MCCA Land Use Prediction Model for the Peruvian Amazon

CAS Advanced Machine Learning 2023/2024
University of Bern, Switzerland

Authors:
Tobias Liechti, tobiliechti@gmail.com
Jan Göpel, jan.goepel@wyssacademy.org

Abstract:

The Peruvian Amazon, a vital part of the world's largest tropical rainforest, faces significant land use and land cover changes (LULCC) due to deforestation, agricultural expansion, and infrastructure development. These changes have severe impacts on biodiversity, climate regulation, and local communities. Accurate prediction of these changes is crucial for effective conservation and sustainable development. Traditional methods like remote sensing and historical trend analysis have limitations, necessitating advanced predictive models. This study explores the integration of Markov Chain models and Cellular Automata (MCCA) to simulate LULCC in the Peruvian Amazon. Markov Chains predict land use transitions based on observed probabilities, while Cellular Automata consider spatial interactions, enhancing the predictive capability of the model. The MCCA model uses diverse data sources, including satellite imagery and socio-economic data, for comprehensive simulations. The model is validated with historical data, showing high accuracy and reliability. The results indicate trends and potential future changes in land use, providing valuable insights for policymakers and conservationists. Despite computational constraints, the study demonstrates the MCCA model's potential in predicting LULCC, highlighting its significance for sustainable management and conservation strategies in the Peruvian Amazon.

1. Introduction

The Peruvian Amazon, a critical component of the world's largest tropical rainforest, is experiencing significant land use and land cover changes (LULCC) driven by various anthropogenic activities. These changes include deforestation, agricultural expansion, and infrastructure development, all of which have profound implications for biodiversity, climate regulation, and local communities. Understanding and predicting land use changes in this region is essential for formulating effective conservation strategies, sustainable development policies, and climate mitigation efforts.

Land use change in the Amazon has been a focal point of numerous environmental studies due to its global significance. The Amazon basin plays a crucial role in the global carbon cycle, acting as a major carbon sink. Deforestation and land conversion to agriculture not only release stored carbon dioxide into the atmosphere, exacerbating climate change, but also disrupt local hydrological cycles and biodiversity (Fearnside, 2005; Malhi et al., 2008). In the Peruvian Amazon, these issues are particularly pronounced due to the region's high rates of forest loss and the presence of diverse ecosystems that are highly sensitive to environmental disturbances (Oliveira et al., 2007).

Predicting future land use scenarios is vital for policymakers and conservationists. Traditional methods of land use analysis, such as remote sensing and historical trend analysis, provide valuable insights but often fall short in projecting future changes under different socio-economic and environmental conditions (Lambin et al., 2003). This gap necessitates the development and application of robust predictive models that can simulate land use dynamics with greater accuracy and reliability.

One promising approach to simulating land use change is the integration of Markov Chain models with Cellular Automata (CA). Markov Chain models are particularly useful for predicting land use transitions based on observed probabilities. They capture the likelihood of change from one land use category to another over time, providing a stochastic framework for modeling land use dynamics (Pontius et al., 2001). However, Markov Chains alone do not consider the spatial configuration of land use changes, which is where Cellular Automata complement this method.

Cellular Automata are spatially explicit models that simulate the evolution of complex systems based on local interactions. In the context of land use change, CA models consider the influence of neighboring land parcels on the state of a given parcel. This spatial component is critical in capturing the patterns and processes of land use changes, such as the spread of agricultural areas or urban sprawl (Clarke et al., 1997). By integrating Markov Chain models with CA, researchers can create hybrid models that leverage the strengths of both approaches: the probabilistic transition capabilities of Markov Chains and the spatial interaction mechanisms of Cellular Automata.

The application of Markov Chain-Cellular Automata (MCCA) models in the Peruvian Amazon offers several advantages. Firstly, it allows for the incorporation of diverse data sources, including satellite imagery, land use maps, and socio-economic data, to generate more comprehensive and accurate simulations (Soares-Filho et al., 2002). Secondly, MCCA models can be calibrated and validated using historical data, enhancing their predictive power (Mas et al., 2014). Lastly, these models facilitate scenario analysis, enabling the exploration of different policy interventions and their potential impacts on land use dynamics.

In conclusion, simulating land use change in the Peruvian Amazon using Markov Chain and Cellular Automata methods is a crucial step towards understanding and managing this ecologically and economically significant region. By providing detailed and accurate predictions of land use trajectories, these models can support the development of strategies aimed at conserving the Amazon's biodiversity, mitigating climate change, and promoting sustainable land management practices.

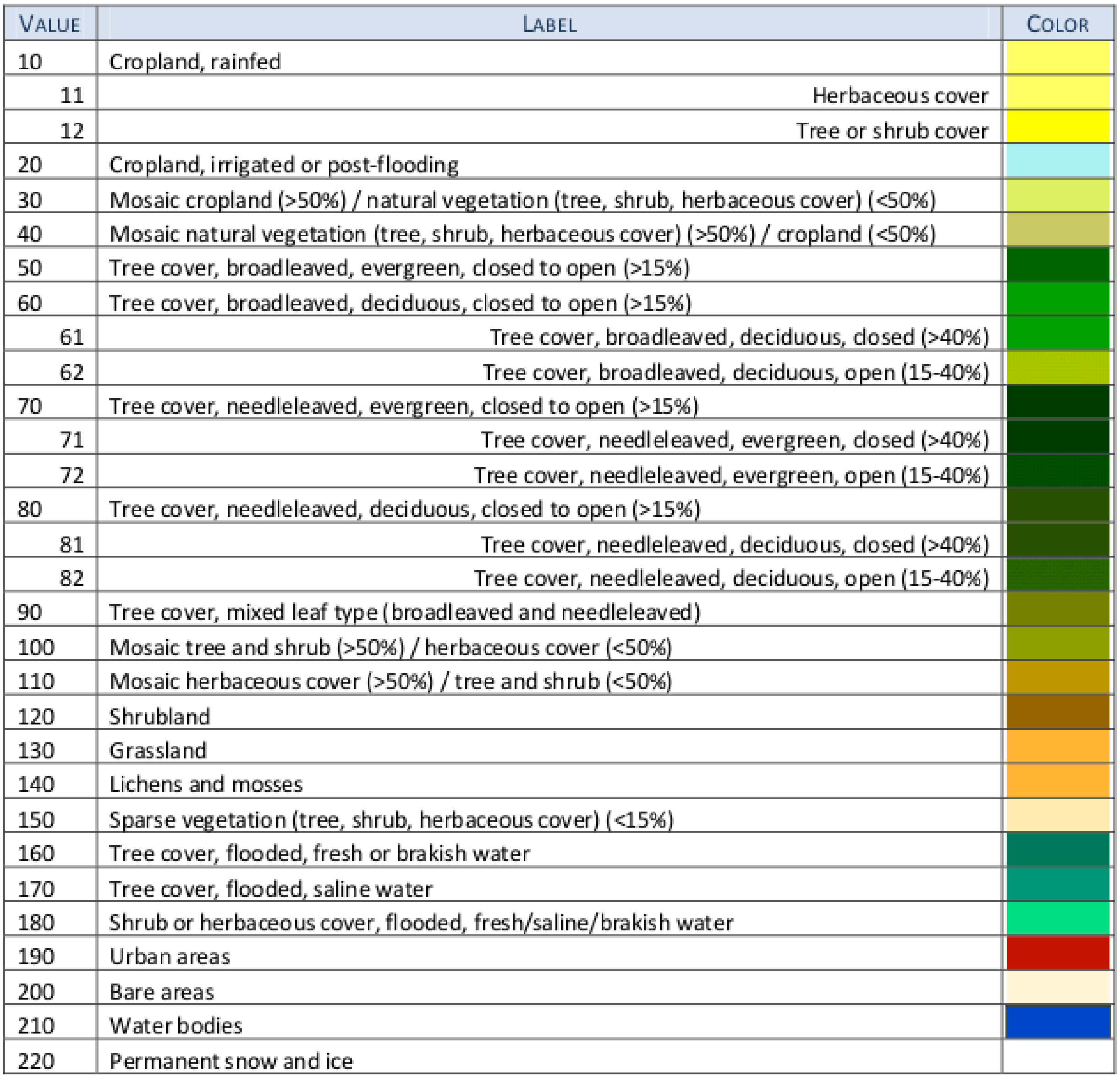
This documentation is to be understood as a description, a protocol of sorts, of our endeavor to create a MCCA land use prediction model to be applied to numerous case studies, regions and constellations of determining factors or drivers of deforestation and LULCC. Our aim is not to simulate future LULCC but to present a method to do just that in the context of specific situations. The first section describes our motivation to do so. The second section describes the data we used as input throughout our exercise while the third section explains the steps necessary to make the required data input align with the modeling technique we chose. In the fourth section we meticulously describe our approach to the employed method and we also in depth discuss the validation results for the study region and time span we chose to analyze. The fifth section illustrates our results while we discuss shortcomings, weaknesses and a way forward in our sixth and final section.

2. Data

a. Land use data

Land cover was obtained from the ESA Climate Change Initiative (CCI). A consistent global annual land cover dataset at 300m spatial resolution from 1992 to 2015 was downloaded via FTP Server Software. Each pixel value corresponds to the label of a land cover class defined based on the UN Land Cover Classification System (LCCS). As can be seen in Table 1, the typology counts 22 global land cover categories.

Table 1: LCCS land cover categories, ESA CCI



Value	Label	Color
10	Cropland, rainfed	🟨
11	Herbaceous cover	🟨
12	Tree or shrub cover	🟨
20	Cropland, irrigated or post-flooding	🟦
30	Mosaic cropland (>50%) / natural vegetation (tree, shrub, herbaceous cover) (<50%)	🟨
40	Mosaic natural vegetation (tree, shrub, herbaceous cover) (>50%) / cropland (<50%)	🟨
50	Tree cover, broadleaved, evergreen, closed to open (>15%)	🟩
60	Tree cover, broadleaved, deciduous, closed to open (>15%)	🟩
61	Tree cover, broadleaved, deciduous, closed (>40%)	🟩
62	Tree cover, broadleaved, deciduous, open (15-40%)	🟩
70	Tree cover, needleleaved, evergreen, closed to open (>15%)	🟩
71	Tree cover, needleleaved, evergreen, closed (>40%)	🟩
72	Tree cover, needleleaved, evergreen, open (15-40%)	🟩
80	Tree cover, needleleaved, deciduous, closed to open (>15%)	🟩
81	Tree cover, needleleaved, deciduous, closed (>40%)	🟩
82	Tree cover, needleleaved, deciduous, open (15-40%)	🟩
90	Tree cover, mixed leaf type (broadleaved and needleleaved)	🟩
100	Mosaic tree and shrub (>50%) / herbaceous cover (<50%)	🟩
110	Mosaic herbaceous cover (>50%) / tree and shrub (<50%)	🟧
120	Shrubland	🟫
130	Grassland	🟧
140	Lichens and mosses	🟧
150	Sparse vegetation (tree, shrub, herbaceous cover) (<15%)	🟨
160	Tree cover, flooded, fresh or brakish water	🟩
170	Tree cover, flooded, saline water	🟩
180	Shrub or herbaceous cover, flooded, fresh/saline/brakish water	🟩
190	Urban areas	🟥
200	Bare areas	⬜
210	Water bodies	🟦
220	Permanent snow and ice	⬜

b. Study Area

The authors chose the Peruvian Amazon as their study area. It is one of the most biodiverse regions on Earth, hosting an incredible variety of plant and animal species. This biodiversity is vital for maintaining ecological balance and resilience. Many species found in the Peruvian Amazon are endemic, meaning they exist nowhere else on the planet. This uniqueness contributes to the overall genetic diversity.

The Amazon rainforest, including the Peruvian portion, acts as a significant carbon sink. It absorbs large amounts of carbon dioxide from the atmosphere, helping to mitigate climate change. The extensive forest cover plays a crucial role in regulating local and global climate patterns. It influences precipitation cycles and helps maintain atmospheric moisture levels.

Nonetheless, the Peruvian Amazon faces threats from deforestation due to agriculture, logging, mining, and infrastructure development. This leads to habitat loss, reduction in biodiversity, and increased carbon emissions.

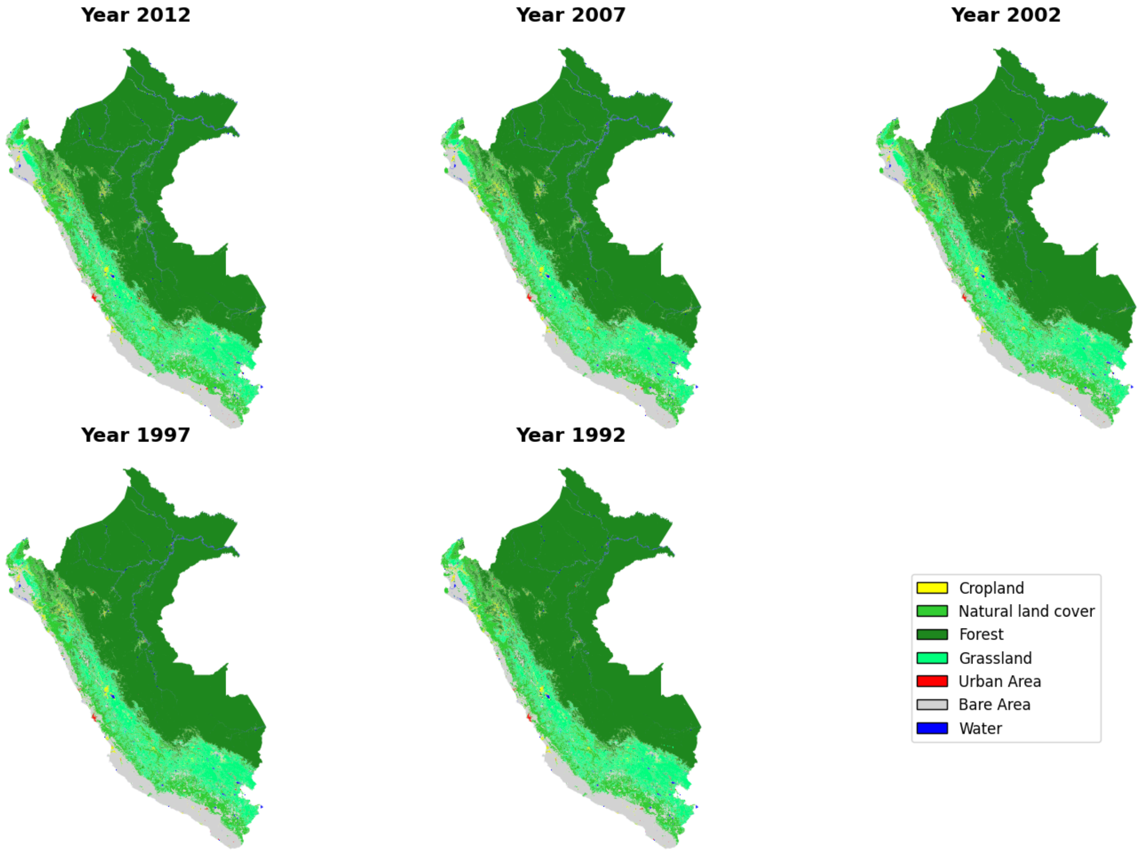


Figure 1: The Peruvian rainforest in 1992, 1997, 2002, 2007, and 2012. Over the years, the maps illustrate a trend of deforestation and an increase in cropland areas.

3. Data Preprocessing

a. Land use data

The available land use data contains a high degree of information about the type of observed land cover as defined by its many land use categories. To reduce computational workload on the one hand and to increase the quantity of class-specific land use changes on the other hand, the authors decided to aggregate the analyzed land use classes according to Table 2.

Table 2: land cover aggregation mapping

In	Out	mapping	land use class
10	1	ValueToValue	cropland
11	2	ValueToValue	natural
30	1	ValueToValue	cropland
40	2	ValueToValue	natural
50	3	ValueToValue	forest
60	3	ValueToValue	forest
61	3	ValueToValue	forest
62	3	ValueToValue	forest
80	3	ValueToValue	forest
100	2	ValueToValue	natural
110	2	ValueToValue	natural
120	2	ValueToValue	natural
130	4	ValueToValue	grassland
150	2	ValueToValue	natural
153	2	ValueToValue	natural
160	3	ValueToValue	forest
170	3	ValueToValue	forest
180	2	ValueToValue	natural
190	5	ValueToValue	urban area
200	6	ValueToValue	barren area
210	7	ValueToValue	water
220	6	ValueToValue	barren area

Especially the early analyzed years (1992 - 2012) were not very dynamic in terms of land use changes targeting to fulfill an economic interest. Only in the latter part of the analyzed time span (2002 - 2012), more pronounced anthropogenic land use changes could be observed. To overcome the only slight changes from year to year, the authors decided to focus on 5-year time slices as their temporal scale of analysis. The annual land use maps, available in a multi-raster-format, were split according to analysis requirements into 4 time slices (1992 - 1996, 1997 - 2001, 2002 - 2006, 2007 - 2011) using a GIS software (ArcGIS Pro 3.2.0).

b. study area

In general, we predicted LULCC for the whole country of Peru. For an analysis of observed and predicted LULCC, we had to single out the relevant ecosystems. For that purpose, we used GIS software (ArcGIS Pro 3.2.0) to overlay the observed and predicted LULCC maps with a map of ecosystem distribution in Peru. This map was obtained from Ministerio del Ambiente de Perù, Direccìon General de ordenamiento territorial (2011). As depicted in Table 3, 4 individual ecoregions have been aggregated to form the Peruvian Amazon, with a total area of 857,852.07 km^2 of Peru's total area of 1.285 million km^2.

Table 3: ecoregions included in analysis of the Peruvian Amazon

ID	shape	domain	name	ecoregion	ha
5	Polygon	Amazonia	Provincia del Páramo	Paramo	159595.00
6	Polygon	Amazonia	Provincia de las Yungas (Selva Alta)	Mountain Forest (Yungas)	20807744.34
8	Polygon	Amazonia	Provincia Amazónica (Selva Alta)	Tropical Rainforest Amazonia	58182777.48
10	Polygon	Amazonia	Provincia Pacífica	Moist Equatorial Forest	5603754.38
12	Polygon	Amazonia	Provincia Amazónica (Selva Alta)	Tropical Rainforest Amazonia	565983.92
13	Polygon	Amazonia	Provincia Amazónica (Selva Alta)	Tropical Rainforest Amazonia	210227.14
14	Polygon	Amazonia	Provincia Amazónica (Selva Alta)	Tropical Rainforest Amazonia	43133.49
15	Polygon	Amazonia	Provincia Amazónica (Selva Alta)	Tropical Rainforest Amazonia	211991.49

c. Data processing steps necessary to maintain spatial integrity

In order to maintain the spatial integrity of data input it was necessary to convert data formats between specific steps of our analysis.

The final product after the prediction process was a GeoTIFF file (raster format), but for the step of validation, a much more precise point shape file was required. Furthermore, for an efficient visualization of results, we required a polygon shape file (the raster files as well as point shape files were too large to be processed on Google Colab).

All necessary conversion steps (raster to point for validation and raster to polygon for visualization) were realized using a GIS software (ArcGIS Pro 3.2.0).

4. Methods

a. Markov Chain transition probability prediction

To analyze the observed land use changes and apply these observed transitions to predict possible future states of a system, the authors decided to employ a Markov Chain Process. A Markov chain is a mathematical system that undergoes transitions from one state to another within a finite or countable number of possible states. It is a random process that satisfies the Markov property, which essentially means that the future state depends only on the current state and not on the sequence of events that preceded it. Furthermore, drivers of change, e.g. socioeconomic and biophysical factors that might impact quantity and location of change, are not taken into account.

Both of the following steps (calculation of transition probabilities and transformation/prediction based on transition probabilities) had to be realized on a local PC, since the 12GB of RAM available on a Google Colab instance were not sufficient to run the scripts.

Required system configuration:
12th Gen Intel(R) Core(TM) i7-12700H   2.30 GHz
64.0 GB RAM

i. LULCC probabilities

As a first step, it was necessary to calculate a change matrix. This change matrix resembles an accounting of which land use classes (number of pixel cells) changed and into which new land use classes (number of pixel cells) as depicted in Table 4.

Table 4: change matrix for LULCC observed between 1992 and 1996

```
change matrix:
          0        1         2         3         4       5         6       7
0  16340475        0         0         0         0       0         0       0
1         0   264134        24       443        61     158         0       6
2         0     1494   2099063       382      4621     652      2877      45
3         0     9992     14524   8616351      1477     129        26   20966
4         0      229       742       368   1581739     154       165      63
5         0        0         0         0         0   18206         0       0
6         0      290      3267         0      1337     279   1051882       0
7         0        2        17       202         2      31         0   64735
```

As a second step, it was necessary to calculate based on the change matrix, the transition probabilities of a certain land use class transitioning into another specific land use class, resulting in a transition matrix, as depicted in Table 5.

Table 5: transition probability matrix of LULCC observed between 1992 and 1996

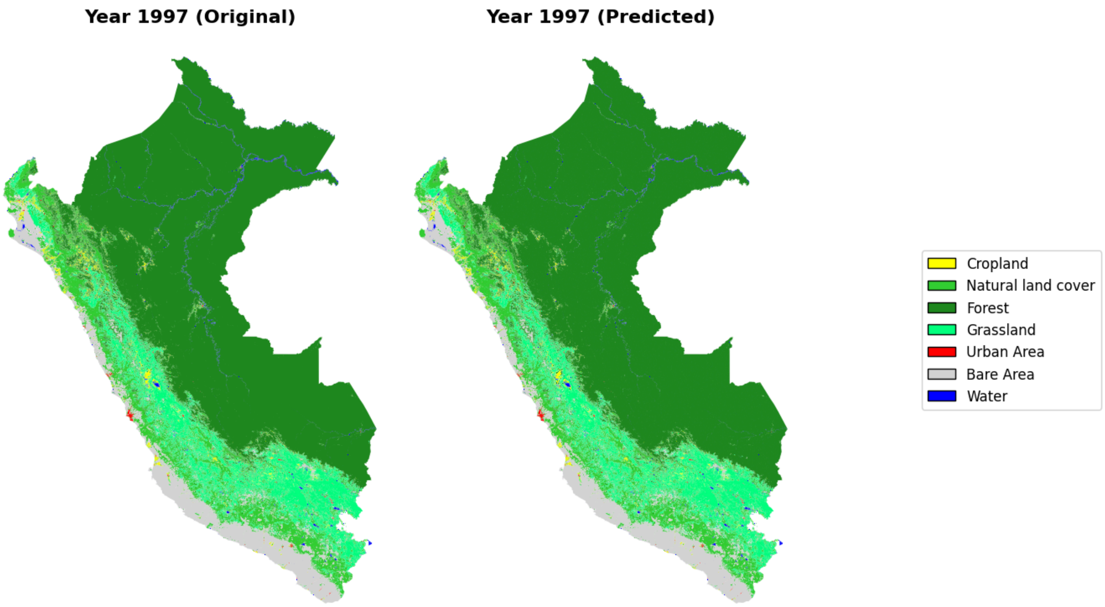
	1	2	3	4	5	6	7
1	0.997387	0.000091	0.001673	0.000230	0.000597	0.000000	0.000023
2	0.000708	0.995225	0.000181	0.002191	0.000309	0.001364	0.000021
3	0.001153	0.001676	0.994562	0.000170	0.000015	0.000003	0.002420
4	0.000145	0.000469	0.000232	0.998913	0.000097	0.000104	0.000040
5	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000
6	0.000274	0.003091	0.000000	0.001265	0.000264	0.995106	0.000000
7	0.000031	0.000262	0.003108	0.000031	0.000477	0.000000	0.996092

The authors repeated this step for all necessary time slices:
1992 - 1996; resulting in a transition matrix to predict land use class distribution for 1997
1997 - 2001; resulting in a transition matrix to predict land use class distribution for 2002
2002 - 2006; resulting in a transition matrix to predict land use class distribution for 2007
2007 - 2011; resulting in a transition matrix to predict land use class distribution for 2012

ii. Transformation of current land use map/prediction of land use class distribution in space

Following the aforementioned steps, the authors applied the calculated transition matrices on the original land use maps for the year that the authors aimed to predict.
As an example:
The transition matrix of the time slice 1992 - 1996 was applied to transform the original land use map of the year 1997 to resemble the simulated or predicted land cover and land use state of the year 1997.
This process resulted in two distinct land cover and land use maps for the respective simulated year as can be seen in Figure 2. After converting them from a raster format to a point shape format, these maps could further be employed for validation purposes.

Figure 2: land use map 1997 original and land use map 1997 predicted


Year 1997 (Original)

Year 1997 (Predicted)

b. Validation

To assess the accuracy of the model's predictions, we performed a validation inspired by Alamanos (2023) and optimized the code to adhere to object-oriented programming principles. We used the following tests:

i. Accuracy classification Score (Accuracy in Table):

It takes the ground truth amount of pixels (correctly classified) and compares it to the overall amount of pixels. The output is normalized and therefore the output a fraction of the correctly classified pixels:

$$Accuracy = \frac{Correctly\ classified\ pixels}{All\ pixels} \tag{1}$$

ii. Mean Absolute Error (MAE):

The MAE measures the error between predicted pixel and actual pixel (land use in our case). It is the sum of the absolute differences divided by the number of pixels (*n*). The current pixel is *p*:

$$MAE = \frac{\sum_{p=1}^{n} |actual_p - predicted_p|}{n} \quad (2)$$

iii. Root Mean Square Error (RMSE):

The RMSE is the quadratic mean of the difference between actual and predicted pixels (land use in our case) and takes the square root of the result. The smaller the RMSE is, the better the prediction of land use.

$$RMSE = \sqrt{\frac{\sum_{p=1}^{n} (actual_p - predicted_p)^2}{n}} \quad (3)$$

iv. Kappa (K):

Cohen’s Kappa measures the inter-rater reliability. In the context of land use categories, it gives us a statistic about the reliability of predicted land use categories versus true land use categories. It also takes into consideration that the agreement is based purely on chance.

$$\kappa = \frac{p_O - p_e}{1 - p_e} \quad (4)$$

Where:

- **PO** is the relative observed agreement between the land use categories in the predicted versus actual scenario (see Almanos, 2023).
- **PE** is the hypothetical agreement between the land use categories, if their agreement is by chance only.

Interpretation:

The Kappa Coefficient is a floating point between -1 and 1. 1 means complete agreement, 0 or lower means that the agreement is based on chance.

v. Confusion Matrix:

Computes the accuracy of the classification based on signal detection theory. According to scikit-learn the confusion matrix C is relative to Cij so that the i is the group of known land use categories and j is the group of predicted land use categories.

- The count of true positives is C11,
- The count of true negatives is C00,
- The count of false positives is C01, and

- The count of false negatives is C10

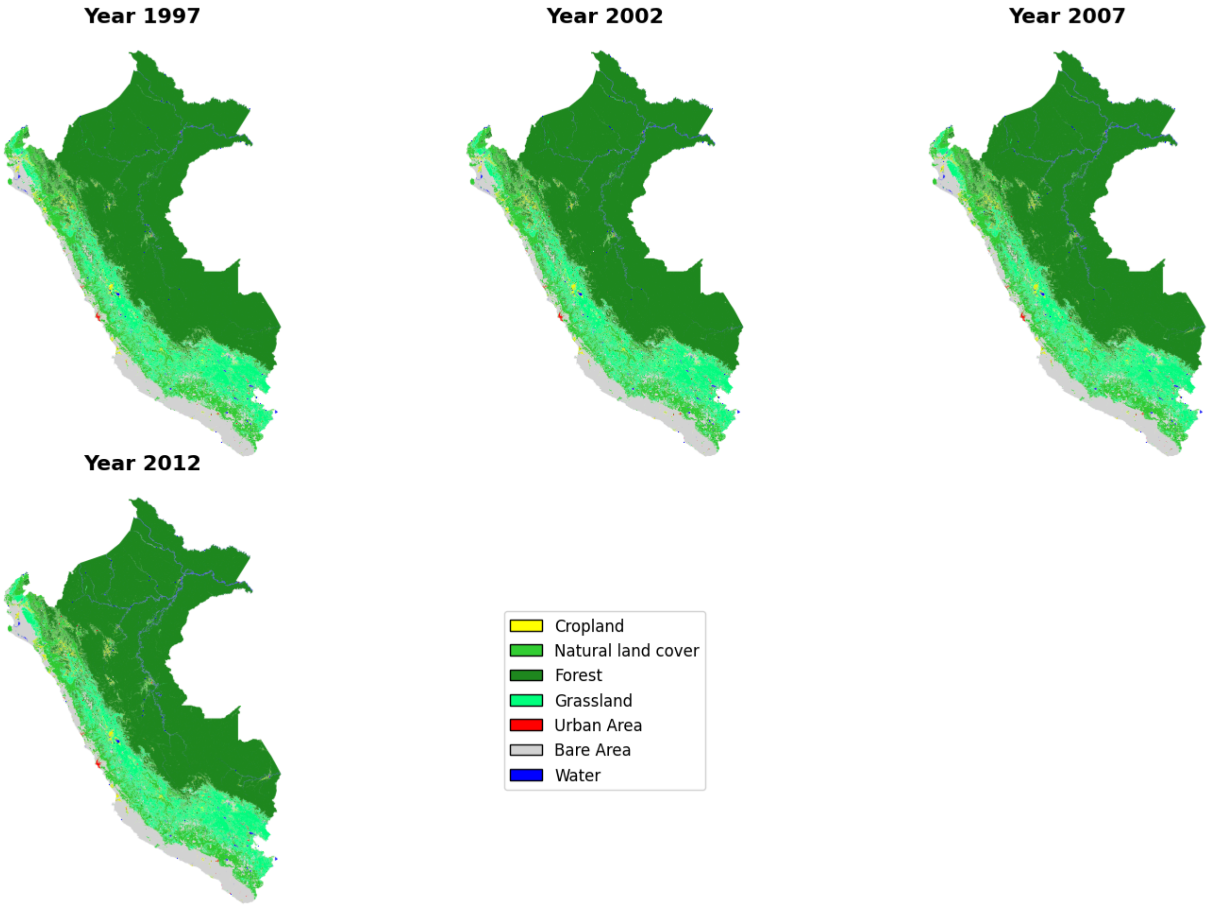
vi. Classification Report

Based on the true and predicted land use classes, the classification Report provides the following:

- **Precision**: Ratio of accurately predicted land use classes (C11) to all predicted positive land use classes (C11 + C01). High precision indicates a low rate of false positive predicted classes.
- **Recall (Sensitivity)**: Ratio of accurately predicted land use classes (C11) to all observed land use classes (C11 + C10). High recall indicates a low rate of false negatives.
- **F1 Score**: The weighted average of precision and recall. It considers both false positives (C01) and false negatives (C10), and is particularly useful for imbalanced datasets where one land use class is dominant (e.g., the Peruvian rainforest).
- **Support**: The number of actual occurrences in the dataset for each known land use category. This contextualizes the other metrics, showing how many instances there are for each class.

5. Results

a. Validation



Year 1997

Year 2002

Year 2007

Year 2012

Figure 3: Predicted LULCC for the Peruvian Amazon

1997

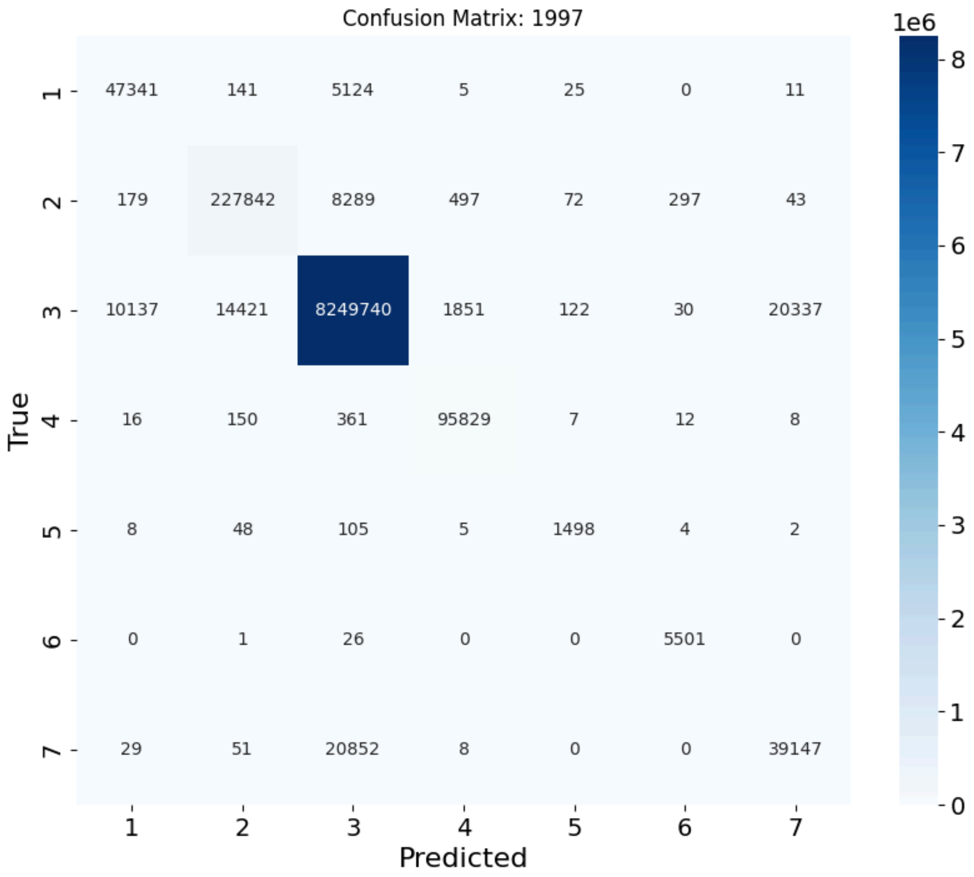


Figure 4: Confusion matrix 1997 with LULCC codes from Table 2

Highest misclassification in class 3 (figure 4). The reason was its large support size (table 6).

Accuracy	0.990483
MAE	0.025712
RMSE	0.294689
KAPPA	0.90608

class	precision	recall	f1-score	support
1	0.820326	0.899216	0.857961	52647.0
2	0.938958	0.960471	0.949593	237219.0
3	0.995805	0.994347	0.995075	8296638.0
4	0.975905	0.994252	0.984993	96383.0
5	0.868910	0.897006	0.882734	1670.0

class	precision	recall	f1-score	support
6	0.941307	0.995116	0.967464	5528.0
7	0.657402	0.651505	0.654441	60087.0

Table 6: Validation statistics Accuracy, MAE, RMSE, KAPPA and Classification Report with precision, recall f1-score and support for the year 1997

As can be seen in Table 6, Class 3 has the highest f1-score. As it is the Forest class, this was to be expected by the support. The second highest f1-score has class 4.

2002

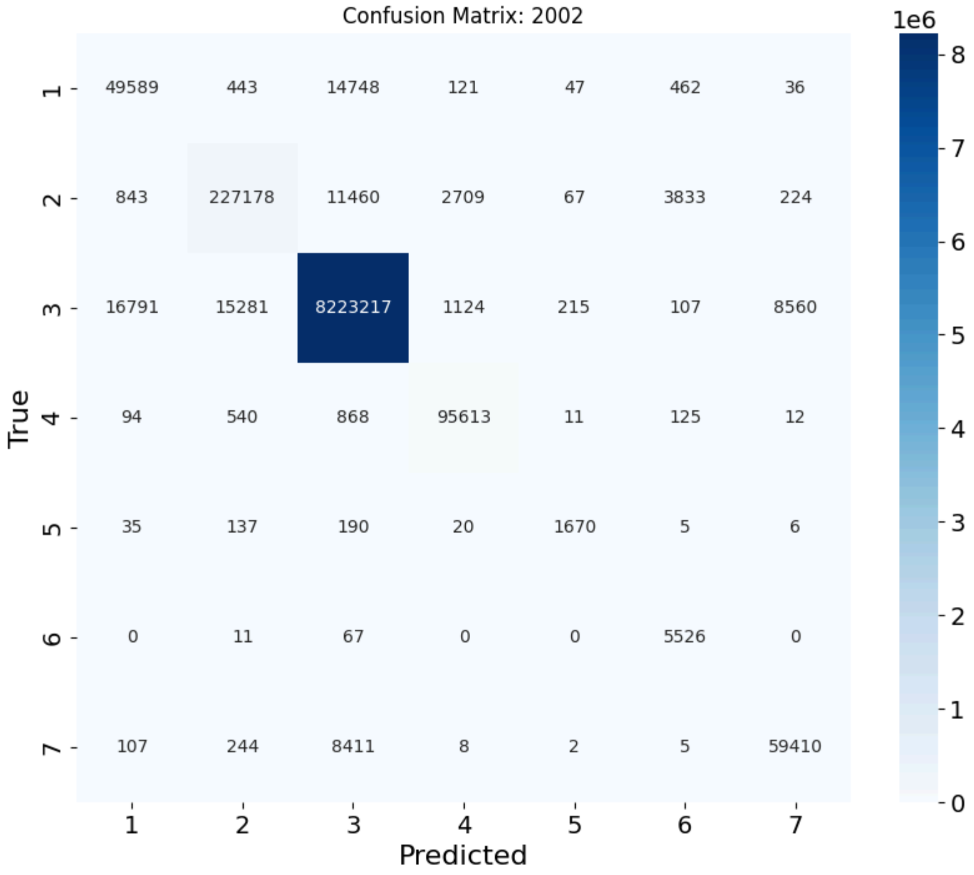


Figure 5: Confusion matrix 2002 with LULCC codes from Table 2

In the confusion matrix for 2002 (figure 5) an improvement in precision and recall can be seen across most classes. However class 6 had a high increase in false negatives (especially class 2).

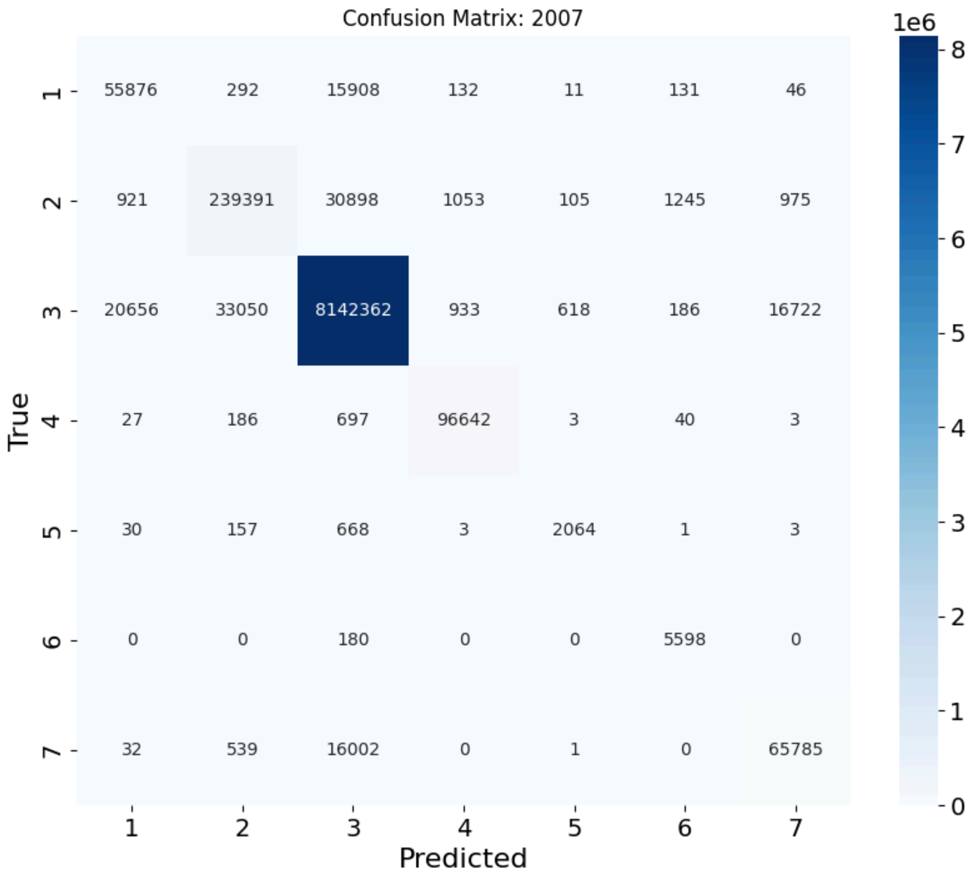
Accuracy	0.989947
MAE	0.021902
RMSE	0.248336
KAPPA	0.906401

class	precision	recall	f1-score	support
1	0.735098	0.757709	0.746232	65446.0
2	0.931691	0.922311	0.926977	246314.0
3	0.995672	0.994909	0.995290	8265295.0
4	0.960018	0.983036	0.971391	97263.0
5	0.830020	0.809501	0.819632	2063.0
6	0.549140	0.986081	0.705432	5604.0
7	0.870502	0.871280	0.870891	68187.0

Table 7: Validation statistics Accuracy, MAE, RMSE, KAPPA and Classification Report with precision, recall f1-score and support for the year 2002

If we compare tables 6 and 7 we can observe that class 6's performance dropped considerably between 1997 and 2002. Classes 1, 2, 4, and 5 dropped moderately in precision and recall. Class 3 did not change much and class 7's precision and recall increased.

2007



	1	2	3	4	5	6	7
1	55876	292	15908	132	11	131	46
2	921	239391	30898	1053	105	1245	975
3	20656	33050	8142362	933	618	186	16722
4	27	186	697	96642	3	40	3
5	30	157	668	3	2064	1	3
6	0	0	180	0	0	5598	0
7	32	539	16002	0	1	0	65785

Figure 6: Confusion matrix 2007 with LULCC codes from Table 2

The most misclassified classes in all the confusion matrices. Notable in 2007 were classes 2, 3 and 7 (figure 6).

Accuracy	0.98372
MAE	0.03339
RMSE	0.305851
KAPPA	0.862325

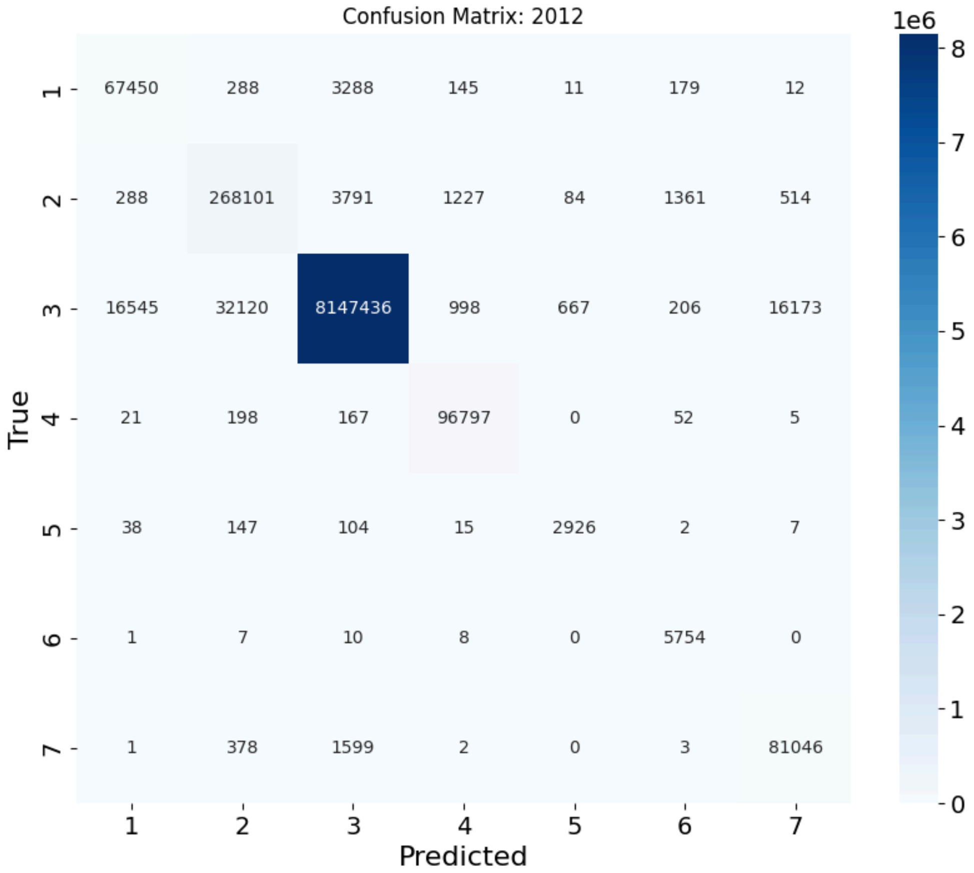
class	precision	recall	f1-score	support
1	0.720590	0.771811	0.745321	72396.0
2	0.874919	0.871819	0.873366	274588.0
3	0.992158	0.991215	0.991686	8214527.0

class	precision	recall	f1-score	support
4	0.978524	0.990205	0.984330	97598.0
5	0.736617	0.705400	0.720670	2926.0
6	0.777392	0.968847	0.862624	5778.0
7	0.787524	0.798759	0.793102	82359.0

Table 8: Validation statistics Accuracy, MAE, RMSE, KAPPA and Classification Report with precision, recall f1-score and support for the year 2007

Comparing tables 7 and 8, class 6 recovered the initial loss of precision, which led to an increase in the f1-score. The classes 2, 5, and 7 declined in precision and recall compared to 2002. Classes 1, 3, and 4 had stable performance metrics.

2012



True \ Predicted	1	2	3	4	5	6	7
1	67450	288	3288	145	11	179	12
2	288	268101	3791	1227	84	1361	514
3	16545	32120	8147436	998	667	206	16173
4	21	198	167	96797	0	52	5
5	38	147	104	15	2926	2	7
6	1	7	10	8	0	5754	0
7	1	378	1599	2	0	3	81046

Figure 7: Confusion matrix 2012 with LULCC codes from Table 2

Compared to 2007, the confusion matrix in 2012 showed a high reduction in false negatives overall (figure 7).

Accuracy	0.990782
MAE	0.018962
RMSE	0.230755
KAPPA	0.925398

class	precision	recall	f1-score	support
1	0.799701	0.945035	0.866315	71373.0
2	0.889994	0.973617	0.929930	275366.0
3	0.998902	0.991879	0.995378	8214145.0
4	0.975855	0.995444	0.985552	97240.0
5	0.793384	0.903365	0.844810	3239.0
6	0.761413	0.995502	0.862863	5780.0
7	0.829056	0.976117	0.896596	83029.0

Table 9: Validation statistics Accuracy, MAE, RMSE, KAPPA and Classification Report with precision, recall f1-score and support for the year 2012

Between the years 2007 and 2012 the precision recall improved notably for classes 1, 2, 5, and 7 (as can be seen by comparing tables 8 and 9). Classes 6, 3 and 4 remained rather stable.

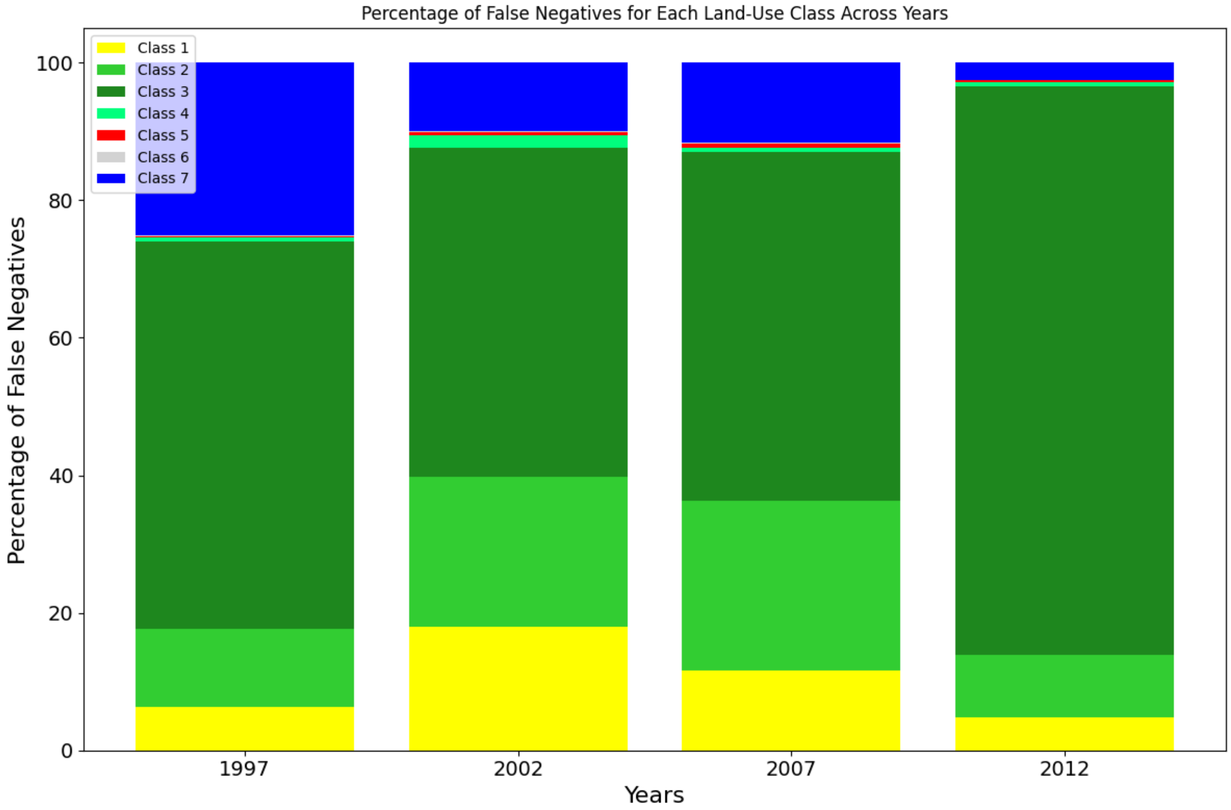


Figure 8 : Proportion of LULCC classified in the wrong class in the Peruvian Amazon

Overall, the accuracy of the classes remained high over the year (except in 2007). The MAE decreased over time, and therefore indicated improved predictive performance. The RMSE followed a similar trend as the MAE and had its highest value in 2007 with the lowest value in 2012. Kohen's Kappa remained almost perfect most of the time, thus indicating a good agreement between actual and predicted LULCC.

The changes between 1997 to 2007 reveal trends in classification performance.
Class 1 declined between 1997 and 2002 but improved from 2007 to 2012. The percentage of false negatives decreased over time.
Class 2 declined between 1997 and 2002 as well but also improved its performance between 2007 and 2012. Most false negatives were allocated in 2002.
Class 3 remained stable and reliable throughout all years. Because of its high support size, the proportion of false negatives is the highest of all classes. The highest proportion of false negatives is in 2012 and in absolute numbers it was in 2007.
Class 4 remained generally stable over the years. The proportion of false negatives is low over the analyzed years.
Class 5 experienced a decline in performance between 1997 and 2007 but improved from 2007 to 2012. The same is true for the rate of False negatives which declined between 2007 and 2012.
Class 6 showed serious fluctuations due to its low sample size in comparison to all the other classes. Its performance declined between 1997 and 2002, increased between 2002 and 2007, and showed a stable performance between 2007 and 2012. The proportion of false negatives is negligible.
Class 7 increased in performance between1997 and 2002, and declined in the following timespan between 2002 and 2007. Its performance recovered between 2007 and 2012. The reduction in false Negatives between 2007 and 2012 signifies a better classification accuracy.

b. Land use evolution over time

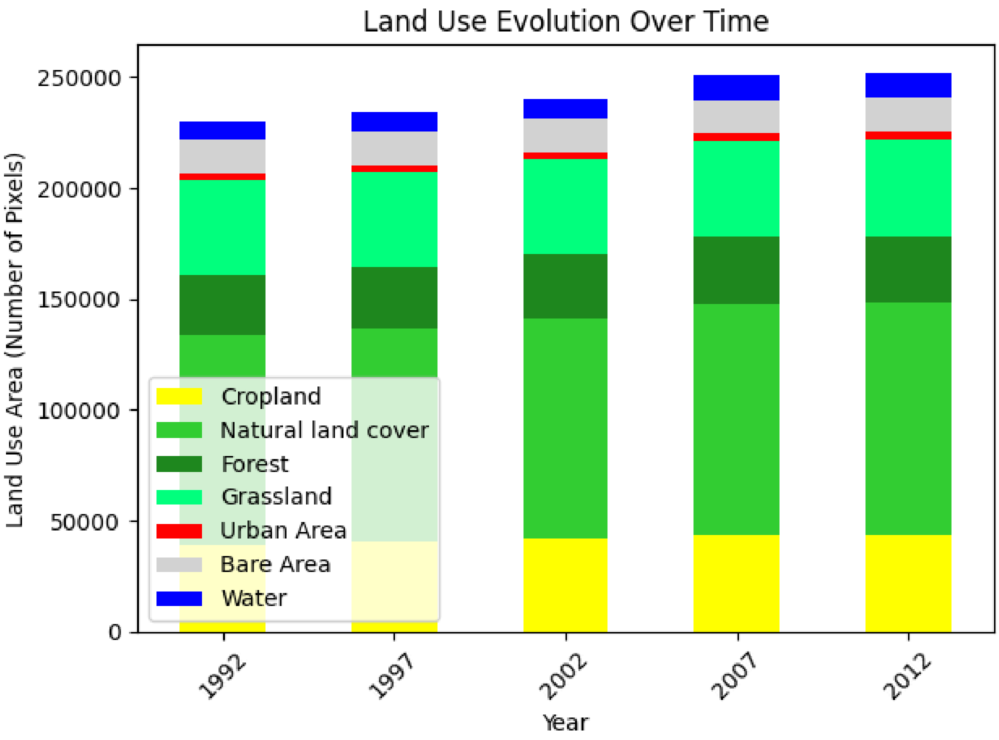


Figure 9 : aggregated Land Use Evolution for the years 1992 until 2012

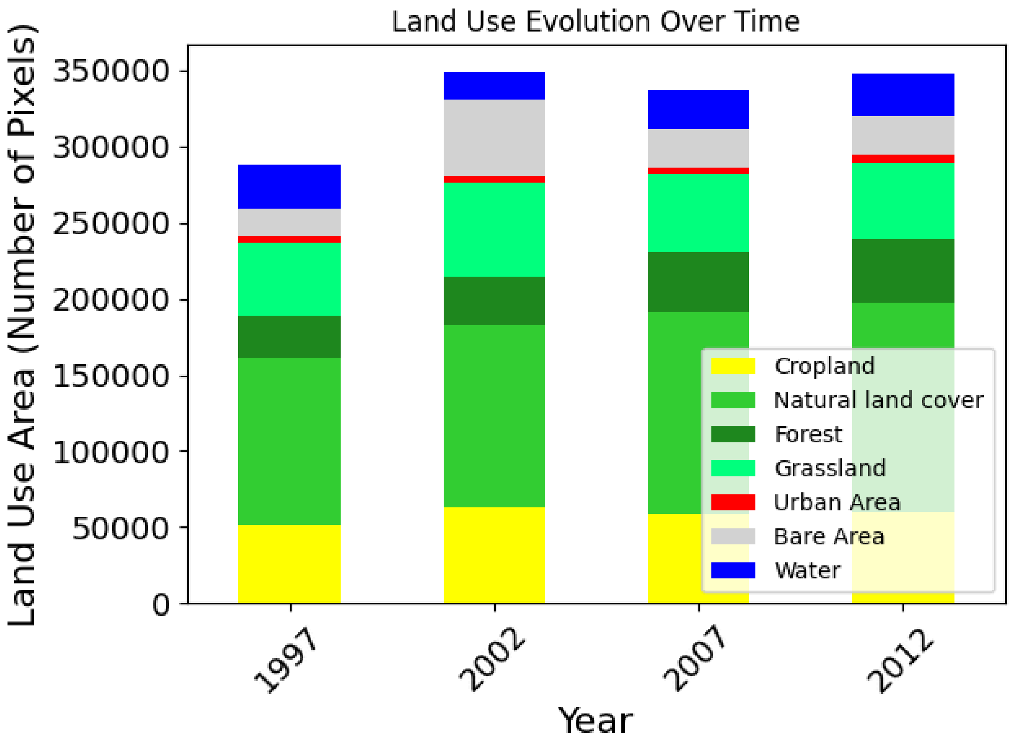


Figure 10 : Predicted Land Use Evolution for the years 1997 until 2012

Aggregated and expected changes in Land Use and Land Cover (LULC) are depicted in Figures 9 and 10, respectively. It is important to note that the GIS function aggregates or simplifies points or pixels into polygons, which can result in significant changes in polygon counts. For instance, if a single pixel of cropland appears in an isolated location, it is converted into a new polygon, thereby increasing the polygon count. If this occurs 1,000 times with individual pixels, it results in 1,000 additional polygons. Our MCCA model tends to generate isolated pixels, which can lead to an inflated polygon count. Therefore, comparing original polygons to predicted polygons may not yield accurate or meaningful insights and could introduce additional uncertainties in the analysis.The data exhibits less fluctuation in the original maps compared to the predicted ones. Nevertheless, the respective functions in the GitHub repository can be used to create an accurate plot if the shapefiles contain pixels and not polygons.

From 2002 to 2007, there is a notable increase in the land use type 'Water,' which is corroborated by the rise in the number of water polygons, indicating a reliable predictor. Conversely, the amount of barren area shows minimal increase between 1997 and 2002, whereas the predicted data indicates a significant rise. The number of polygons representing urban areas remains relatively stable during both the observed and projected periods, with a steady increase aligning with the model's predictions. Additionally, the polygons representing natural land cover, forest, and grassland show an upward trend, reflected in the total polygon count. The expected increase for these categories continues from 1997 to 2002, after which the predicted grassland decreases. In the original map data, the number of polygons representing cropland steadily increases, but the predicted data shows a different pattern: an initial increase from 1997 to 2002, followed by a decrease between 2002 and 2007, and another increase from 2007 to 2012.

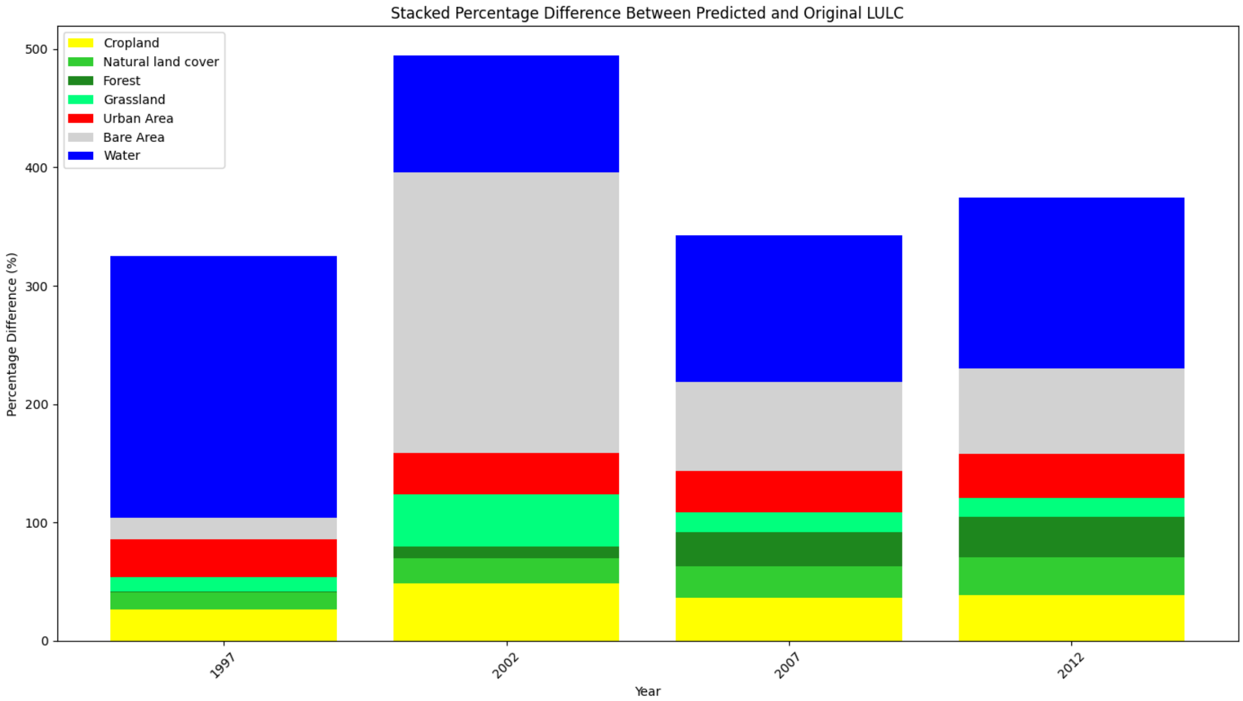


Figure 11 : Difference between predicted and actual land use (in %) for the years 1997 until 2012

Figure 11 illustrates the discrepancy between the predicted and actual LULCC. The predicted land use significantly overestimates the actual land use, with variations ranging from 0.5% (Forest in 1997) to 237% (Bare Area in 2002). This overestimation could be attributed to the model's inability to accurately account for the predominant rainforest in the region. The dense and extensive nature of the rainforest may lead to complexities that the MCCA model does not fully capture, resulting in inflated predictions for certain land use types (such as Water or Bare Area).

6. Discussion

We have created a LULCC simulator based on the described MCCA methodology and tested it by simulating 20 years of LULCC in Peru. From this simulation, we have extracted the LULCC that occurred within the biome of the Peruvian Amazon and validated four predicted maps of land use and land cover employing several metrics that depict a more or less complete picture for specific cases of spatially explicit modeling of LULCC.

In recent years, the integration of Markov Chain models and Cellular Automata has proven to be a robust approach for modeling LULCC (Arsanjani et al., 2013; Aburas et al., 2016). Markov Chains provide a statistical framework to estimate transition probabilities between different land use states, while Cellular Automata add a spatial dimension to this process, capturing local interactions and spatial dependencies (Chen et al., 2014; Houet et al., 2016). This combined approach has been successfully applied in various geographical contexts, demonstrating its utility in projecting future land use scenarios (Mas et al., 2014; Huang et al., 2019).

a. Limitations of this study

First and foremost, our LULCC simulator does not take into account socioeconomic drivers per se. Socioeconomic developments are reflected by the LULCC our model used to learn from, from time step to time step. Observed LULCC is driven by supply and demand for arable land, infrastructure (e.g., irrigation infrastructure), commodities in the domestic sense, and even more so by internationally traded goods. Further, all this evolves driven by international trade and the price-building mechanisms of a global exchange of goods. Thus, our model learns LULCC and incorporates all these aspects indirectly. This limitation aligns with findings from other studies which emphasize the importance of incorporating socioeconomic variables to improve model accuracy and explanatory power (Verburg et al., 2004; Pontius & Malanson, 2005).

This cannot be understood as a shortcoming or limitation in its purest sense, as it is inherent to the method of choice, but it could lead to situations where socioeconomy can only be understood as a black box, not enabling us to assess direct causalities between socioeconomic drivers and LULCC. Moreover, we cannot, thus far, use our LULCC simulator to simulate specific socioeconomic scenarios, as this requires detailed socioeconomic input and a learning or modeling process based thereon. Recent advancements suggest integrating more sophisticated learning processes, such as machine learning techniques, which can incorporate multiple layers of socioeconomic and biophysical data, potentially transforming the simple MCCA process into a more comprehensive Hidden Markov Chain/CA approach (Li et al., 2017; Feng et al., 2020).

Secondly, we were relatively constrained by computational requirements. Our simulation, based on .tif files, was resource-intensive. Even more resource-intensive was the validation processes (based on point shape files). This file type allows for integrating highly complex information while maintaining the spatial integrity of our data. But that all occurs at the cost of increased resource intensity. Our model runs can only be validated on high-performance machines or on an HPC altogether. The code could likely be improved by specific methods that help to minimize resource requirements. This challenge is consistent with other studies highlighting the computational intensity of high-resolution LULCC models and the need for efficient processing techniques (Guan et al., 2011; Vliet et al., 2016).

Also, the handling of the aforementioned multitude of file types made it necessary to outsource several steps of data preprocessing or processing in between analysis steps to other software solutions. For the tasks of converting output to specific formats, we mostly had to employ GIS software (ArcGIS Pro 3.2.0). There are libraries and methods to integrate these processes into our model, but all of them would need implementation and testing, as not all of them would have fit our requirements, especially the requirement to handle large, spatially explicit datasets. Considering these factors, we could have dedicated more time to model development if we hadn't faced such strong resource requirements and numerous steps of format conversions (even with specialist software, these proved to be quite time-intensive).

Third, the simplification of land use and land cover classes led to some artifacts in our model. For instance, the classes of natural vegetation identified as flooded classes (which we integrated into the general land cover classes related to the under-classes, e.g., flooded forest to forest) appeared to impact land use changes in our water-only class, as they are depicted as water or terrestrial land cover in the land use data product employed by the authors. Also, the decision to analyze time slices of 5-year LULCC might have contributed to this situation. A robust study might have to find a more elegant way to aggregate land use and land cover classes, which might help overcome this issue. Additionally, a more robust method to harmonize or organize the analyzed LULCC in its temporal dimension might have helped solve this problem. Similar issues have been documented in the literature, where the aggregation of land cover classes and temporal resolution can introduce significant uncertainties (Chen & Pontius, 2010; Silva et al., 2017).

Fourth and lastly, the original intent of our project has not been fully satisfied. A land use simulator only helps in meaningful ways if it allows for an assessment of cause-effect relationships (which our simulator doesn't, as discussed in preceding paragraphs) or if it allows for an analysis of possible future LULCC. The next step in our endeavor would have been to transfer the calculated transitions of the years 2007 to 2011 to our simulated or predicted 2012 land use map to simulate LULCC from 2012 to 2016, resulting in a land use map for 2016.

These steps would then be repeated until the prediction reached a predefined point in the future. We present a framework for doing so, with the apparent flexibility to adapt it, by creating and processing relevant and fitting input, to specific case study regions or contexts of choice. This approach is consistent with methodologies proposed by Soares-Filho et al. (2002) and Mas et al. (2014), who emphasize iterative and adaptive modeling frameworks for LULCC.

b. A proposal for future studies

As can be anticipated based on the preceding section, the authors have identified two ways the model can be improved.

First, the model could be considerably enhanced in its ability to reflect socioeconomic drivers of change by integrating a hidden layer network in the process of calculating transition probabilities. Several socioeconomic (e.g., population development, GDP development, poverty or development indicators such as the Human Development Index, poverty index) and biophysical drivers (elevation, slope, precipitation, temperature, to name just a few possibilities) could be considered in the learning process and finally, in the determination of the spatially explicit components of LULCC. Further and maybe most importantly, it would enable the simulation of a priori defined socioeconomic scenarios that could help shed light on possible future constellations of LULCC and their complex network of drivers, thus allowing for an analysis of the causality between LULCC and its drivers. This approach has been successfully applied in recent studies incorporating machine learning techniques to improve the predictive accuracy of LULCC models (Feng et al., 2020; Mansour et al., 2021).

Secondly, we propose to further develop our modeling exercise to depict future land use changes following the fourth point of the preceding discussion chapter. Only this would qualify our model, in its current state, to contribute in a meaningful way to current research goals and discussions. Future developments could draw on methodologies for scenario-based modeling as demonstrated by Van Vliet et al. (2016) and Verburg et al. (2019), enabling the exploration of multiple future land use pathways under different policy and socio-economic contexts.

References

Aburas, M. M., Ho, Y. M., Ramli, M. F., & Ash'aari, Z. H. (2016). Improving the accuracy of land use/cover change prediction using a combination of Markov chain and cellular automata models. Environmental Monitoring and Assessment, 188(5), 1-17.

Alamanos, A. (2023). A Cellular Automata Markov (CAM) model for future land use change prediction using GIS and Python. DOI: 10.13140/RG.2.2.20309.19688. Available at: https://github.com/Alamanos11/Land_uses_prediction

Arsanjani, J. J., Helbich, M., Kainz, W., & Boloorani, A. D. (2013). Integration of logistic regression, Markov chain and cellular automata models to simulate urban expansion. International Journal of Applied Earth Observation and Geoinformation, 21, 265-275.

Chen, H., & Pontius, R. G. (2010). Diagnostic tools to evaluate a spatial land change projection along a gradient of an explanatory variable. Landscape Ecology, 25(9), 1319-1331.

Chen, Y., Li, X., Liu, X., & Ai, B. (2014). Modeling urban land-use dynamics in a fast developing city using the modified logistic cellular automaton with a patch-based simulation strategy. International Journal of Geographical Information Science, 28(2), 234-255.

Clarke, K. C., Hoppen, S., & Gaydos, L. (1997). A self-modifying cellular automaton model of historical urbanization in the San Francisco Bay area. Environment and Planning B: Planning and Design, 24(2), 247-261.

Fearnside, P. M. (2005). Deforestation in Brazilian Amazonia: History, rates, and consequences. Conservation Biology, 19(3), 680-688.

Feng, Z., Liu, D., & Yang, Y. (2020). Modeling dynamic urban land use change by integrating an improved cellular automata model with the CLUE-S model. Environmental Modelling & Software, 124, 104600.

Guan, D., Gao, W., Watari, K., & Fukahori, H. (2011). Land use change of Kitakyushu based on landscape ecology and Markov model. Journal of Geographical Sciences, 21(1), 150-168.

Houet, T., Pigeon, G., & Bégin, Y. (2016). Bridging a gap between cellular automata and agent-based models for integrated modeling of social-ecological systems. Environment and Planning B: Planning and Design.

Lambin, E. F., Geist, H. J., & Lepers, E. (2003). Dynamics of land-use and land-cover change in tropical regions. Annual Review of Environment and Resources, 28, 205-241.

Malhi, Y., Roberts, J. T., Betts, R. A., Killeen, T. J., Li, W., & Nobre, C. A. (2008). Climate change, deforestation, and the fate of the Amazon. Science, 319(5860), 169-172.

Mas, J. F., Kolb, M., Paegelow, M., Olmedo, M. T. C., & Houet, T. (2014). Inductive pattern-based land use/cover change models: A comparison of four software packages. Environmental Modelling & Software, 51, 94-111.
Oliveira, P. J. C., Asner, G. P., Knapp, D. E., Almeyda, A., Galván-Gildemeister, R., Keene, S., ... & Smith, R. C. (2007). Land-use allocation protects the Peruvian Amazon. Science, 317(5842), 1233-1236.

Pontius, R. G., Huffaker, D., & Denman, K. (2001). Useful techniques of validation for spatially explicit land-change models. Ecological Modelling, 135(2-3), 147-186.

Soares-Filho, B. S., Cerqueira, G. C., & Pennachin, C. L. (2002). DINAMICA—a stochastic cellular automata model designed to simulate the landscape dynamics in an Amazonian colonization frontier. Ecological Modelling, 154(3), 217-235.