 2016). The hold-out approach permits assessment of how the trained classifier works with unseen data (testing group).

I then sampled the imagery to collect the bandwidths across the locations of my training set. The bandwidths correspond to the unique spectral signature of each land type. As the training set refers to 2012 data, it would be most suitable to sample 2012 imagery. However, Landsat 5 only contains imagery until the beginning of 2012 and does not cover the summer months that my data corresponds to. Landsat 7 could have been used, however, 22% of images are missing due to instrument failure (Alexandridis *et al.*, 2013). When examining error, it was clear that this failure had an effect on images covering Latvia. I therefore chose to substitute 2012 images with 2011 images. Consequently, my training points will not precisely match up with the imagery. However, large-scale land use change is not expected to occur between 2011 and 2012, as this period was a stable time for Latvia’s economy and therefore, land use was kept constant (Skribane and Jekabsone, 2013). For instance, abandoned area coverage changed less than 0.1% between 2011 and 2012 (Arika and Mazure, 2017). Upon retrieving the unique bands, the classifier was ready to be constructed and trained.

**2.5 Random forest classification**

There are two main inputs when constructing a random forest classifier: the number of decision trees and number of predictor variables per node of each decision tree. I set the number of decision trees to 30 with consideration of the computational burden on GEE (Nomura and Mitchard, 2018). The default number of predictor variables used at each node to form the binary rules within the classifier is commonly set to the square root of the number of input variables (Gislason *et al.*, 2006). Here, the number of input variables is the number of bands (6) used to classify the image. Limiting the number of predictor variables reduces the computational complexity, as well as the correlation between decision trees (Gislason *et al.*, 2006).

I trained the classifier using the bandwidths gathered from sampling the input imagery and applied this classification across Latvia. For each year, I applied the trained classifier to the new imagery. Each classified image was exported as both an image and a GEE asset at 30 metre resolution. Exporting as an asset allows for data to be used in other scripts by the same user, which was important for preventing GEE exceeding memory capacity during data collection.

**2.6 Classification accuracy and error**

I tested classification accuracy and error in two ways: (1) the resubstitution method and (2) the hold-out method. Accuracy was obtained as the percent of pixels classified and error was given in the form of a confusion matrix. A confusion matrix compares the ground truthed data to the classification output on a class by class basis (Murayama and Thapa, 2011) and is an indication of the classifier’s performance (Dougherty, 2013). Table 2.3 represents a conceptual example of a confusion matrix, where the features that were classified correctly and incorrectly are reported.

**2.6.1 Resubstitution accuracy and error**

The resubstitution method tests the classifier’s ability to resample bandwidths that were used to train the classifier originally (Verbyla and Litvaitis, 1989). Resubstitution is known to produce optimistically biased results on the overall accuracy of the classifier (Verbyla and Litvaitis, 1989), therefore, the results of the leave-out method should be considered more representative.

**2.6.2 Hold-out accuracy and error**

The hold-out method makes use of the test set left aside when forming the training sample. This process depicts the efficiency with which the trained model classifies unseen data (Suthaharan, 2016). I resampled the input imagery using my test set, ultimately reclassifying the image. From this, a test accuracy and error matrix can be obtained.

**2.7 Data collection and processing**

As my classification is computationally intensive, memory constraints prevented both the classification and the calculations to be completed in the same script. I imported each year’s classified image and applied a grid composed of 124 equally sized cells (689.5 km2) onto Latvia. Collecting data by cell prevented exceeding the memory capacity of GEE and improved the replication within my study. A larger grid of five blocks was also imposed manually using R statistical software to allow for regional examination.

**2.7.1 Area**

For each year, I collected the number of pixels per cell of each of my three key land use classes. I saved each file as a CSV and aggregated them using R statistical software, version 3.5.3 (R Core Team, 2019). I calculated area in square kilometres using the following formula:

**2.7.2 Transition**

To calculate the number of pixels undergoing a land use transition, I first outlined the transitions I needed to gather data for, shown in Table 2.4.

The process will be explained using the abandoned to intensive transition as an example. First, taking the first year’s imagery, solely land that is abandoned is selected. Using the next year’s imagery, only intensive land is selected. I then overlaid the intensive land imagery atop the abandoned imagery. With the *and* command in GEE, I was able to select pixels that were both abandoned land in the first year *and* intensive land in the next year. This process was completed for all transitions for a yearly time step, meaning from one year to the following year, for each of my study years (1989-2011). I saved each file as a CSV and aggregated them using R statistical software, version 3.5.3 (R Core Team, 2019). I calculated area in square kilometres using the same formula in section 2.7.1.

**2.8 Statistical analyses**

All data analysis was performed using R statistical software, version 3.5.3 (R Core Team, 2019). Examples of key chunks of code used is available in Appendix X. Complete scripts of all code used to complete this project is available on GitHub (https://github.com/izzyrich/dissertation).

**2.8.1 Data setup**

To test my three questions, I aggregated yearly data into groups, representing the before and after period of each SPE event (Table 2.5). Area was averaged for each cell in each time period to obtain one value for each cell. Although this lowers sample size, aggregation was necessary to help meet model assumptions.

**2.8.2 Linear mixed models**

Linear mixed models (LMMs) were used due to the hierarchical nature of the data, with values grouped by location. LMMs help account for the lack of independence of each data point and consider what values are true replicates. In my data, smaller cells would be true replicates, but as all assumptions of a LMM were violated without data aggregation, cell was not used as a random effect. Deviations from normality and homoscedasticity were tested using visual assessment with Q-Q plots and histograms (Arnau *et al.*, 2013). No apparent violations were observed when data was averaged across cells. I used the R package *lme4* to run all LMMs (Bates *et al.*, 2015). All LMMs followed the same structure, with time period as the fixed effect and grid as the random effect:

Separate models for each time period combination were run for each land use type and transition.

To determine model fit, pseudo-R2 values were calculated to examine the amount of variation explained by the fixed effects, as well as the whole model (Nakagawa and Schielzeth, 2013). Using R2 values allows for between-model comparison, helping to assess the relative importance of time period as a predictor for each land use type and transition. Variables were considered significant when the error around the slope was low and when the *p*-value was under 0.05 (Zhu, 2012). Pseudo *P*-values were calculated using the *lmerTest* package (Kuznetsova *et al.*, 2017). Significance was further examined through effect size examination, where effect size is defined as the average group difference (Zhu, 2012).

**2.8.3 Additional analysis: breakpoint analysis**

Breakpoint analysis examines the location of tipping points, where properties in a dataset change dramatically (Toms and Villard, 2015). Breakpoints are determined by a change in the relationship between the response and explanatory variable, visible through a sharp change in the slope. It therefore can be equivalent to a piecewise linear regression. Breakpoints were determined using the *segmented* R package (Muggeo, 2017). Such analysis does not allow for a LMM design, so linear models were used. Data was thus summed by year to obtain a total yearly sum of area for each land use type and transition. A separate model was run for each land use type and transition using the following structure:

Breakpoints were examined visually to determine if they align with the SPE events.

**3. Results**

**3.1 LUC visibility within three years following SPE events**

**3.1.1 SUC**

Within three years following SUC, abandoned land cover significantly decreased (Table 3.1 and Graph), implying a rejection of H1 where abandoned land cover is predicted to visibly increase. Time period explained a very low proportion of the variation within the abandoned land cover data (R2M = 6%). By adding grid as a random effect, a greater proportion of variation observed was explained (R2C = 30%).

I found no significant relationship between time period and both intensive and extensive land (Appendix), implying an acceptance of the null hypothesis (H10).

**3.1.2 EUA**

Within three years following EUA, abandoned land cover significantly decreased (Table 3.2 and graph), causing an acceptance of H2. Intensive land cover significantly increased within three years following EUA (Table 3.2), allowing acceptance of H2. The time period predictor explained little variation in both land use models, with a slight increase when grid was introduced as a random effect (Table 3.2).

There was no significant relationship between time period and extensive LUC (Appendix). This implies an acceptance of H10 for extensive land, indicating that extensive LUC is not visible directly following either SPE event.

**3.1.3 Other potential predictors of LUC: location**

When data is not aggregated, cell can be included as a nested random effect, which provides finer scale data. In all cases, this increased total model fit (R2C) and significance (decreased *p*-value) as compared to the models solely including the larger grid as a random effect (Appendix). However, the variation explained by the fixed, time period, effect (R2M) was only increased for the extensive LUC LMM for the EUA time period (Appendix). Assumptions were not met for models including cell as a random effect, with no transformation helping to mitigate normality and homoscedasticity issues. Results still may indicate the importance of location in determining LUC and LUC visibility (graph).

**3.2 Land use transition visibility within three years following SPE events**

**3.2.1 SUC**

Within three years following SUC, the amount of intensive land transitioning to abandoned land significantly decreased (Table 3.3 and graph). A negative relationship for this transition causes a rejection of H2. Again, the fixed effect explains a very small proportion of variation in the data, with model fit considerably improved when considering grid as a random effect (Table 3.3).

**Table 3.3** – Significant LMM ran between time period and transition to abandoned land area.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Land use** | **Estimate** | | **± SE of estimate** | **t-value** | | ***p*-value** | | **R2M** | **R2C** |
| **I 🡪 A** | -3.02 | 0.70 | | | -3.76 | | <0.001 | 0.04 | 0.24 |

I found no significant relationship between SUC and the transition from intensive to extensive land (Appendix), causing null hypothesis acceptance (H20).

**3.2.2 EUA**

For both transitions to intensive land from extensive and abandoned land, no significant relationship was found (Appendix). By accepting the null hypothesis, it is implied that neither transition is visible directly following EUA.

**3.3 Time lag on LUC visibility following SPE events**

**3.3.1 SUC**

When examining the influence of a time lag following SUC on LUC and transitions, all LUC and transitions were significant (Table 3.4, Appendix).

For abandoned land, land cover significantly decreased, with a steeper slope than that seen directly following SUC (Table 3.1 & 3.4). For this model, a greater proportion of the data’s variation is explained by the fixed effect (R2M = 14%). By moving the time window, extensive land use has a weakly negative, significant relationship with time period, with the fixed effect being the sole explainer of variation in the model (Appendix). This relationship is weak, with a slight decrease in extensive land cover following SUC (Table 3.4). Intensive land use change also becomes significant, with a strong negative relationship indicating a loss of intensive land (Table 3.4). A slightly weaker relationship is seen for the transition from intensive to abandoned land, as compared to the change observed directly following SUC (Table 3.3 & 3.4). The transition to extensive from intensive land becomes significant with this new time window as a predictor (Table 3.4). A very weak relationship is seen, with slight less transitioning to extensive land following SUC.

**3.3.2 EUA**

When testing the influence of a time lag following EUA on LUC and transitions, time lag was a significant predictor for intensive LUC only. A steep positive relationship indicates a strong move towards intensive land use following EUA (Table 3.5). The slope (estimate) of this relationship is weaker than that seen directly following EUA (Table 3.2).

**3.3.3 Breakpoints**