Master thesis

Regionalisation of the Brazilian Amazon basin for improved land change modelling

Merret Buurman

Supervisor: Prof. Dr. Edzer Pebesma
Second supervisor: Prof. Dr. Gilberto Câmara
Institute for Geoinformatics

Abstract

The Brazilian Amazon rain forest is the world's largest tropical forest and one of the places with the highest biodiversity on Earth. Since the 1970s, large forest areas have been removed, resulting in cutting about 18 % of the original forest. Understanding the causes of deforestation is necessary to set up adequate public policies to control this process. However, since the Brazilian Amazon forest occupies an area larger than Europe, there are significant regional differences in the causes of forest removal.

In order to better understand theses causes, this study investigates how much statistical models of deforestation can be improved by subdividing the study area into subregions (regionalisation).

Various sets of subregions are created using a graph-based regionalisation algorithm. Multiple linear regression models are fitted to all subregions. The performance of the regionalisation is evaluated using an error measure and compared to the results of fitting a single model to the entire study area. Results show that in general the predictions are improved by applying regionalisation. Subdividing into 9 regions improves more than subdividing into 3 regions. The best results are obtained by regionalisation using deforestation or land cover attributes. Care must be taken when using subregions in predicting deforestation for different points in time than the one for which the models were fitted.

Contents

1 Introduction **1**

2 Methodology **5**

2.1 Overview 5

2.2 Study area and data 5

2.2.1 Protected and indigenous areas (2 variables) 8

2.2.2 Farm size percentages (4 variables) 8

2.2.3 Transport costs (2 variables) 8

2.2.4 Agricultural attributes (3 variables) 9

2.2.5 *MODIS* land cover (6 variables) 9

2.2.6 *PRODES* land cover (3 variables) 11

2.2.7 *TerraClass* land cover (8 variables) 11

2.3 The statistical model used in this study 12

2.4 Evaluation metric 14

2.5 Regionalisation: Dividing space into subregions 16

2.5.1 Overview of regionalisation techniques 16

2.5.2 The *SKATER* algorithm 17

2.5.3 Advantages and disadvantages of the *SKATER* algorithm . 18

3 Results **20**

3.1 Partitions that were created and compared 20

3.1.1 Spatial meaningful partitions 21

3.1.2 Spatial random partitions 22

3.1.3 Non-spatial subsets 22

3.2 Comparison scenarios 22

3.3 Performance of the partitions under scenarios A and B (modelling for understanding the status quo) 24

3.4 Performance of the partitions under scenarios C (modelling for estimating future deforestation) 26

3.5 Ranking of the partitions in various scenarios 29

4 Discussion **36**

4.1 General discussion of the effects of regionalisation 36

4.2 Discussion of the best partitions and comparison with reference partitions 37

4.2.1 Why does the partition by the explanatory variables not perform so well? 40
4.2.2 Why does the partition by the agricultural variables not perform so well? 40
4.2.3 The partitions into 9 regions 41
4.2.4 Comparison with the Becker regions and the federal states . 41
4.3 Discussion of the findings from scenario C 42

5 Conclusions and further work **49**

A Graphics of the errors of all partitions in the scenarios A, B, C1 and C2, and maps of land cover variables used for regionalisation **56**

List of Tables

1 The explanatory variables used for statistical modelling. 7
2 The aggregation of the IGBP global vegetation classification classes used in *MODIS*. 10
3 Overview over the 66 partitions 23
4 Amounts of accumulated deforested area in 2002 and 2012, estimated and real. 28
5 Best partitions in all scenarios 30

List of Figures

1 Map of accumulated deforestation in the Amazon area. 3
2 Rates of deforestation in the Brazilian Amazon rain forest from 1998-2012, broken down by state. 3
3 The study area 6
4 The variable selection process. 13
5 The effect of subsetting on the R^2 15
6 The chaining effect in the *SKATER* algorithm 18
7 Contiguity definition in the *SKATER* algorithm 19
8 The partition proposed by Becker (source: [5], adapted). 20
9 Comparison of the adjusted R^2 and the overall error in 2002 25
10 Boxplots of the error depending on the number of regions of the partitions in 2002 26
11 Boxplots of the error depending on the randomness of the partitions in 2002. 26
12 Boxplots of the error depending on the number of regions and the randomness of the partitions in 2002. 27
13 The misestimation of the overall deforestation sum in the different partitions and in the individual subregions 29
14 Visualisation of the eight rankings. 31
15 Visualisation of the eight rankings by the partitions' performances relative to the undivided study area. 33
16 Visualisation of the eight rankings, sorted by the performance in scenarios A and B. 34
17 Rankings of the partitions into 3 regions in scenarios C1 and C2 . . 35
18 Rankings of the partitions into 9 regions in scenarios C1 and C2 . . 35
19 The accumulated deforested area and the deforestation rates in 2002 and 2012 (source INPE/PRODES). 38
20 The eight best partitions into 3 regions for scenarios A and B. 39
21 The partition created by all explanatory variables and the two explanatory variables that contributed mostly to this spatial pattern: Indigenous areas and the percentages of farms in size class 0 to 0.2 ha. 44
22 The partition created by using the attributes planted corn and soy area and number of cattle in the years 2002-2012. 45
23 The variables corn, soy, cattle in 2002 and 2012. 46

24 The eight best partitions into 9 regions for scenarios A and B 47
25 The federal states of Brazil overlapping the study area. 48
26 The errors of all partitions in 2002 (scenario A) 57
27 The errors of all partitions in 2012 (scenario B) 58
28 The errors of all partitions for predicting 2012 based on models fitted to 2002 (scenario C1) 59
29 The errors of all partitions for predicting 2012 based on models fitted to 2002, using a correction factor by region (scenario C2) . . . 60
30 The deviations of the predicted deforestation amounts from the real values for 2012 for all the partitions. 61
31 Examples for the *MODIS* land cover data used for partitioning the study area. 62
32 The *TerraClass* land cover data used for partitioning the study area (source: INPE, adapted). 63

1 Introduction

The Brazilian Amazonia rain forest covers an area of 4 million km^2 [13]. This large area of pristine rainforest has high biodiversity [7] and provides important ecosystem services. At the same time, there is a high pressure on this region. Brazil's rising economy and increasing population [4] leads to a higher demand of land for economic use. During the last decades, close to 740.000 km^2 have been deforested [19].

Obviously, there is a conflict between environmental preservation and economic development. In order to reconcile these two important things, good public policies and informed decision-making is crucial. Good decision-making depends on good data and careful analysis.

It is important to know which driving factors drive the deforestation. For this aim, much research has on the dynamics and the drivers of deforestation has been done. Many studies rely on statistical analysis for analysing the driving factors of deforestation. They have related deforestation to various variables that are assumed to represent drivers of change. Most statistical analysis of deforestation [1, 22, 3, 32, 28, 26, 33] take accumulated deforestation as the dependent variable, and use data derived from census as the independent variables. These independent variables include, for example, production of agricultural commodities, cattle raising activities, number of settlements, transportation costs, land tenure, fertility and protected areas.

It is important to distinguish between *proximate causes* and *underlying driving forces* [24]. Proximate causes are those associated to an individual's decision to transform the land cover, which include pasture expansion for cattle production, large-scale agriculture, timber industry and smallholder settlements. The underlying forces are those factors that work at a larger scale, such as demographic and technological change, global trade and policy and institutional factors. When doing a statistical analysis of causes of deforestation, it is usually the case that one has data that comes from remote sensing images or from census data collection. Such data are mostly related to proximate factors. As a result, there is an limited explanatory power related to statistical analysis, in that we may be able to study the proximate causes and it is much harder to represent the underlying factors.

In her paper "Geopolítica da Amazônia" [5], the geographer Becker draws a more comprehensive picture of the role of the Amazon area. She explains the dynamics of the Amazon region in a geopolitical context and emphasizes the role of political and societal actors. For a long time, the Amazon was seen as a large area

that needs to be occupied in order to ensure national territorial integrity. Nowadays, international actors such as market pressures and environmental preservation organisations have increased their influence on what is happening locally. Becker underlines the role of market pressures and the global interests at stake in the Amazon as a consequence of the globalisation of the economy, but also in the light of the trend towards mercantilisation of nature, shown e.g. by the Kyoto protocol and attempts of commercialising biodiversity and water. She emphasizes the importance of the different actors and the ability and potential of political and local actors to restrict the free reign of market forces in the area, to counter the influence of actors such as agribusiness companies, which have had a big influence in the recent decades.

A large part of the history of land use change in the Amazon area is related to agricultural expansion, carried out by different types of actors, which have different effects [31]. From 1990 to 2005, 110,000 km^2 were deforested in Mato Grosso state in the southern part of Amazonia. Such deforestation was associated to a large migration from farmers from the South of Brazil. This resulted in a large expansion of the soy production area and contributed to Brazil's exports. In 2008, Brazil produced 58 million tons of soybeans. Mato Grosso accounted for 15 million tons (25% of total). The other states in Amazonia have no significant contribution to the production of grains [27]. Soy and corn expansion of production in Amazonia has slowed down since 2000. Several reasons for this are identified: (a) decreased migration from the South; (b) emphasis on productivity improvements instead of area expansion; (c) newly available areas have worse connection to markets and unfavorable soil conditions; (d) external market pressures for avoiding further deforestation. From 1970 to 1980, Mato Grosso's population almost doubled from approximately 600,000 to 1,130,000 people. From 1980 to 1990, it increased to 2,000,000 people. Growth was smaller in the next decade, reaching 2,500,000 people in 2000. Less migrants means less pressure for new land. Furthermore, Greenpeace and ABIOVE (Brazilian Association for Vegetable Oil) have signed an agreement in 2006 (the Soy Moratorium), where the soybeans exporters have declared that they would not carry out any more deforestation. The Soy Moratorium has been renewed yearly since 2006. Soy and corn production account for about 5% of the total deforestation in Amazonia [19].

Deforestation data from the *Instituto Nacional de Pesquisa Espacial* (INPE, the Brazilian National Space Research Institute), broken down by state, is shown in figure 2. It shows that most of the forest being cleared since 2006 is located in the state of Pará, whose rural economy is largely based on cattle raising and intensive

logging [6, 31]. These activities are unsustainable. Recent research showed that relative standards of living, literacy and life expectancy increase as deforestation begins but then decline as the frontier evolves [34].

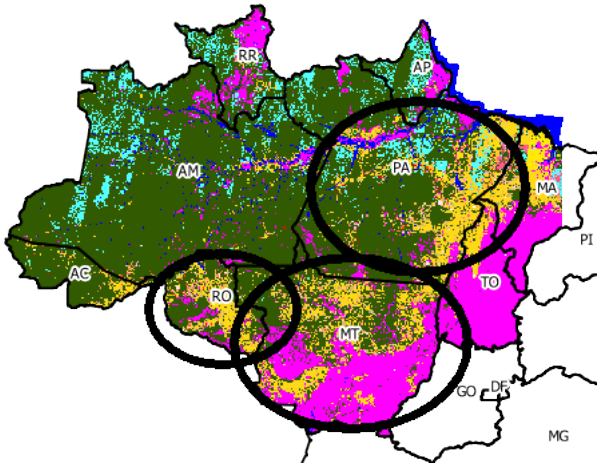

Figure 1: Map of accumulated deforestation in the Amazon area. Remaining forest is shown as green, deforestation in shades of yellow to red, savanna in magenta, clouds in blue. The highlighted areas are Pará (in the north; cattle, timber and settlements), Rondônia (southwest; cattle and settlements) and Mato Grosso (southeast; large soybeans production) (source: PRODES/INPE, adapted).

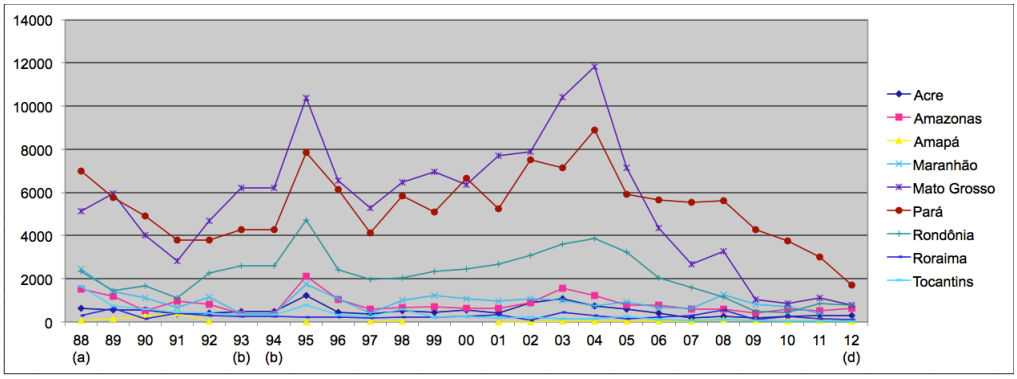

Figure 2: Rates of deforestation in the Brazilian Amazon rain forest from 1998-2012, broken down by state (source: PRODES/INPE).

The three main areas of occupation in Amazonia are the states of Pará, Rondônia, and Mato Grosso, shown on figure 1. In each state, there are different dominant driving forces for deforestation. In Mato Grosso, deforestation is associated to grains and cattle raising. In Pará, expansions combine farms for cattle, illegal timber extraction, and land speculation. In Rondônia, cattle raising is combined with small-scale settlements. Thus, in each of the states where there has been most deforestation, the causes are different [27, 6, 31].

Those studies show that deforestation is a highly spatially heterogeneous process that calls for analyzing the regions separately. For this reason, [35] acknowledges this and splits the entire Amazon basin into 47 subregions, using different cellular automata model for a deforestation forecast 2050. In [1], three subregions are used, which were introduced by Becker [5]. She identifies three macroregions: The Consolidated Arch (*Arco de povoamento con-solidado*), Central Amazonia and Occidental Amazonia. The Arch includes areas from northeastern Amazonia (Belém, Pará, Maranhão) over the eastern and southeastern area and includes the southern border of the Amazon rainforest until Mato Grosso and Rondónia. It is the more urban area, with cities, population and infrastructure well established. Occidental Amazonia is the most preserved, with its population concentrated in Manaus. Central Amazonia is assumed to be the most vulnerable area. Infrastructures axes cut across it and the most recent deforestation frontiers are located here.

In [1], these regions are used for fitting separate statistical models. They find different driving factors to be important in the various subregions. For example, protected areas are more relevant in the Arch than in Central Amazonia, while not showing a significant effect in Occidental Amazonia. From this, they deduce recommendations on localisations of potential protected areas.

This work shows the usefulness of subdividing the Brazilian Amazon rainforest, which occupies an area larger than Europe. However, the regions proposed by Becker have not been evaluated or compared against other regionalisations. In [1]'s work, the models fitted to the subregions show that different driving forces are dominant in the various areas, but it is not clear by how much the explanation of the spatial pattern of deforestation improved by these subregions. Thus, two questions arise: Does the explanation of deforestation improve by regionalisation, and how much? And is there regionalisations that do so more than the ones proposed by Becker?

Those two questions are addressed in this work. We propose to evaluate the usefulness of subdividing the space into federal states, into the regions proposed by Becker, in statistical deforestation modelling. We aim to find out whether other regionalisations perform better, and find out by how much the regionalization improves the goodness of fit. We will do this by creating regionalizations in an objective way based on various attributes known to be related to land change in Amazonia and evaluating their quality using linear statistical models of deforestation.

2 Methodology

2.1 Overview

The objective of this study is to evaluate the influence of regionalisation on the result of statistical deforestation modelling. For this, various subdivisions of the study area into regions are compared.

To evaluate how much regionalisation improves the results of deforestation models, the following steps are carried out repeatedly. The study area is split into subregions. A statistical model is fitted to each subregion. Using these models, the deforestation is predicted for each subregion separately. The combined predictions of a set of subregions are compared to the real deforestation, measured by satellite imagery.

In the following, we call a combination of subregions that cover the whole study area a partition. 66 partitions were created. They were compared among each other and to the deforestation predicted by applying a model to the entire study area.

This section is structured as follows. First, the study area, its representation in a spatial database and the spatial data used for regionalising and for deforestation modelling are described (section 2.2). The statistical model itself is described in section 2.3 and the metric used for evaluating the models is introduced in section 2.4. The regionalisation algorithm that is used for creating the partitions is introduced in section 2.5.

2.2 Study area and data

The study is applied to the Brazilian part of the Amazon rainforest, called the Amazon Rainforest Biome. This is a part of the Brazilian administrative region Legal Amazon. Approximately 20 % of the Legal Amazon area is covered by other ecosystems, mainly by the Cerrado, a savannah ecosystem, and by a small part of the Pantanal, a wetlands area (see fig. 3). The PRODES dataset 2.2.6, which is used as reference in this study, does not map deforestation in these areas, so the regression model would be biased by including these areas in the study. The reason to restrict the study to the Brazilian part of the Amazonian rainforest, ignoring the neighbouring countries, is the availability and homogeneity of data in Brazil.

The Rainforest biome has an area of approximately 4,196,943 km^2 and covers approximately 49.29 % of the Brazilian national territory [13]. The source of the

Figure 3: The biomes of Brazil (grey areas, AMZ = Amazon rainforest, CER = Cerrado, CAAT = Caatinga, PTN = Pantanal, MAT = Mata Atlântica, PMP = Pampa) and the Legal Amazon area (thick black line). In this study, only the Amazon rainforest biome, painted in dark grey, is used (source: IBGE, adapted).

spatial data on biomes and the Legal Amazon is the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística*, IBGE).

The study area is divided into 6836 grid cells of 25 by 25 km containing all the variables used for modelling and for regionalisation. To make sure that all of them have the same area, all the data was reprojected to an equal-area Albers Conic projection centered on the study area. 28 variables are used for statistical modelling and/or regionalisation. 14 of them are available for each year during the studied period (2002-2012), 14 were only available for single years. The variables that were used for the statistical models are listed in tab. 1.

Potential explanatory variable		**Year**	**Source**
1	percentage of farms 0-0.2 ha	2006	IBGE (Census)
2	Percentage of farms 0-0.2 ha (log)		
3	Percentage of farms 0.2 - 5 ha		
4	Percentage of farms 0.2 - 5 ha (log)		
5	Percentage of farms 5 - 500 ha		
6	Percentage of farms 5 - 500 ha (log)		
7	Percentage of farms >500 ha		
8	Percentage of farms >500 ha (log)		
9	Number of heads of cattle	Yearly	IBGE (PPM survey)
10	Number of heads of cattle (log)		
11	Area planted with corn	Yearly	IBGE (PAM survey)
12	Area planted with corn (log)		
13	Area planted with soy		
14	Area planted with soy (log)		
15	Percentage of indigenous area	Yearly	FUNAI
16	Percentage of indigenous area (log)		
17	Percentage of proctected area	Yearly	MMA
18	Percentage of proctected area (log)		
19	Transport cost to state capitals	2008	PNLT/INPE
20	Transport cost to state capitals (log)		
21	Transport cost to export ports		
22	Transport cost to export ports (log)		

Table 1: The explanatory variables used for statistical modelling.

In the following, a short description of each used variable is given. The scripts used to create the database and the resulting grid cells are available on the at-

tached DVD.

2.2.1 Protected and indigenous areas (2 variables)

These attributes indicate how much of the cell is covered by protected respectively indigenous areas in a specific year. The range is from 0 to 1. The data is freely available from the Brazilian Ministry of Environment (*Ministério do Meio Ambiente*) [21] and the Brazilian National Indian Foundation (*Fundação Nacional do Índio*) [20] as polygon data, from which the fractions per cell were computed for each year.

2.2.2 Farm size percentages (4 variables)

The farm size variables indicate how much of the farm area falls into each of four farm size classes (0 to >0.2 ha, 0.2 to >5 ha, 5 to >500 ha, 500 ha and above). The values are percentage of the overall farm area, so the values of the four farm size attributes add up to 100. The data is available by municipality. Each grid cells inside a municipality gets the same value. For grid cells that are shared between several municipalities, a weighted average between the values of the participating municipalities was computed. The data is based on a agricultural census in the year 2006 and freely available from the Brazilian Institute of Geography and Statistics (*Instituto Brasileiro de Geografia e Estatística*, IBGE) [10].

2.2.3 Transport costs (2 variables)

The transport cost was computed for each cell by an algorithm developed by INPE. There are two types of transport costs: (a) The cost to the closest export port, and (b) the cost to the capital of the state. Both attributes are in Brazilian Reais (BR$). The costs were determined by assigning different per-kilometre costs to the different types of roads in the network and computing the cumulative cost until reaching the nearest export port or state capital. The cost of the distance from each grid cell's centroid to the closest road is twice as high as the highest road cost to symbolize off-road transport. Due to connectivity problems after cartographic reprojection of the input data, four grid cells received exaggeratedly high costs. They were assigned their neighbour cell's cost values manually. The underlying data about the transport network is from the Brazilian National Transports and Logistics Plan (*Plano Nacional de Transporte e Logística*) from 2008 as was provided by INPE.

2.2.4 Agricultural attributes (3 variables)

These attributes indicate how much area inside a grid cell is used for corn and soy culture (in ha) and how many bovines (cattle) are present (number of heads). The area of planted corn and soy and the cattle numbers are available by municipality from IBGE. Cattle data is from the *Produção da Pecuária Municipal* (PPM) [15] and plant data from the *Produção Agrícola Municipal* (PAM) [14], which are a yearly agricultural samples. The data is freely available from IBGE [11][12].

To downscale the municipality-level values to the individual grid cells, information from satellite-based land cover data (*MODIS* land cover, see section 2.2.5) and information on protected and indigenous areas was used. The planted area was assumed to be homogeneously distributed over all the area classified as ”croplands” in a municipality, excluding the area that falls into indigenous or protected areas. The cattle was assumed to be homogeneously distributed over all the area classified as ”pasture” or ”natural pasture” in a municipality, also excluding protected and indigenous areas. The data by municipality was available yearly. The distribution was carried out for each year using yearly land cover data and the yearly indigenous and protected areas.

2.2.5 *MODIS* land cover (6 variables)

For each grid cell, the percentage of the area covered by croplands, pasture, forest, natural pasture, water and other land cover was computed from yearly *MODIS* satellite-based land cover data. *MODIS* (Moderate Resolution Imaging Spectroradiometer) is an instrument for satellite-based land cover imaging on board of several satellites of the *US National Aeronautics and Space Administration* (NASA), from which various land cover datasets are derived. The *MODIS* data used in this study is the *MODIS* Land Cover Type product (*MCD12Q1*, [29]) which is a yearly land cover mapping with the resolution of 500 m. The data uses the land cover classes of the International Geosphere-Biosphere Programme (IGBP) global vegetation classification scheme. We aggregate them according to our necessity according to tab. 2. The *MODIS* land cover variables were used for the regionalisation (not for the statistical models). Their values range from 0 to 100 and add up to 100 in each grid cell. The *MODIS* imagery used in this study was provided by INPE as a mosaic covering the whole country.

Aggregated land cover classes	**Original *MODIS* land cover classes**
Forest	Evergreen needleleaf forest Evergreen broadleaf forest Deciduous needleleaf forest Deciduous broadleaf forest Mixed forest
Croplands	Croplands Cropland/natural vegetation mosaic
Pasture	Grasslands Permanent wetlands
Natural pasture	Closed shrublands Open shrublands Woody savannas Savannas Barren or sparsely vegetated
Urban and built-up	Urban and built-up
Water and remaining	Water Snow and ice (Unclassified) (Fill Value)

Table 2: The aggregation of the IGBP global vegetation classification classes used in *MODIS*. Not all of the classes in the right column neccessarily exist in the study area, e.g. snow and ice. The class *Permanent wetlands* was included in the aggregated class *Pasture* to cover the Pantanal region before that biome was excluded from the analysis.

2.2.6 *PRODES* land cover (3 variables)

The *PRODES* dataset [16] provides yearly mapping of deforested areas. For statistical modelling, the accumulated deforested area for each year between 2002 and 2012 is used. In regionalisation, the yearly deforestation rates are used, too. The rates are the areas that are deforested in a specific year. Furthermore, we use the accumulated deforestation divided by the area considered suitable for forest. Area suitable for forested is the whole cell area except the area classified as non-forest, water and cloud by *PRODES*. This removes the bias introduced by low deforestation values in grid cells that mainly consist of water or savannah vegetation. The *PRODES* data is freely available from INPE [18].

2.2.7 *TerraClass* land cover (8 variables)

The *TerraClass* land cover dataset for the year 2010 provides information about the land cover in the areas classified as deforested in the *PRODES* dataset. In *PRODES*, areas that are once mapped as deforested are not re-analyzed in subsequent years. Thus, *PRODES* cannot capture reforestation or secondary vegetation. The *TerraClass* project was created to analyse the land use of deforested areas after deforestation. It considers the areas classified as deforested in the *PRODES* dataset and assigns them one of the classes secondary vegetation, reforested, agriculture (distinguishing between annual and permanent agriculture and oil palm culture), pasture (distinguishing between various degrees of degradation), urban, mining and non-forest [19]. Secondary vegetation encompasses regenerated tree and shrub vegetation after a human induced removal (clear-cut) of the original forest. Small patches of secondary vegetation after selective logging activities are not mapped, as they are considered forest by *PRODES* and thus excluded from *TerraClass* mapping. Reforested areas are areas that underwent planting of tree species for commercial exploitation.

In this study, eight aggregated classes were used for regionalisation: Annual agriculture, permanent agriculture, secondary vegetation, non-forest, forest including reforested areas, pasture, other land cover (including urban areas, water and mining) and outside study area (including non-observed areas and cloud cover). Areas classified as ”agropecuária” (agriculture) were included in the class *Permanent agriculture*. The TerraClass 2010 data is freely available from INPE ([17]). The mosaic of the entire study region used in this study was provided by INPE.

2.3 The statistical model used in this study

The objective of this study is to evaluate whether and how much subdividing the study area improves deforestation modelling by linear regression models. In this section we present the details of the multiple regression models that were used.

Regression models have been used in various studies to relate the deforestation in the Brazilian Amazon basin to spatial variables that are assumed to represent the underlying drivers of change (proxies).

In land change modelling, we are more interested in finding out what drives the changes in land use/cover than in explaining a static pattern. Thus, it would make sense to select the change in forest cover during a specific period as the response variable and model its dependence on some explanatory variables. This way, the different drivers of change that are relevant during a specific time period could directly be identified. However, during short time periods, only relatively little area is deforested. Even though a large area may have a potential for deforestation, the demand for deforestation may not be that high during that limited time period, so much of the area with a large potential is not deforested inside the time period. By modelling the accumulated deforestation, the correlation with the driving factors is much clearer, as a larger fraction of the area with high potential for deforestation is actually deforested. Thus, we select the accumulated deforested area (available in the *PRODES* dataset) as the response variable.

The explanatory variables of a deforestation model are variables that are assumed to be proxies for deforestation drivers. They do not drive deforestation themselves, but they represent underlying drivers of deforestation. In this study, data on farm sizes (4 variables), transport costs (2 variables), protected and indigenous areas (2 variables), number of cattle and planted area of soy and corn (3 variables) are potential explanatory variables (see tab. 1). Variables that are strongly interdependent with the response variable, such as land cover attributes derived from satellite imagery, are not considered for explanatory variables.

Not all of the proxy variables necessarily have a high correlation with the accumulated deforestation in the whole study area. As we are fitting models to different subregions of the study area, different driving forces may be more or less important. Which of the above-mentioned proxies are applied as explanatory variables in the statistical models in this study is determined in several steps, see fig. 4. First, the log-transformations of all 11 variables are computed, as the relationships between the drivers and the deforestation is often not linear. These 22 variables are then checked for high correlations with each other. From any

pair of variables that has an absolute correlation above 0.85, the one with the lower correlation with the independent variable is excluded. Each time a model is fit to a subregion –396 times for the year 2002 and 396 times for the year 2012 – , the remaining variables are used in an automated stepwise variable selection process to select the variables that are relevant for that area. The selection starts by fitting the model using all offered variables. In each subsequent step, one variable is selected to be excluded. The variable that is excluded is the one that improves the model least, as evaluated by Akaike's Information Criteria. The variable selection is repeated for each subregion so that the optimal combination of explanatory variables for each subregion is found. Table 1 lists the available variables and shows which ones are used as candidates for explanatory variables. From these, the relevant ones for each subregion are selected.

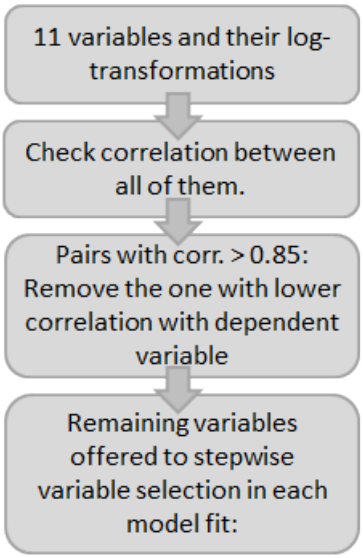

Figure 4: The variable selection process.

A term representing spatial autocorrelation is not included in the model. Deforestation is known to be strongly autocorrelated. However, if we explain deforestation as a function of itself, such an outcome is of little use for understanding the drivers of deforestation [30] and for finding ways of reducing it, which is the ultimate goal of land change research. Furthermore, the effect of spatial autocorrelation is expected to be reduced by the regionalisation.

There is a number of assumptions that apply to linear regression models and which do not hold in this case. The observations that are used for fitting the model are not independent, as they are spatially neighbouring grid cells with a strong spatial autocorrelation. It is also more than questionable whether the

relationship between the response variable and the explanatory variables is linear and has a constant variance.

2.4 Evaluation metric

When linear regression models are compared, the usual measure is the determination coefficient R^2, or the adjusted R^2 for multiple regression, which tell us how much of the variation in the dependent variable can be explained by the explanatory variables [23, 36].

The R^2 metrics describe individual statistical models. When we divide the study area in subregions, each subregion will be associated to a different model. We are not interested in the performance of each individual model but of the combination of models for the whole study area. We have a set of several models that we want to compare with other sets of several models. We could use an average of the adjusted R^2 as a goodness of fit metric. However, since the subregions' area sizes are very different, we would need to apply weighting. The problem is that it would be hard find a balanced set of weights. If we weight adjusted R^2 of the subregions by area, large areas are given a lot of weight, but they do not necessarily have high deforestation amounts. Thus, their adjusted R^2 gets a lot of weight, while they are not of much relevance in the deforestation estimation. Thus, using an area-weighted adjusted R^2 does not provide a good criteria.

Another reason for not using the adjusted R^2 is that it depends on the range of data the model is fitted to. When fitting a model to subsets of data, the correlation in those subsets may be lower than the overall correlation, but the quality of the fit regarding the whole dataset (i.e. the combined residuals of the various models) could still be better than when applying a single model.

As a demonstration, consider fig. 5. On the left, a single model is fitted to a point cloud, with an R^2 of 0.85. After splitting the point cloud into two subsets, the models fitted to them have R^2 values of 0.24 and 0.76, although the lines are fitted closer to the subsets. Compared to the small range of the data, the variability is quite high.

For comparing the quality of fit of the different regionalisations, we need a single number that requires no weighting. This value should be directly comparable between the models fitted to the various partitions and the whole study area.

For this, we use the statistical models fitted to that region to predict the accumulated deforestation (i.e. the response variable) in each grid cell of each sub-

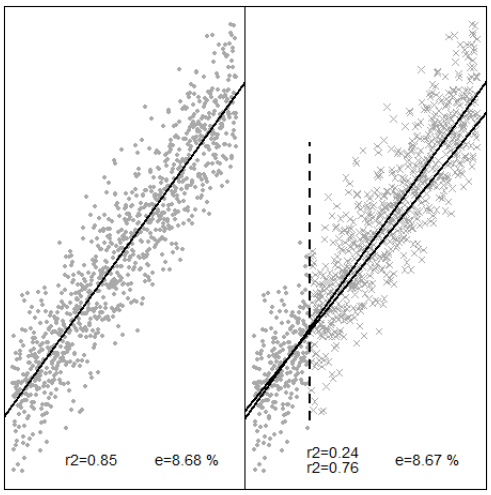

Figure 5: The effect of subsetting on the R^2: On the left, a single linear model is fitted to the data (R^2=0.85). On the right, the data is split and two linear models are fitted (R^2= 0.24 and R^2=0.76). The R^2 of both models on the right is lower than the R^2 of the single model, although the fitted lines are closer to the data.

region separately. The predictions are combined for the whole study area and compared to the deforestation that really occurred. We compute the difference between the predicted and the real deforestation amount in each grid cell (residuals). The absolute values of these residuals are summed and divided by two to get the overall amount of misplaced deforestation. We divide the overall amount of misplaced deforestation by the sum of really occurred deforestation to find out the percentage of the total amount that was misplaced.

$$e = \frac{0.5 \sum_{i=1}^{n} abs(d_{reali} - d_{predi})}{\sum_{i=1}^{n} d_{reali}} * 100 \quad (1)$$

where

d_{reali} = Real accumulated deforestation in the grid cell i

d_{predi} = Predicted accumulated deforestation in the grid cell i

The residuals include both overestimation and underestimation. If the overall amount of deforestation (predicted and real) is the same – as it happens in linear models, as the sum of residuals is zero – each underestimation in one place results in an overestimation in another place. By counting both, we count double. That is why the factor 0.5 is introduced in the equation. As a demonstration, consider the following example: 100 units of deforestation were predicted, of which half is placed incorrectly. Then there are 50 units of overestimation where this deforestation was wrongly placed, and 50 units of underestimation where it should

have been placed. Without the factor 0.5 in the error equation, 100 wrong units would be counted, and the error would be 100 % – although of the 100 units of deforestation, only 50 were placed wrongly and an error of 50 % would be more meaningful.

This error metric has the following characteristics. If none of the predicted values deviates from the real values (perfect fit), its value is 0. Its value is 100 % if all deforestation is placed in the wrong location and if the predicted and real amounts of deforestation are the same. (It can exceed 100 % if negative predictions occur – in this case, the absolute values of the residuals have no limit and can exceed the total of deforestation.)

If the predicted and real amounts of deforestation differ, this metric should not be applied. Its values are not meaningful in those cases. For example, if all deforestation is placed in wrong places, smaller overall predicted amounts would lead to smaller errors. An underestimation of the total amount would have a smaller error than an overestimation of the total amount, even if the amount of mis-estimations is the same. Thus, meaningful comparisons of the model performance cannot be achieved anymore.

Going back to the example in fig. 5, we can see that the errors are 8.68 % and 8.67 %. As the observations in both subsets stem from the same process, we would expect approximately the same error value, independently of how many subsets we apply. Of course it decreases slightly because of better fitting to local subsets, but the overall values are comparable. Strong error decreases would indicate a different behaviour of the subsets, e.g. because the underlying process might be nonlinear.

2.5 Regionalisation: Dividing space into subregions

2.5.1 Overview of regionalisation techniques

Regionalisation is the division of an area into regions so that the regions are as homogeneous as possible inside and the difference between the regions is maximized. The area is composed of spatial entities, which are aggregated based on their similarity in one or several attributes. For example, a nation is composed of municipalities which can be aggregated into regions based on their similarity in population density.

Grouping a large number of objects into a subgroups is a common task in data analysis and frequently done by clustering analysis. In regionalisation, there is an additional constraint: The groups have to be contiguous, i.e. the entities forming

them have to be neighbours in space. For this, several types of methods exist [25]. Some methods use a non-spatial clustering technique first and then apply this spatial constraint subsequently. Other methods incorporate space by using a spatial closeness measure into the non-spatial clustering process. A third group of methods uses trial-and-error to optimize a random initial regionalisation, and a fourth group of methods is using the spatial constraint as a basis, in the sense that the spatial contiguity already affects the choice of objects whose similarity is evaluated. The *SKATER* algorithm, which is used in this study, belongs to this latter group.

2.5.2 The *SKATER* algorithm

The *SKATER* algorithm [2] is a graph-based regionalisation algorithm. It is performed in three steps. First, a connectivity graph is created from the spatial entities of the study area. In this graph, each node is a spatial entity and it is connected to all adjacent spatial entities by edges. The cost of the edges represents the dissimilarity of the spatial objects. The dissimilarity of the objects is measured by using the Euclidean distance in attribute space between the attribute vectors of both spatial objects.

As a next step, a spanning tree is created from this graph. This is a subgraph of the connectivity graph which contains all the *n* nodes, but only *n-1* edges, so that all nodes are connected to each other, and the removal of any one edge leads to dividing the spanning tree into two separate subgraphs. Spanning trees are not unique. Various spanning trees can be constructed from the connectivity graph. The one used in the *SKATER* algorithm is the one with the minimal sum of dissimilarity over all edges, called the minimum spanning tree. It is constructed by starting at one node and adding one node after another to the tree. At each step, the node that is added is the one with the least expensive direct connection edge to one of the previously chosen nodes. Unless there are neighbours of a node that have the same dissimilarity, the minimum spanning tree is unique.

Finally, the subdivision into regions is achieved by iteratively removing edges from the minimum spanning tree. Each removal results in disconnected subgraphs, which correspond to the disconnected (but adjacent) regions. Each removal subdivides a region into the most homogeneous subregions. The heterogeneity of a region is measured by the intracluster square deviation, which is the sum of squared deviations of the attribute values of each object from the average attribute values of all objects in that region. So at each step, the edge should be

removed that splits the graph into subgraphs with the lowest intracluster square deviations. As evaluating this for each and every possible edge is computationally intensive, a heuristic is applied. It starts with evaluating the edge removal that splits the graph into subgraphs of similar size, and then examines neighboring edges. The selection of the neighboring edges to be evaluated next is based on a balancing function that aims at finding the most homogeneous subregions as well as avoiding regions that are very unbalanced in size.

2.5.3 Advantages and disadvantages of the *SKATER* algorithm

The advantage of this algorithm is that the spatial constraint is inherent to the clustering procedure [2]. In contrast, when using algorithms where space is merely included as one of the attributes along with the other attributes, it is more difficult to ensure spatial adjacency [25]. Another advantage of the *SKATER* algorithm is that the number of regions can be controlled, as the regions are obtained by subsequently splitting the graph [25]. This is an advantage in this study, as we want to evaluate the effect of splitting the study area into predetermined numbers of regions.

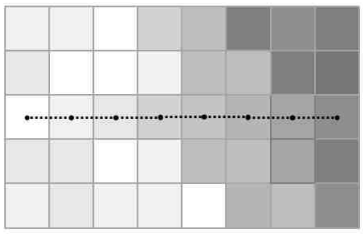

Figure 6: The chaining effect in the *SKATER* algorithm.

[25] argues that an important shortcoming of the *SKATER* approach is the so-called chaining effect. During the generation of the minimum spanning tree, nodes to be added are chosen based on their similarity to nodes already chosen. For this, only the similarities between two single nodes are considered. This can result in chains of contiguous points where the first and the last are not necessarily similar (see fig. 6).

[25] also criticizes *SKATER*'s contiguity definition. In *SKATER*, only the edges that connect directly adjacent nodes are used to compute the similarity between clusters. However, two clusters might be connected by an edge which is a very low cost, but connects two nodes that are not directly adjacent, but in adjacent regions (see fig. 7). Using *SKATER*, such clusters cannot be merged, as those

edges are not present in the connectivity graph. [25] argue that they should be connected.

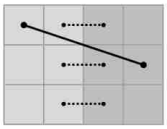

Figure 7: Contiguity definition. In the *SKATER* algorithm, only the dashed edges can be used for merging regions. The black one is discarded.

They propose a family of algorithms that are generalised extensions of *SKATER* and which avoids these shortcomings and which they also argue to be more efficiently implemented. However, they are available as Java implementation operating on shapefiles, so applying them would have meant a considerable effort compared to the implementation of *SKATER* in the TerraLib library [8], as all the data were kept in a TerraLib database.

3 Results

3.1 Partitions that were created and compared

A total number of 66 partitions were compared among each other, and to the results obtained by applying a model to the entire study area. Half of the partitions split the study area into 3 subregions and the other half split it into 9 regions (see tab. 3). These numbers of regions were chosen in order to compare the regions with the partitions proposed by Becker [5] that were described in the introduction 1 (see fig. 8) and the federal states (9 states covering the area).

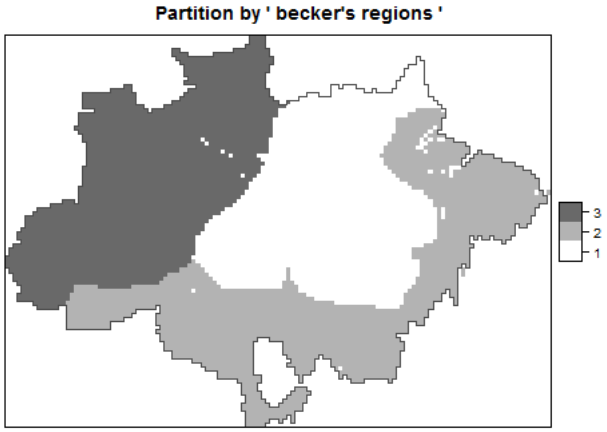

Figure 8: The partition proposed by Becker (source: [5], adapted).

For creating the partitions, we use the implementation of *SKATER* available in the TerraLib library, as it is efficient and easily applied on the data stored in a TerraLib database [8]. All attributes are scaled to range between 0 and 1000 to give them all the same weight in the dissimilarity measure. We define the spatial adjacency by a maximum distance of 26 km between the grid cells' centroids. As all grid cells measure 25 by 25 km, this ensures that the direct neighbours of each grid cell are chosen (Von-Neumann-neighbourhood), but not the grid cells that touch the grid cells at the corners (Moore-neighbourhood), whose centroids are at a distance of $\sqrt{2 * 25} = 35.4$ km. This way, we avoid regions that are only connected by a corner. Whenever available, the attributes were used for several years to ensure that the regions are not only maximally similar in space, but also in time.

Applying the algorithm on the data shows that it tends to produce regions of very heterogeneous sizes, even when applied to uniformly distributed random data, even though the heuristic explained in section 2.5 was designed to avoid this. For this reason, the regionalisations were re-run a second time, this time

with a enforced minimum region size. In this case, the implementation does not allow to specify the resulting number of regions, so several tentatives had to be carried out to reach partitions of 3 and 9 regions.

3.1.1 Spatial meaningful partitions

These partitions are created by applying the regionalisation algorithm to one or a combination of spatial attributes.

1. *Corn, soy and cattle*: Area planted with corn (ha), area planted with soy (ha) and heads of cattle; yearly values from 2002 through 2012. Three attributes, 11 years.

2. *Deforestation rates*: Yearly deforestation rates from 2002 through 2012. One attribute, 11 years.

3. *Accumulated deforestation*: Yearly accumulated deforested area from 2002 through 2012. One attribute, 11 years.

4. *Accumulated deforestation per forest area*: Yearly accumulated deforestation from 2002 through 2012 divided by the area that is suitable for forest. One attribute, 11 years.

5. *Farm sizes*: The percentages of farm area in four farm size classes (percentage). Four attributes, one year (2006).

6. *MODIS land cover*. The six land cover classes croplands, pasture, forest, natural pasture, water and other land cover from the *MODIS* land cover dataset were used on a yearly basis. Six attributes, 11 years.

7. *Residuals 2002*: The residuals of the statistical model fitted to the whole study area in 2002. One attribute, one year (2002).

8. *Residuals 2012*: The residuals of the statistical model fitted to the whole study area in 2012. One attribute, one year (2012).

9. *Transport costs*: Transport costs to state capitals and to export ports. Two attributes, one year (2008).

10. *TerraClass land cover*. The eight land cover classes aggregated from the TerraClass 2010 dataset were used. Eight attributes, one year (2010).

11. *Full explanatory variables*. In this partition, all the variables used as explanatory variables in the statistical models were used: Indigenous areas, protected areas, soy and corn areas, cattle heads, farm size percentages, transport costs. The log-transformations were excluded. Eleven attributes, 11 years (except for farm sizes percentages and transport costs, which are only available for one year).

To avoid having to write "the partition created using the set of attributes *MODIS land cover*", we will designate it by "the partition *MODIS land cover*".

3.1.2 Spatial random partitions

20 spatial random partitions are created by assigning random values drawn from a uniform distribution to the grid cells and then applying the regionalisation algorithm on these random values. By this method, we expect to obtain random subregions. However, if we do not enforce homogeneous sizes, this process results in 2 or 8 extremely small regions, and the remainder of the study area forms the last region. Therefore, approximately homogeneous region sizes were enforced.

3.1.3 Non-spatial subsets

For comparing the performance of the spatial subregions with completely random (i.e. non-spatial) subsets of the data, the process of modelling, predicting and evaluating was applied to random non-spatial subsets of the study area. The study area is randomly split 50 times into 3 subsets of approximately the same number of grid cells, 50 times into 9 subsets of the same number of grid cells, 50 times into 3 subsets of random sizes and 50 times into 9 subsets of random sizes.

3.2 Comparison scenarios

All the partitions described in the previous section are compared in three scenarios. In scenario A, the models are fitted to the values of 2002 and predictions are made using the same data. Analogously, in scenario B, the models are fitted to the values of 2012 and predictions are made for 2012. In scenario C, the models fitted to the data of 2002 are applied to the data of 2012, simulating a predicting of future deforestation from the 2002 perspective, but knowing about the deforestation drivers in 2012. In scenario A and B, we evaluate the goodness of fit of the models themselves. In scenario C, we evaluate how well the relationship between

Type of partitions	**Partitions in to 3 regions**	**Partitions into 9 regions**
Becker	1	
Federal states		1
Spatial Random	10	10
Meaningful spatial	11	11
Heterogeneous region sizes	11	11
Random non-spatial	50	50
Heterogeneous region sizes	50	50

Table 3: Overview over the 66 partitions

the accumulated deforestation and the proxies for the driving factors, expressed by the model coefficients, holds for the situation 10 years later. Scenario C has two sub-scenarios. In C1, a correction factor is applied to the predictions before computing the error, to ensure the correct overall demand amount. This is important for the error metric to provide meaningful results (see section 2.4). In C2, the correction factor is applied by subregion.

3.3 Performance of the partitions under scenarios A and B (modelling for understanding the status quo)

Subdividing the study area into smaller subregions reduces the error of the statistical models of accumulated deforestation, measured by the error metric presented in section 2.4. The model fitted to the entire study area has an error of 26.2 % in 2012, meaning that 26.2 % of the overall amount of deforestation is placed in the wrong grid cell, and 21.7 % in 2002. All the spatial partitions outperform these results (see figs. 26, 27 in the appendix for an overview of the errors of all partitions).

Fig. 9 shows the adjusted R^2 of the models of various partitions compared with the error measure. As mentioned, it is difficult to conclude the performance of the combined models using their adjusted R^2 values because the variation between the subregions of one partition can be high and no clear trend is visible when comparing the partitions.

The best results for 3 and 9 regions in 2002 is reached by the partitions using *TerraClass land cover* (errors of 18.5 % for 3 regions and 16.8 % for 9 regions), which has a better fit than the Becker regions (22.0 %) and the federal states (19.2 %).

The best result for 3 regions in 2012 is reached by the partition *Accumulated deforestation* (15.7 %), which has a better fit than the Becker regions (e = 18.1 %). For 9 regions, *Accumulated deforestation (homogeneous region sizes)* with an error of 14.4 % is the best, outperforming the partition of the federal states, which reach an error of 15.7 %.

In general, estimations using a partition into 9 regions perform better than partitions using 3 regions (see fig. 10). On average, partitioning into 3 regions results in an error of 22.1 % (2012: 18.0 %), partitioning into 9 regions in an error of 19.3 % (2012: 15.8 %)(see also fig. 12).

Fig. 11 shows the distributions of the error values for the meaningful spatial partitions (right) compared to the spatial random partitions (second from the right) and to the non-spatial random subsets of the data (second on the left), in 2002. While completely random partitions have similar error values as a single the model fitted to the entire study area, the error when using spatially random partitions are similar to the one using meaningful partitions (see also fig. 12). The pattern in 2012 is very similar with slightly lower overall errors. The error of the whole study area is 26.2, by partitioning into 3 and 9 subregions, we reach average errors of 22.1 and 19.3, respectively.

The partitions that performed best for 3 regions in 2002 are (1) *TerraClass land*

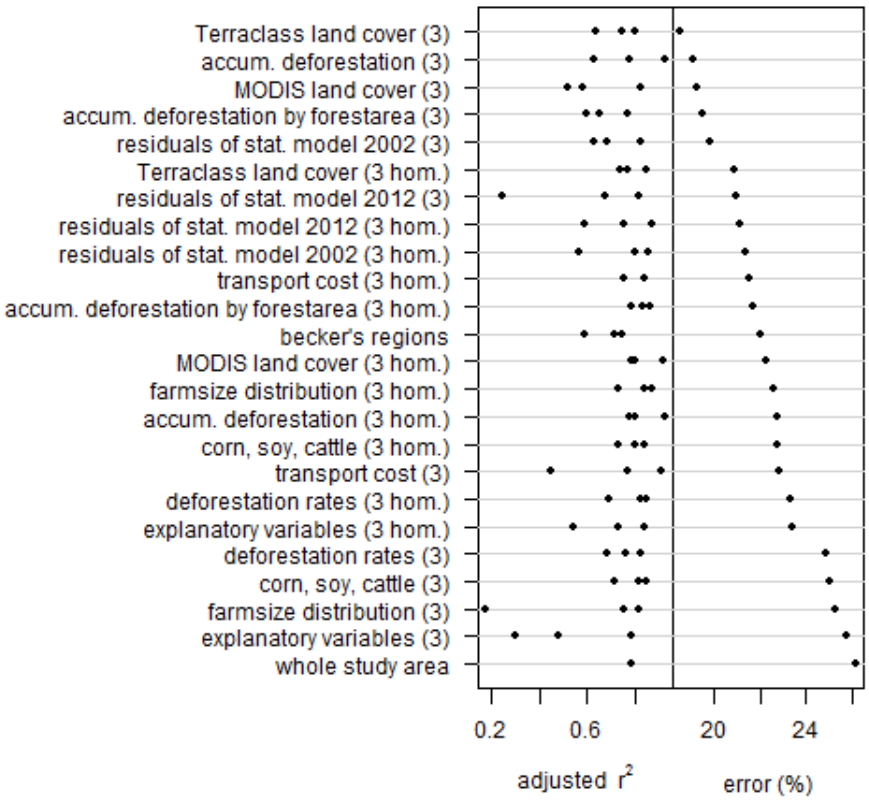

Figure 9: Comparison of the adjusted R^2 of the individual subregions' models (left) and the errors of the entire partitions (right) in 2002.

cover, (2) *Accumulated deforestation*, (3) *MODIS land cover* and (4) *Accumulated deforestation per forest area*. For 9 regions, it is (1) *TerraClass land lover*, (2) *MODIS land cover*, (3) *Residuals 2002* and (4) *Accumulated deforestation (homogeneous region sizes)*. Figures showing the errors of all the partitions of 2002 and 2012 can be found in the appendix (figs. 26,27). Tables of the errors are available on the DVD attached to this study.

In 2012, the best regionalisations for 3 regions are (1) *Accumulated deforestation*, (2) *TerraClass land cover*, (3) *Accumulated deforestation per forest area* (4) *MODIS land cover*. For 9 regions, it is (1) *Accumulated deforestation (homogeneous region sizes)*, (2) *Accumulated deforestation (heterogeneous region sizes)*, (3) *TerraClass land cover (homogeneous region sizes)* and (4) *TerraClass land cover (heterogeneous region sizes)*.

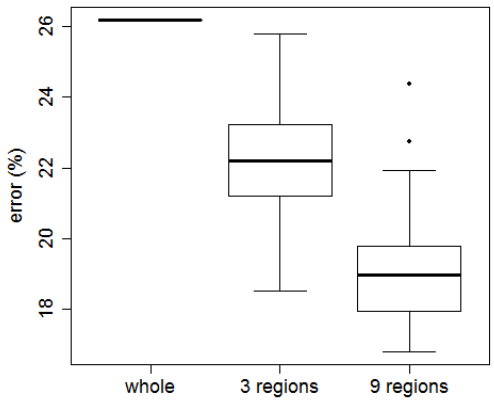

Figure 10: Boxplots of the error depending on the number of regions of the partitions in 2002. The box indicates the interquartile range and the whiskers extend to values up to 1.5 times the interquartile range. More extreme values are plotted as dots. The pattern in 2012 is essentially the same, so its boxplot is not shown.

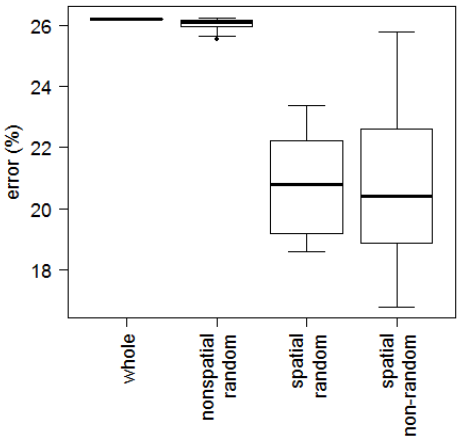

Figure 11: Boxplots of the error depending on the randomness of the partitions in 2002. The pattern in 2012 is essentially the same, so its boxplot is not shown.

3.4 Performance of the partitions under scenarios C (modelling for estimating future deforestation)

In this section, the effects of regionalisation on estimating future deforestation are presented. For this, the models fitted to the 2002 data are applied to the explanatory variables' values of 2012.

When using a model to predict for the dataset it was fitted to, the total amount of deforestation is estimated correctly, as sum of residuals is zero. When applying the model to a different data set, i.e. with different values of the explaining variables, this is not given. For the entire study area, applying the 2002 model on the 2012 data results in slightly overestimating the amount of accumulated deforestation in 2012 by a factor of approximately 1.005 (estimated: 673,084 km^2, real: 669,526 km^2, see tab. 4). The real amount in 2002 was 539.262 km^2, so the increase in accumulated deforestation in reality was 24.16 %, while the estimated increase in accumulated deforestation was 24.82 %.

The overall deforestation amount is less well captured when regionalisation

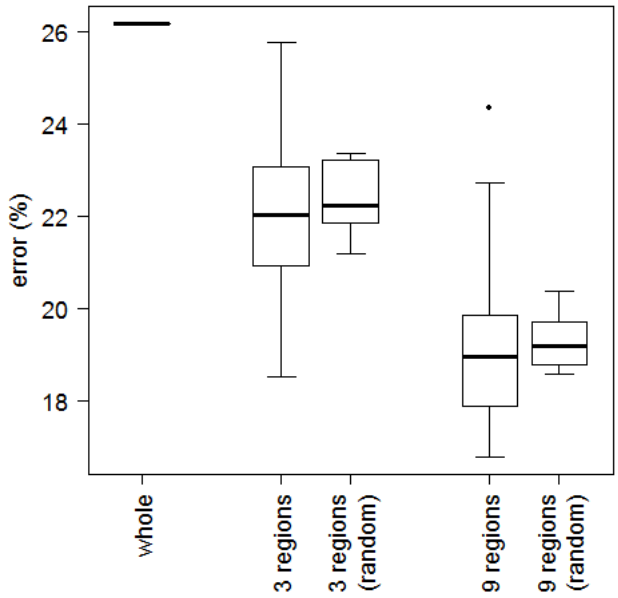

Figure 12: Boxplots of the error depending on the number of regions and the randomness of the partitions in 2002. The pattern in 2012 is essentially the same, so its boxplot is not shown.

is used. By applying the 2002 regional models to the 2012 data, the estimation of the accumulated deforestation of the whole study area is often heavily over- or underestimated: In six partitions, the predicted sum of deforestation is negative – a result that is absurd, as there cannot be less than 0 km^2 of deforested area. In another two partitions, the predicted sum of deforested area is overestimated by more than double (by factor 3 and factor 5.2). In all of these eight partitions with extreme misestimations of the amount, at least one of the regions over- or underestimates the deforestation amount by a factor of more than 20. Fig. 13 shows by how many times the overall deforestation amount is over- or underestimated in the various partitions and in the individual subregions. It shows that in many cases, a bad overall estimation is caused by one extreme region. This extreme behaviour will be discussed in section 4.3.

In the remaining 58 partitions, the predicted sum deviates on average approximately 59,500 km^2 from the real sum, with the deviations ranging from -437,601.1 to 317,236.8 km^2.

The four best amount estimations – and the only ones outperforming the whole study area – are three random partitions (twice into 3 regions, and once into 9 regions), and the partition *Full explanatory variables (3 homogeneous regions)*, the latter deviating -2520 km^2 from the real amount (see tab. 4). Fig. 30 (in the

Amounts of accumulated deforested area	
Real amount 2002	539,262 km^2
Real amount 2012	669,526 km^2
Predicted for 2012 (model fitted to entire study area)	673,084 km^2
Predicted for 2012 (best partition, *Full explanatory variables)*	667,006 km^2

Table 4: Amounts of accumulated deforested area in 2002 and 2012, estimated and real.

appendix) shows the deviations of the predicted deforestation amounts from the real values for all the partitions.

As mentioned in section 2.4, applying the error measure to predictions whose total predicted amount of deforestation differs from the real amount leads to misleading error values – partitions with underestimation are privileged, the more extreme the underestimation, the smaller the error. Thus, before computing the error, we apply a correction factor on the predictions to enforce equal predicted and real amounts. Because of this, the error only allows comparing the correctness of the spatial allocation of the models. These results are the scenario C1.

The spatial allocation is achieved by the partition *Residuals 2002* with an error of 16.3 % (9 regions). The best partition into 3 regions – and the second best overall – is *TerraClass land cover* (16.6 %). Becker's regions and the federal states perform rather badly (errors of 31.4 % and 38.1 %). They perform even worse than the undivided study area, which has an error of 22.4 %. In both the scenarios A and B, the undivided study area performs worst. Now, even with the correction factor, 28 of 66 partitions perform worse than it.

This is because the correction factor keeps the relative differences between the regions constant. The partitions that misestimate strongly the overall amount of deforestation do so because one of the regions misestimates extremely, while the other regions have a normal behaviour. Thus, by applying the correction factor, the regions that have extreme overestimations are downscaled. At the same time, the predictions in the regions whose predictions had approximately correct magnitudes also get downscaled, so they become less realistic. In the cases where the sum of predictions was negative, the whole pattern is inverted – deforestation is predicted in the most unlikely places, and the highly deforested areas get negative predictions. This effect leads to error values exceeding 100 % (up to almost 420 %), even with a correction factor. The 10 partitions with the worst errors are

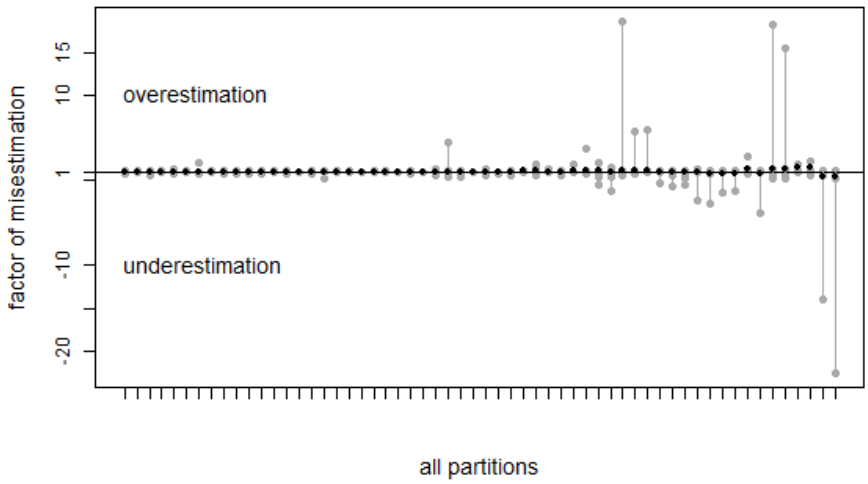

Figure 13: The misestimation of the overall deforestation sum in the different partitions (black dots, excluding the eight outliers) and in the individual subregions (grey dots and lines). While the estimation of deforestation sum in the whole partitions varies between factor 0.35 and 1.47, the estimation in the individual subregions can reach factor 20. This graph should demonstrate the extreme behaviour of some subregions compared to the entire study area . The names of the partitions are not relevant for this, but can be look up in fig. 30 in the appendix.

the 10 that have the poorest prediction of the overall deforestation amount.

To reduce the influence of the outlier regions, a correction factor by region is applied in scenario C2. Now there are no more extreme outliers, the error values range from 16.2 % to 39.1 %. Most partitions that performed very poorly before perform similarly to the other partitions now. While there is no change in the best four partitions, the ranking of the other partitions changed a lot. Still, 25 % of the partitions perform worse than the undivided study area (error of 22.4 %), including the Becker regions and the federal states (22.8 % and 31.0 %).

3.5 Ranking of the partitions in various scenarios

In each of the four scenarios above – (A) models fitted to 2012 and applied to 2012, (B) models fitted to 2002 and applied to 2002, and (C) models fitted to 2002 and applied to 2012 (with correction factor (C1) and correction factor by subregion (C2)) –, the ranking of the partitions and their performance compared to the undivided study areas are different. Tab. 5 shows on overview over the best partitions in the four scenarios.

The partition *TerraClass land cover* is the best in four out of eight comparisons –

	Scenario A (2002)	**Scenario B (2012)**	**Scenario C1** (pred. 2012, correction factor applied)	**Scenario C2** (correction factor applied by region)
Best partition (3 regions)	18.52 % (*TerraClass land cover*)	15.74 % (*Accum. deforestation*)	16.52 % (*TerraClass land cover*)	16.65 % (*TerraClass land cover*)
Best partition (9 regions)	16.78 % (*TerraClass land cover*)	14.35 % (*Accum. deforestation, hom. region size*)	16.30 % (*Residuals 2002*)	16.18 % (*Residuals 2002*)

Table 5: Best partitions in all scenarios

but it scores 14th in scenario C1 for 9 regions. *Accumulated deforestation* is best in only one of eight comparisons, but is always among the 7 best (ranks: 1, 2, 2, 4, 4, 4, 6, 7). *Accumulated deforestation (homogeneous region sizes)* is the best in scenario B (2012) with 9 regions, but scores badly overall (ranks: 1, 4, 12, 12, 12, 13, 15, 20). In scenario B (2012) for 3 regions, *Accumulated deforestation* scores best, but is only marginally better than *TerraClass land cover* (errors of 15.7 % vs. 15.9 %). So for finding out a general trend, it is better to include not only the 'winners', but the entire rankings. Fig. 14 allows to visually seize the whole distribution of all the rankings by colour coding. It shows the ranks of all partitions in the eight comparisons from table 5. The best partitions are coloured in the darker shades of grey. The table is sorted by the sum of the ranks, so that partitions that rank well in several comparisons are on top.

It is visible that the rankings in the various comparisons can differ a lot in some cases, but a general trend is visible, especially for the scenarios A and B. Among the partitions that perform well in many cases are the one that are done by land cover – *TerraClass land cover*, *MODIS land cover*, and *Accumumlated deforestation*. The partitions by variables that were used as explanatory variables in the statistical models are found on the lower ranks.

Fig. 15 shows another table of the partitions, this time ordered by how well the partition performed relative to the undivided study area. A negative value means that this partition performed worse than the undivided study area. The

3 regions A_2002	9 regions A_2002	3 regions B_2012	9 regions B_2012	3 regions C1	9 regions C1	3 regions C2	9 regions C2	
2	7	1	2	4	4	4	6	accum. deforestation
1	1	2	4	1	14	1	8	Terraclass land cover
5	3	5	10	5	1	5	1	residuals of stat. model 2002
3	2	4	7	3	13	2	7	MODIS land cover
4	5	3	5	2	12	3	11	accum. deforestation by forestarea
13	6	11	8	10	3	9	3	MODIS land cover (hom.)
7	12	13	16	6	2	6	2	residuals of stat. model 2012
10	14	7	9	7	5	7	9	transport cost (hom.)
15	4	12	1	12	20	12	13	accum. deforestation (hom.)
6	8	10	3	23	22	16	5	Terraclass land cover (hom.)
9	13	9	12	9	17	10	15	residuals of stat. model 2002 (hom.)
8	11	6	11	8	21	8	21	residuals of stat. model 2012 (hom.)
11	9	8	6	22	19	23	4	accum. deforestation by forestarea (hom.)
14	10	15	14	11	23	11	12	farmsize distribution (hom.)
18	18	17	17	14	6	14	10	deforestation rates (hom.)
19	17	18	18	13	7	13	18	explanatory variables (hom.)
16	16	16	15	15	18	15	14	corn, soy, cattle (hom.)
17	19	19	19	16	10	17	20	transport cost
23	22	23	21	17	8	18	16	explanatory variables
20	23	20	23	19	11	21	17	deforestation rates
22	21	22	22	18	9	19	22	farmsize distribution
21	20	21	20	20	15	22	19	corn, soy, cattle

Figure 14: Visualisation of the eight rankings (4 scenarios, distinguished by number of regions). The darker the colour, the better the partition performs. The table is sorted by the columns' sums.

higher the (positive) value, the better the partition performed and the darker the grey shade.

The figure shows that partitioning improved the overall goodness of fit most in 2002, using 9 regions. The next best improvement is provided by partitioning into 9 regions in 202. Negative values only occur in the scenarios C1 and C2. *TerraClass land cover*, *Accumulated deforestation by forest area* and *MODIS land cover*, which are among the best in scenarios A, B and parts of C perform worse than the whole study area in scenario C1 for 9 regions.

Because of the large range of values, scenario C1 gets it a lot of weight in the ordering. If we sort by only the scenarios A and B (see fig. 16), the colour pattern shows how different the scenarios C's ranking is from scenarios A and B. They agree in the low performance of the partitions by the explanatory variables, but disagree strongly in the performance of the land cover related partitions *TerraClass land cover (homogeneous region sizes)*, *Accumulated deforestation (homogeneous region sizes)* and *Accumulated deforestation by forest area (homogeneous region sizes)*.

The rankings of scenario C1 and C2 in 3 regions are very similar (see fig. 17). For 9 regions, they differ a lot (see fig. 18).

3 regions A_2002	9 regions A_2002	3 regions B_2012	9 regions B_2012	3 regions C1	9 regions C1	3 regions C2	9 regions C2	
7.1	8.27	5.96	7.14	5.07	4.98	5.23	4.67	accum. deforestation
6.39	8.65	4.6	6.41	4.61	6.11	4.64	6.23	residuals of stat. model 2002
5.24	7.23	3.77	5.66	3.72	5.4	3.8	5.46	residuals of stat. model 2012
3.94	8.3	3.99	6.88	2.48	5.39	2.69	5.42	MODIS land cover (hom.)
7.67	9.41	5.77	7.07	5.82	-6.35	5.76	3.82	Terraclass land cover
6.69	8.33	5.62	7.01	5.5	-3.22	5.62	2.8	accum. deforestation by forestarea
6.93	8.93	5.4	6.89	5.43	-5.52	5.66	4.39	MODIS land cover
4.67	7.05	4.39	6.73	3.12	3.84	3.43	3.44	transport cost (hom.)
2.91	6.26	3.47	5.66	1.92	2.87	1.9	2.86	deforestation rates (hom.)
2.78	6.38	3.32	5.43	2	0.51	1.96	-4.05	explanatory variables (hom.)
3.32	5.05	2.67	4.96	0.98	-2.07	0.6	-5.06	transport cost
0.41	3.44	0.35	3.69	0.15	0.5	0.23	-0.37	explanatory variables
1.35	1.83	1.69	2.19	-1	-2.26	-0.56	-1.52	deforestation rates
0.91	4.27	0.8	3.29	-0.03	0.05	-0.08	-7.83	farmsize distribution
1.17	4.92	1.09	4.02	-1.33	-7.66	-1.71	-4.99	corn, soy, cattle
4.83	7.11	4.26	6.32	2.49	-77.29	2.45	1.59	residuals of stat. model 2002 (hom.)
3.41	6.64	3.48	5.83	1.76	-107.8	1.68	2.11	corn, soy, cattle (hom.)
3.46	8.62	3.89	7.34	2.1	-167.5	2.31	2.39	accum. deforestation (hom.)
5.02	7.33	4.5	6.38	2.73	-173.14	3.05	-6.23	residuals of stat. model 2012 (hom.)
4.5	8.26	4.31	7.01	-12.5	-163.53	-8.56	5.32	accum. deforestation by forestarea (hom.)
5.29	8.26	4.17	7.08	-58.65	-329.01	1.4	5.2	Terraclass land cover (hom.)
3.59	7.49	3.5	5.95	2.34	-396.2	2.32	2.53	farmsize distribution (hom.)

Figure 15: Visualisation of the eight rankings (4 scenarios, distinguished by number of regions). The partitions are ranked by their performance relative to the undivided study area. The darker the colour, the better the partition performs. The values are the differences in error to the undivided study area. The table is sorted by the columns' sums.

3 regions A_2002	9 regions A_2002	3 regions B_2012	9 regions B_2012	3 regions C1	9 regions C1	3 regions C2	9 regions C2	
7.67	9.41	5.77	7.07	5.82	-6.35	5.76	3.82	Terraclass land cover
7.1	8.27	5.96	7.14	5.07	4.98	5.23	4.67	accum. deforestation
6.93	8.93	5.4	6.89	5.43	-5.52	5.66	4.39	MODIS land cover
6.69	8.33	5.62	7.01	5.5	-3.22	5.62	2.8	accum. deforestation by forestarea
6.39	8.65	4.6	6.41	4.61	6.11	4.64	6.23	residuals of stat. model 2002
5.29	8.26	4.17	7.08	-58.65	-329.01	1.4	5.2	Terraclass land cover (hom.)
4.5	8.26	4.31	7.01	-12.5	-163.53	-8.56	5.32	accum. deforestation by forestarea (hom.)
3.46	8.62	3.89	7.34	2.1	-167.5	2.31	2.39	accum. deforestation (hom.)
5.02	7.33	4.5	6.38	2.73	-173.14	3.05	-6.23	residuals of stat. model 2012 (hom.)
3.94	8.3	3.99	6.88	2.48	5.39	2.69	5.42	MODIS land cover (hom.)
4.67	7.05	4.39	6.73	3.12	3.84	3.43	3.44	transport cost (hom.)
4.83	7.11	4.26	6.32	2.49	-77.29	2.45	1.59	residuals of stat. model 2002 (hom.)
5.24	7.23	3.77	5.66	3.72	5.4	3.8	5.46	residuals of stat. model 2012
3.59	7.49	3.5	5.95	2.34	-396.2	2.32	2.53	farmsize distribution (hom.)
3.41	6.64	3.48	5.83	1.76	-107.8	1.68	2.11	corn, soy, cattle (hom.)
2.91	6.26	3.47	5.66	1.92	2.87	1.9	2.86	deforestation rates (hom.)
2.78	6.38	3.32	5.43	2	0.51	1.96	-4.05	explanatory variables (hom.)
3.32	5.05	2.67	4.96	0.98	-2.07	0.6	-5.06	transport cost
1.17	4.92	1.09	4.02	-1.33	-7.66	-1.71	-4.99	corn, soy, cattle
0.91	4.27	0.8	3.29	-0.03	0.05	-0.08	-7.83	farmsize distribution
0.41	3.44	0.35	3.69	0.15	0.5	0.23	-0.37	explanatory variables
1.35	1.83	1.69	2.19	-1	-2.26	-0.56	-1.52	deforestation rates

Figure 16: Visualisation of the eight rankings (4 scenarios, distinguished by number of regions), sorted by the performance in scenarios A and B. The partitions are ranked by their performance relative to the undivided study area. The darker the colour, the better the partition performs. The values are the differences in error to the undivided study area.

1	1	Terraclass land cover
2	3	accum. deforestation by forestarea
3	2	MODIS land cover
4	4	accum. deforestation
5	5	residuals of stat. model 2002
6	6	residuals of stat. model 2012
7	7	transport cost (hom.)
8	8	residuals of stat. model 2012 (hom.)
10	9	MODIS land cover (hom.)
9	10	residuals of stat. model 2002 (hom.)
11	11	farmsize distribution (hom.)
12	12	accum. deforestation (hom.)
13	13	explanatory variables (hom.)
14	14	deforestation rates (hom.)
15	15	corn, soy, cattle (hom.)
16	17	transport cost
17	18	explanatory variables
18	19	farmsize distribution
23	16	Terraclass land cover (hom.)
19	21	deforestation rates
20	22	corn, soy, cattle
22	23	accum. deforestation by forestarea (hom.)

Figure 17: Rankings of the partitions into 3 regions in scenarios C1 and C2: There is hardly any change.

1	1	residuals of stat. model 2002
2	2	residuals of stat. model 2012
3	3	MODIS land cover (hom.)
4	6	accum. deforestation
5	9	transport cost (hom.)
6	10	deforestation rates (hom.)
13	7	MODIS land cover
14	8	Terraclass land cover
12	11	accum. deforestation by forestarea
19	4	accum. deforestation by forestarea (hom.)
8	16	explanatory variables
7	18	explanatory variables (hom.)
22	5	Terraclass land cover (hom.)
11	17	deforestation rates
10	20	transport cost
9	22	farmsize distribution
18	14	corn, soy, cattle (hom.)
17	15	residuals of stat. model 2002 (hom.)
20	13	accum. deforestation (hom.)
15	19	corn, soy, cattle
23	12	farmsize distribution (hom.)
21	21	residuals of stat. model 2012 (hom.)

Figure 18: Rankings of the partitions into 9 regions in scenarios C1 and C2: There is substantial change between the two rankings, except for the best three partitions.

4 Discussion

4.1 General discussion of the effects of regionalisation

Regionalisation into subregions reduces the overall error of the models. This was expected as by subdividing the area, we can fit the models more precisely to the local behaviour.

This implies that the modelled phenomenon varies across space. If the modelled phenomenon were constant over space, the subregions would capture the same behaviour, resulting in similar models and a similar overall error – like the non-spatial random subsets do in this study. So the low error in the spatial partitions compared to (non-spatial) random subsets confirms that the relationship of the deforestation with the explanatory variables varies across space.

This spatial variation is nothing new. Subregions have been used in deforestation modelling before. However, it was not clear how much fitting local models improves the results, as measures evaluating to the individual models, such as the adjusted R^2, cannot capture the overall performance. [1] fitted models to the three *Becker* regions and obtained lower adjusted R^2 values for each of the subregions than for the undivided study area. While they could use and interpret the difference in the regression coefficients, it was not clear how much better the modelling of deforestation got by the regionalisation.

The results presented in this study allow for evaluation how much we win by subsetting space, quantitatively. The best partitions resulted in a decrease of the error of between 6 % and 9.4 % compared to using just a single model for the entire study area.[1] At the same time, the partition *Becker regions* got 4.2 % better than the undivided study area and the best partition into 3 regions got 7.7 % better.

While the regionalisation clearly outperforms the models fitted to non-spatial random subsets, the error of the spatial random partitions is not clearly worse than the meaningful spatial partitions. This is not surprising, since the random partitions are spatially connected and thus capture some of the effects of local spatial autocorrelation that exist in the data. There is not an infinite number of possible subdivisions of the area into 3 or 9 contiguous even-sized subregions, so

[1] This is the difference in error between partition and undivided study area. It is not a decrease *by* 6-9.4 %, which would mean that the error of the undivided study area is 100 %. For example, the partition *TerraClass land cover* (9 regions) in 2002 has a difference in error of 9.4 % compared to the undivided study area. This means that the amount of deforestation that was misplaced using the undivided study area, and is now correctly placed, is 9.4 % of the total deforestation sum.

it is not unlikely to capture meaningful regions by chance.

4.2 Discussion of the best partitions and comparison with reference partitions

In this section, we discuss the performance of the spatial partitions. The most important maps of partitions and of other spatial variables are included in the text. For the maps of all other variables and partitions, as well as the maps of the residuals of the models fitted to all the partitions, please refer to the DVD attached to this study.

Overall, the partitions using variables related to land cover and deforestation perform better than the partitions using the variables derived from the census. The latter capture regions that are relatively homogeneous in the combination of driving factors. The former are homogeneous in their deforestation amount. If the relationship between deforestation and explanatory variables were linear and strongly correlated, these would result in similar partitions. This is consistent with the high spatial autocorrelation of deforestation. By regionalising into regions that are homogeneous in the amount of accumulated deforestation, we approximate the effect of considering both global and local spatial autocorrelation factors in the model.

The slightly better performance of the deforestation-related partitions indicates the driving factors used in the statistical model do not completely explain the amount of deforestation. Decision-making on land use change is related to different factors, including economic, cultural and institutional one. It is difficult for census-based variables such as agricultural area or number of bovines to capture deforestation related to land speculation. In many cases in Amazonia, people decide to cut forest based on future expected revenues when the land is resold to farmers [28]. This results in a moving frontier effect in regions with high accumulated deforestation. Speculators sell their land to incoming farmers and expand the frontier, leading to more deforestation which will only be captured later in time by economic-related variables [9].

As we have mentioned before, the accumulated deforestation reflects the land cover change of several decades, while the driving factors are driving the deforestation in the current moment. As mentioned by several studies, the spatial pattern of deforestation changes over time. Thus, the explanatory variables, which reflect the recent state of driving forces, may be more related to the deforestation rates than to the accumulated deforestation. Then, areas with similar

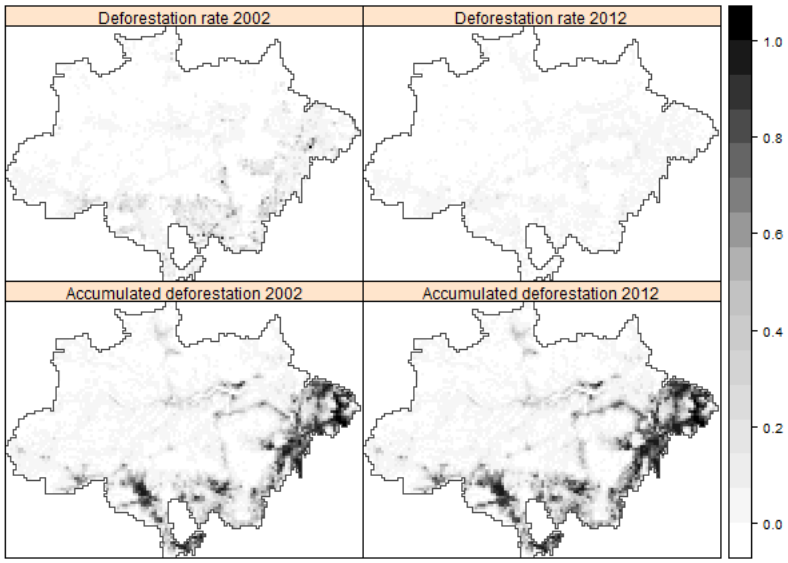

Figure 19: The accumulated deforested area and the deforestation rates in 2002 and 2012 (source INPE/PRODES).

deforestation rates but different accumulated deforestation would show different relationship between accumulated deforestation and the driving factors.

It is not surprising that the *TerraClass* partitions perform well, as they captures both – the amount of forest is related to the dependent variable, and the other land cover classes which are closely related to the agricultural driving factors. Furthermore, the *TerraClass* dataset is closely related to the *PRODES* dataset, as they share the areas mapped as forest and non forest. The *MODIS* dataset also combines information about forest and agricultural land covers, but is independent of the *PRODES* dataset, which might explain its lower performance. Other regionalisations that perform relatively well are the ones using the residuals of the models fitted to the entire study area. This makes sense, as they capture regions that have a behaviour deviating from the rest of the study area, that's why they have high residuals.

Fig. 20 shows eight partitions that are highly ranked in scenarios A and B. The partitions *Accumulated deforestation*, *TerraClass land cover*, *Accumulated deforestation by forest area*, *Accumulated deforestation* and *MODIS land cover* are the four best partitions into 3 regions in 2002 as well as in 2012 (with slightly different rankings).

They look very similar, as can be seen in fig. 20.

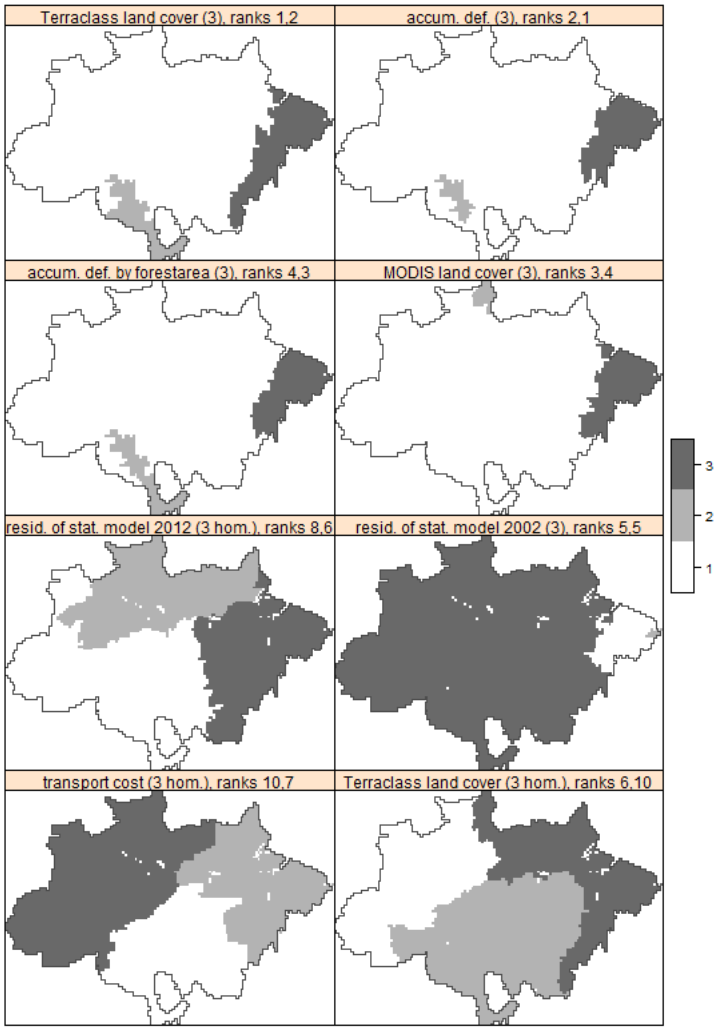

Figure 20: The eight best partitions into 3 regions for scenarios A and B.

It seems that by partitioning into these regions, we capture something that we do not capture with other regionalisations or by using the undivided study area. These regions have a more homogeneous response to deforestation drivers than other regions, such as the Becker regions used in previous studies, which is shown in fig. 8.

The four best partitions have in common that they separate the northeastern tip of the study area (western Maranhão, parts of Tocantins, eastern Pará – for a map of the federal states, please consider fig. 25) from the rest. This region is characterized by a high amount of cattle (visible in fig. 23, also confirmed by the studies cited in the introduction) and a high amount of accumulated deforestation, as it has been being deforested for many years19. Three of the partitions consider a small region in the south of the study area (parts of Rondônia,

southwestern Mato Grosso) a distinct region, which has even more cattle than the northeastern tip. In both regions, indigenous and protected areas play a very little role. Soy and corn are also present in these areas, but their region of dominance is rather in central Mato Grosso. MODIS land cover, instead of selecting the southern tip, selects the Boa Vista area in Roraima. That region has high values in the agricultural variables, but not a high accumulated deforestation, as the area is dominated by non-forest vegetation (32 in the appendix).

which also has high deforestation and much corn and cattle. Probably, a regionalisation into four areas, combining these patterns, would perform very well.

If the agricultural attributes play a large role, why are the partitions using the explanatory variables not performing so well?

4.2.1 Why does the partition by the explanatory variables not perform so well?

The partition using the full explanatory variables highlights two regions - one is the Xingú indigenous area, and the other is an area close to Colombia, which coincides with the municipality São Gabriel da Cachoeira, which is very distinct from the surrounding areas regarding the percentage of very small farm sizes, and with the indigenous area Alto Rio Negro, see fig. 21. Variables that have very abrupt changes is space, such as variables on municipality basis or clearly delimited indigenous areas, have a strong influence on the regionalisation a lot, as the algorithm easily discerns the high dissimilarity values associated with the abrupt borders. Particularly, as all variables are scaled to 0-1000 for regionalisations, this municipality is very dissimilar from all other regions. The percentage of small farms value is not particularly high, but higher than in the other municipalities, so the rescaling of the attributes to 0-1000 before regionalisation gives it a heavy weight.

4.2.2 Why does the partition by the agricultural variables not perform so well?

The partitions using only the agricultural attributes is shown in fig 22. It does capture the area in Rondônia as a distinct region, but not the northeastern area. Instead, a large area in the southeastern study area is merged as one region, including Tocantins, parts of Pará and most of Mato Grosso.

The agricultural variables represent what is happening during the years 2002-2012. We can see that the northeastern area is not highlighted very much in the agricultural variables from 2002 to 2012 23. Cattle is similarly abundant in the south as in the northeast. Corn and soy area greatly emphasize the southeast,

especially more recently. So the regionalisation created based on the agricultural variables of each year over 10 years highlights the southeast instead of the northeast.

Partitions based on land cover reflect the spatial pattern of accumulated deforestation. The recent deforestation (yearly deforestation rates) is slightly more intensive the in southeast, coinciding with the agricultral dynamics, but the northeast has a high accumulated deforestation, as it has been deforested for decades. This left the northeast with a very distinct land cover pattern (see figs. 31 and 32 in the appendix). So the partitions using land cover classes reflect past changes rather than the recent dynamics. This is further confirmed by the presence of secondary vegetation in the northeastern area, (as seen by *TerraClass* dataset, see 32 in the appendix), indicating that the deforestation is not a recent phenomenon and parts of the area are already abandoned.

This explains the different partitions created by regionalisation using land cover and using agricultural attributes and their differing performance. The statistical models used in this study model accumulated deforestation, not yearly deforestation rates. So by using partitions that single out the northeast region, we capture a region with very distinct behaviour in the modelled phenomenon, which leads to good results in the modelling. Furthermore, the agricultural variables, especially soy and corn, are less present in northeastern region than in other areas with high accumulated deforestation (e.g. Rondônia, Mato Grosso). Thus, this region has a distinct behaviour of the dependent variable in relation to the agricultural explanatory variables. This is why modelling it as a separate region improves the results.

4.2.3 The partitions into 9 regions

The partitions into 9 regions look less similar among each other, see fig. 24. But the areas identified in the partitions into 3 regions – northeastern area, Rondônia/ Western Mato Grosso cattle area, Roraima area in the north, and the southeastern area with intensive corn and soy culture – can be seen again. All of the partitions in fig. 24 highlight the northeastern area. The best partitions catch all of areas identified above.

4.2.4 Comparison with the Becker regions and the federal states

The observations above could explain the relatively poor performance of Becker's regions and the federal states.

The Becker partition includes both southeast and northeast in one large region. It captures much area that has less accumulated deforestation than the very intensive northeastern, e.g. the area west of the city of Belém and Marajó island. The Roraima area is included in western region, and not distinguished an individual region. Overall, the Becker regions comprise a more heterogeneous area than the partitions created by the regionalisation algorithm.

The federal states have the same effect. While the areas in north (Roraima) and southwest (Rondônia) are separate, the northwestern deforestation hotspot is shared by Tocatins, Maranhão and Pará. Pará also covers much area which does not show a high accumulated deforestation, so this region is quite heterogeneous regarding deforestation.

4.3 Discussion of the findings from scenario C

In this scenario, the models were used for predicting a different moment, by using models fitted in moment t_1 to predict the behaviour of a different moment t_2. In this case, the goodness of the partitions is not just expressing in which subregions the deforestation response to the regressors is most spatially homogenous, but also in which it stays most constant over the 10 years time lapse.

The comparison between the predictions using the 2002 models show that estimating predicted total amount of deforestation by statistical models is particularly prone to misestimations in small subregions, which strongly decreases the benefit of partitioning space.

Decreasing area sizes increase the chance of getting a subregion where few regressors play a big role, while others are relatively spatially homogeneous – especially if the regions are chosen to be spatially homogeneous in (some) regressors. The spatially homogeneous regressors may also have a large influence on the overall amount of deforestation, but due to their low variation, the have no big effect in the model. Then if there is a strong change in the regressor, this effects the overall amount heavily. For example, a strong increase in regressor A may not influence the overall amount much, as at the same time, there is a strong decrease of regressor B. If the regressor B plays no big role in the model fitted, as it was spatially constant in t_1, we misestimate the amount extremely.

Such an over-reliance on one regressor causes the strong outliers in predicted amounts in scenario C. The eight partitions that predict a negative amount or an amount more than double the real amount do so because (at least) one of the regions misestimates the deforestation amount by a factor of more than 105 or less

than -23. This always occurs in the region that covers the area of the state Roraima in the north of the study area, around the city Boa Vista, and the models fitted to those regions are almost always depending strongly on the planted soy area. Other areas with high residuals are the Santarém region and the Paragominas region, 300 km east of Belém. All of these regions show a high increase in soy area between 2002 and 2012.

Another factor contributes to the strong misestimation of the overall deforestation amount in that area. The models fitted to those regions are relying heavily on soy area, but for some reason, the best fit was obtained by using at the same time the soy area and its log-transformation, with very high magnitude of coefficient values and opposed signs. Such a curve can have a precise fit in a small range of soy values, but can have rather extreme behavior with increasing soy values. So in these areas, where soy increased a lot, the predicted values 'go crazy'. This explains why the demand in the Roraima region is sometimes heavily overestimated and sometimes heavily underestimated.

To prevent this kind of behaviour, the variable selection process could be adapted or the use of a variable and its log-transformation at the same time could be prohibited. But even disconsidering these extreme cases, the prediction of demand remains less precise in smaller regions compared to whole cell space. For example, in the states, the area with the highest residuals is Pará, but the model relies on various variables.

This problem occurs mainly in partition to 9 subregions. In 3 subregions, the regionwide correction factor does not change much, indicating that no single area has extreme behaviour.

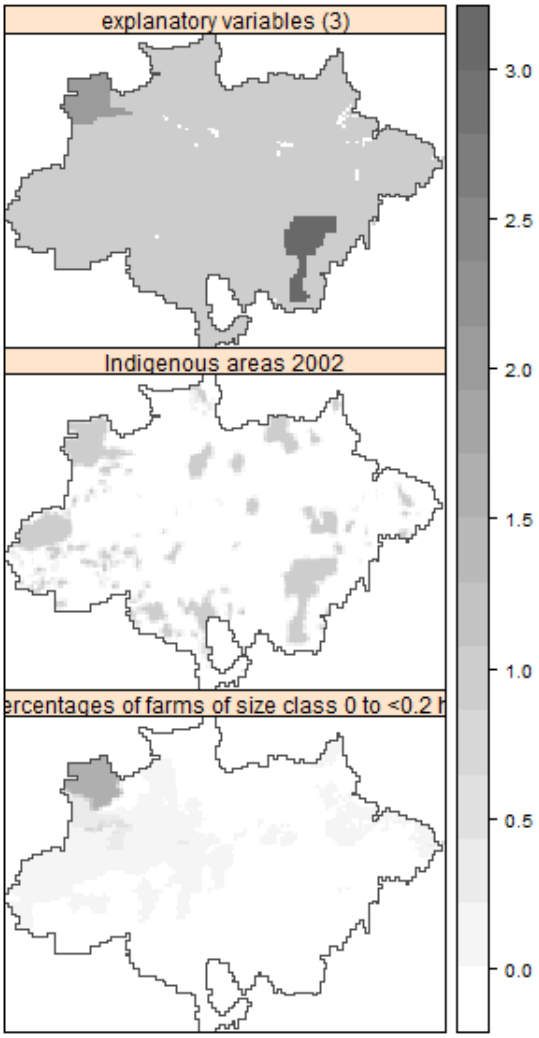

Figure 21: The partition created by all explanatory variables (on top) and the two explanatory variables that contributed mostly to this spatial pattern: Indigenous areas (middle) and the percentages of farms in size class 0 to 0.2 ha (bottom) (sources: FUNAI, IBGE).

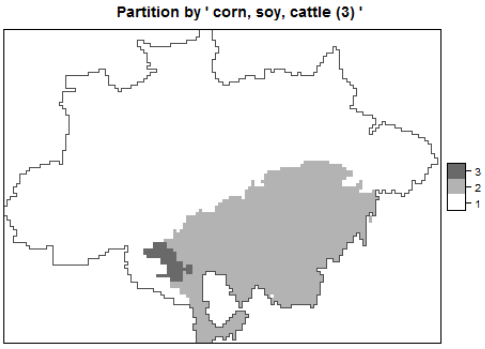

Figure 22: The partition created by using the attributes planted corn and soy area and number of cattle in the years 2002-2012.

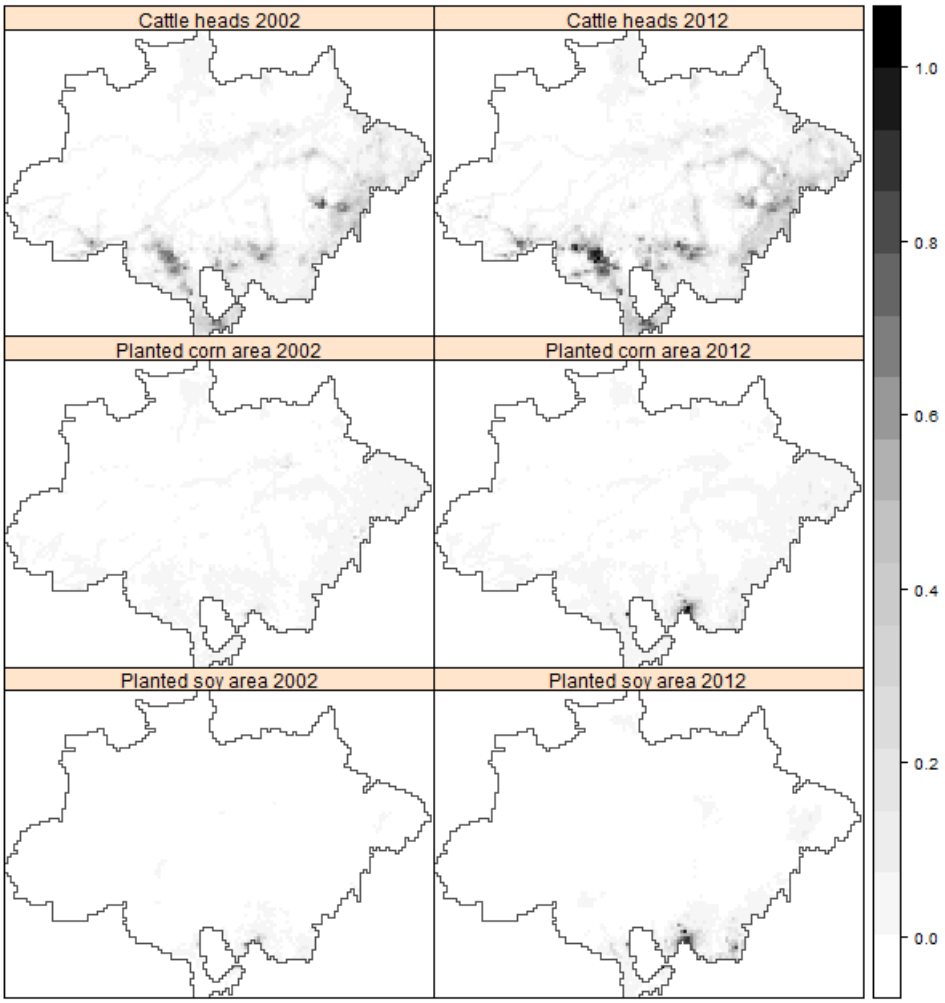

Figure 23: The variables planted corn area, planted soy area and number of cattle in 2002 and 2012, used for statistical modelling and for regionalisation (source: IGBE, adapted).

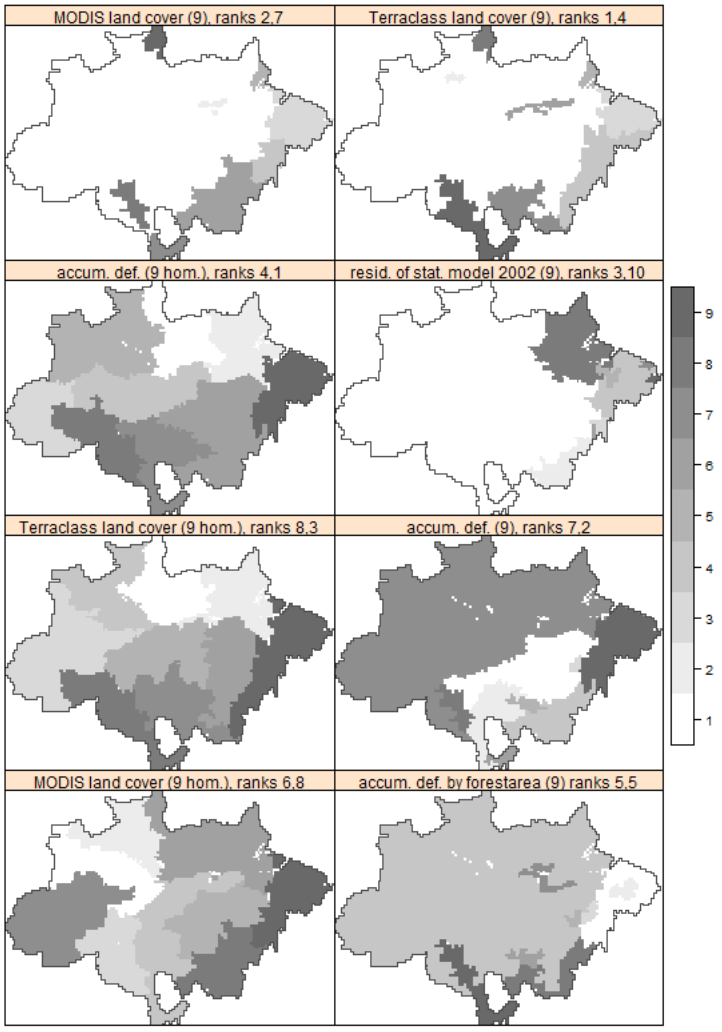

Figure 24: The eight best partitions into 9 regions for scenarios A and B. The ranks in both scenarios are noted above each map.

Figure 25: The federal states of Brazil. The study area is coloured in dark grey. The nine federal states that cover or overlap with it are: Pará (PA), Amazonas (AM), Acre (AC), Rondônia (RO), Roraima (RR), Amapá (AP) and parts of Maranhão (MA), Mato Grosso (MT) and Tocantins (TO) (surce: IBGE, adapted).

5 Conclusions and further work

This thesis studies whether regionalisation methods can improve statistical modelling of deforestation in Amazonia. To the best of our knowledge, this is the first study where different regionalisations are evaluated to find out what criteria work best.

The regionalisation procedure succeeded in identifying regions in space that show distinct behaviour regarding deforestation. Several deforestation hotspots that have been mentioned in the literature on land change in the Amazon region were identified: The cattle-raising hotspot in Rondônia, the corn- and soy-intensive central Mato Grosso area, and the northeastern region. Furthermore, it singled out a region that is not a deforestation hotspot, but shows the opposite contrast to the rest of the area – the Boa Vista area, with high agricultural dynamics but low accumulated deforestation, contrasting with the usual relationship of land cover change in the Amazon. This shows that we were successful in detecting spatial patterns relevant for land change in Amazonia. We showed that regions related to land cover attributes give the best results, and that they are better suitable for modelling accumulated deforestation than regionalisations by federal states or the regions proposed by Becker.

Overall, the subregions capture different relationships between deforestation and its driving factors, at least when these relationships are assumed to be (and modelled as) linear. It is unlikely that the relationship between deforestation and driving factors is strictly linear in reality. If the process is non-linear, the subregional models can better approximate local parts of the relationship, thus the errors reduce.

Even in the unlikely case if we had the perfect set of regressors, regionalisation can improve the results if the process is non-linear. Another reason for the better performance of subdivided space is that we may lack relevant explaining variables which vary spatially, so we capture their effect by using the subregions.

The different effect of the driving factors in space can help to find locally better adapted ways of fighting deforestation. Thus, having a clearer idea of which regions have a homogeneous response to the driving factors can be a first step towards a more local adapted understanding of deforestation.

For predicting future deforestation, using subregions, especially smaller ones, risky. Choosing the appropriate partition can be tricky, as it cannot be excluded that a partition that performs well for modelling a status quo (as in scenarios A and B) performs poorly in prediction for subsequent points in time. Thus, spatial

regionalisation when applying models for future estimation should only be used in ex-post analysis, where the reality for the target point in time (t_2) is known and the effect of the regionalisation can be evaluated. The benefit of regionalisation is then to find subregions that have similar responses to driving factors over time. Furthermore, in an ex-post analysis where the distribution of deforestation in t_2 is known, correction factors can be applied by region to correct the total amount of deforestation, so that the error describes exclusively the quality of the prediction of the spatial pattern in the subregions. Then, comparing the error reveals regions in which the spatial distribution of deforestation responds constantly to the driving factors, while the overall demand for deforestation is assumed to be independent of the spatial factors, but rather given externally.

From our results, we can point out some directions of future work in using regionalisation for land use studies. The first issue concerns choice of explanatory variables. If possible, one should use explanatory variables that capture the regional differences. In the case of Brazilian Amazonia, we could include additional variables such as land tenure situation, population, productivity factors, logging data, land prices and bank credits.

The other director for future work is the use of a different regionalisation methods. As we discussed in the Methods section, SKATER tends to create regions that are very different in size. Future studies could use alternatives such as region growing or REDCAP.

We split Amazonia into 3 and 9 regions. We could extend the study to test other region numbers. the results of this study indicate that using more regions tends to improve the goodness of fit, but is more prone to problematic effects due to overfitting.

Our regionalisation were able to confirm knowledge about deforestation hotspots in Pará, Rondônia and Mato Grosso. A more detailed characterisation of the obtained regions could be done to see what conclusions can be drawn from them. For this, an improved variable selection could be performed, reducing the risk of overfitting that occurred in some of the regions in this study.

A next step would be to find out whether these findings are valid over a larger range of applications. This could answer the questions whether those regions perform well because they really have an inherently different behaviour in reality, or whether the performance of the regionalisations depend on the model used or the variables included.

It would also be interesting to include not only deforestation, but various land cover / land use transitions. Land use/cover change can be expressed as a sparse

matrix including all possible transitions between various land uses/covers. Deforestation comprises various possible transitions. By examining them separately, the regional differences are expected to be even more emphasized and regionalisation would be more useful.

To sum up, we found out that regionalisation helps understanding the spatial differences of land change processes, especially if the area under study is large. Regionalisation is also useful to find gaps in the explanatory variables used in statistical models of land change, since when a large regions is broken into subregions, variables which may not be significant at the global scale turn out to be relevant to explain local processes. In all, regionalisation has proven to be a useful and valuable tool in land use change studies.

References

[1] Ana Paula Dutra Aguiar, Gilberto Câmara, and Maria Isabel Sobral Escada. Spatial statistical analysis of land-use determinants in the brazilian amazonia: Exploring intra-regional heterogeneity. *Ecological Modelling*, 209(2):169–188, 2007.

[2] Renato M Assunção, Marcos Corrêa Neves, Gilberto Câmara, and Corina da Costa Freitas. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20(7):797–811, 2006.

[3] Juliano Assunção, Clarissa Gandour, and Rudi Rocha. Deforestation slowdown in the legal amazon: Prices or policies? Technical report, Climate Policy Initiative / PUC-Rio, 2012.

[4] World Bank. Population of brazil. `http://data.worldbank.org/indicator/SP.POP.TOTL/countries/BR?display=graph`. Accessed: 2014-04-25.

[5] Bertha K Becker. Geopolítica da amazônia. *Estudos avançados*, 19(53):71–86, 2005.

[6] M.S. Bowman, B.S. Soares-Filho, F.D. Merry, D.C. Nepstad, H. Rodrigues, and O.T. Almeida. Persistence of cattle ranching in the brazilian amazon: A spatial analysis of the rationale for beef production. *Land Use Policy*, 29(3):558–568, 2012.

[7] Thomas M Brooks, Russell A Mittermeier, Gustavo AB da Fonseca, Justin Gerlach, Michael Hoffmann, John F Lamoreux, Cristina Goettsch Mittermeier, John D Pilgrim, and ANDA SL Rodrigues. Global biodiversity conservation priorities. *science*, 313(5783):58–61, 2006.

[8] Gilberto Camara, Lubia Vinhas, Karine Ferreira, Gilberto Queiroz, Ricardo Cartaxo Modesto Souza, Antônio Miguel Monteiro, Marcelo Tilio Carvalho, Marco Antonio Casanova, and Ubirajara Moura Freitas. *TerraLib: An open-source GIS library for large-scale environmental and socio-economic applications*, pages 247–270. Springer, Berlin, 2008.

[9] Sergio Souza Costa. *Regional scale agent-based modelling of land change: Evolving institutional arrangements in frontier areas*. PhD thesis, 2012.

[10] Instituto Brasileiro de Geografia e Estatística (IBGE). Download igbe censo agropecuário 2006 dataset. http://www.sidra.ibge.gov.br/bda/acervo/acervo2.asp?e=v&p=CA&z=t&o=11. Accessed: 2014-04-25.

[11] Instituto Brasileiro de Geografia e Estatística (IBGE). Download of the produção agrícola municipal dataset. http://www.sidra.ibge.gov.br/bda/acervo/acervo2.asp?e=v&p=PA&z=t&o=11. Accessed: 2014-04-25.

[12] Instituto Brasileiro de Geografia e Estatística (IBGE). Download of the produção pecuária municipal dataset. http://www.sidra.ibge.gov.br/bda/acervo/acervo2.asp?e=v&p=PP&z=t&o=24. Accessed: 2014-04-25.

[13] Instituto Brasileiro de Geografia e Estatística (IBGE). Mapa de biomas e de vegetação. http://www.ibge.gov.br/home/presidencia/noticias/21052004biomashtml.shtm. Accessed: 2014-04-25.

[14] Instituto Brasileiro de Geografia e Estatística (IBGE). Produção agrícola municipal. culturas temporárias e permanentes. http://www.ibge.gov.br/home/estatistica/economia/pam/2009/PAM2009_Publicacao_completa.pdf, 2009. Accessed: 2014-04-25.

[15] Instituto Brasileiro de Geografia e Estatística (IBGE). Produção pecuária municipal. ftp://ftp.ibge.gov.br/Producao_Pecuaria/Producao_da_Pecuaria_Municipal/2011/ppm2011.pdf, 2011. Accessed: 2014-04-25.

[16] Instituto Nacional de Pesquisas Espaciais (INPE). Description of prodes project. http://www.obt.inpe.br/prodes/index.php. Accessed: 2014-04-25.

[17] Instituto Nacional de Pesquisas Espaciais (INPE). Description of terraclass 2010 project. http://www.inpe.br/cra/projetos_pesquisas/terraclass2010.php. Accessed: 2014-04-25.

[18] Instituto Nacional de Pesquisas Espaciais (INPE). Download of the prodes dataset. http://www.dpi.inpe.br/prodesdigital/dadosn/mosaicos/2012/PDigital2000_2012_AMZ_gtif.zip.

[19] Instituto Nacional de Pesquisas Espaciais (INPE). Levantamento de informaçoões de uso e cobertura da terra na amazônia 2010. http://www.inpe.br/cra/projetos_pesquisas/sumario_terraclass_2010.pdf. Accessed: 2014-04-25.

[20] Fundação Nacional do Índio (FUNAI). Download of the indigenous areas dataset. `http://mapas2.funai.gov.br/portal_mapas/shapes/terra_indigena.zip`. Accessed: 2014-04-25.

[21] Ministério do Meio Ambiente (MMA). Download of the protected areas dataset. `http://mapas.mma.gov.br/ms_tmp/ucstodas.shp`. Accessed: 2014-04-25.

[22] Giovana Espindola, Ana Paula Aguiar, Edzer Pebesma, Gilberto Camara, and Leila Fonseca. Agricultural land use dynamics in the brazilian amazon based on remote sensing and census data. *Applied Geography*, 32(2):240–252, 2012.

[23] Künstler R Pigeot I Fahrmeir, L and G Tutz. *Statistik*. Springer, 2004.

[24] Helmut J Geist and Eric F Lambin. Proximate causes and underlying driving forces of tropical deforestation tropical forests are disappearing as the result of many pressures, both local and regional, acting in various combinations in different geographical locations. *BioScience*, 52(2):143–150, 2002.

[25] Diansheng Guo. Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7):801–823, 2008.

[26] Angelsen A. Kaimowitz, D. *Economic Models of Tropical Deforestation: A Review*. Center for International Forestry Research (CIFOR), Bogor, Indonesia, 1998.

[27] M.N. Macedo, R.S. DeFries, D.C. Morton, C.M. Stickler, G.L. Galford, and Y.E. Shimabukuro. Decoupling of deforestation and soy production in the southern amazon during the late 2000s. *Proceedings of the National Academy of Sciences*, 109(4):1341–1346, 2012.

[28] Sergio Margulis. Causes of deforestation in brazilian amazon (world bank working paper 22). Technical report, World Bank, 2004.

[29] MODIS. Description of the modis land cover dataset. `https://lpdaac.usgs.gov/products/modis_products_table/mcd12q1`. Accessed: 2014-04-25.

[30] KP Overmars, GHJ De Koning, and A Veldkamp. Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, 164(2):257–270, 2003.

[31] Pablo Pacheco and René Poccard-Chapuis. The complex evolution of cattle ranching development amid market integration and policy shifts in the brazilian amazon. *Annals of the Association of American Geographers*, (22 May 2012), 2012.

[32] Alex Pfaff. What drives deforestation in the brazilian amazon? evidence from satellite and socio-economic data. *Journal of Environmental Economics and Management*, 37:26–43, 1999.

[33] E. Reis and R. GuzmÃ¡n. *An econometric model of Amazon Deforestation*, pages 172–91. University College London Press, London, 1994.

[34] ASL Rodrigues, RM Ewers, L Parry, C Souza Jr, A Verissimo, and A Balmford. Boom-and-bust development patterns across the amazon deforestation frontier. *Science*, 324(5933):1435, 2009.

[35] Britaldo Silveira Soares-Filho, Daniel Curtis Nepstad, Lisa M Curran, Gustavo Coutinho Cerqueira, Ricardo Alexandrino Garcia, Claudia Azevedo Ramos, Eliane Voll, Alice McDonald, Paul Lefebvre, and Peter Schlesinger. Modelling conservation in the amazon basin. *Nature*, 440(7083):520–523, 2006.

[36] Thomas H Wonnacott and Ronald J Wonnacott. *Introductory Statistics*. Wiley, 1990.

A Graphics of the errors of all partitions in the scenarios A, B, C1 and C2, and maps of land cover variables used for regionalisation

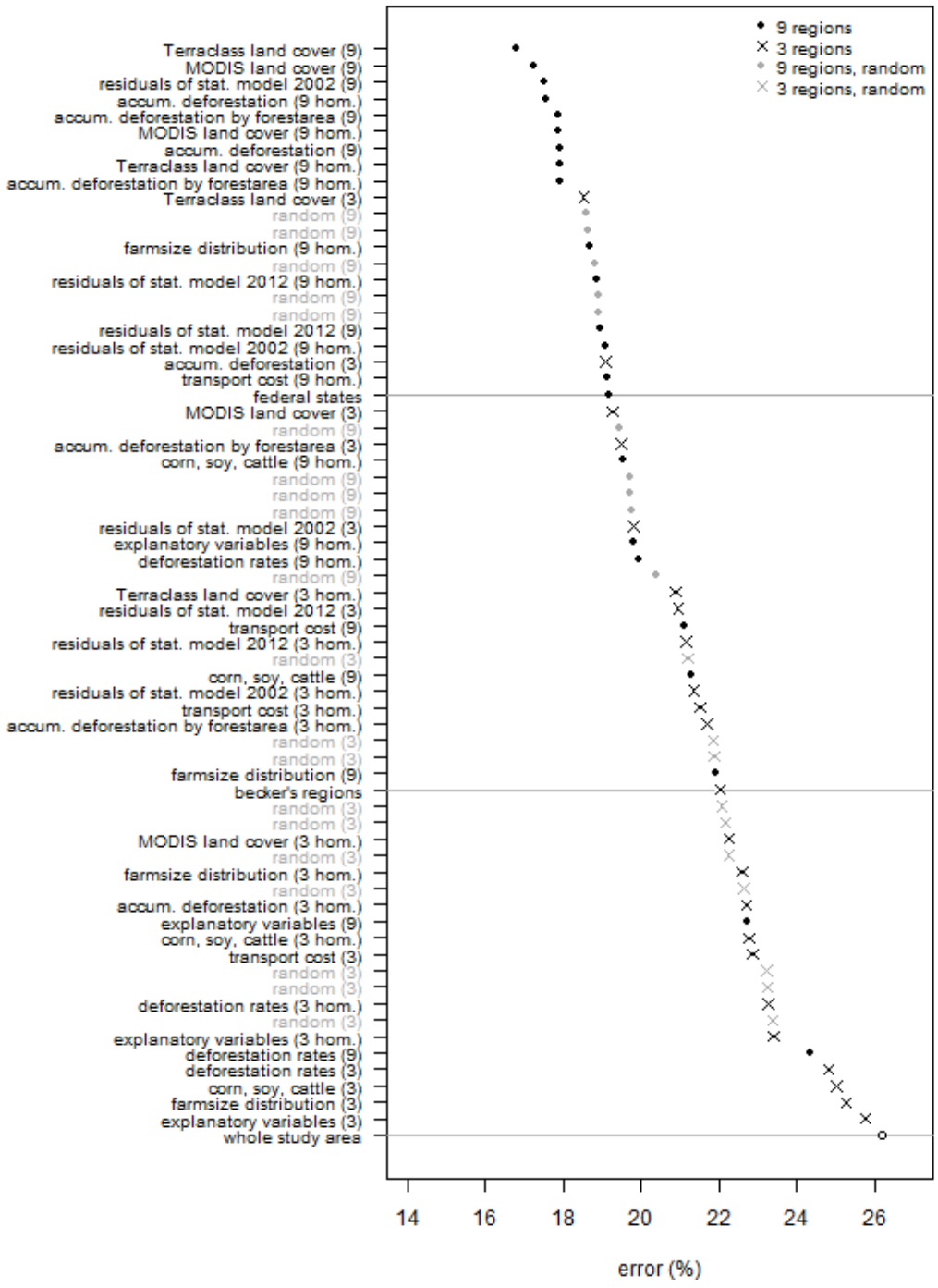

Figure 26: The errors of all partitions in 2002 (scenario A). The partitions on top are performing best. The horizontal grey lines show the performance of the federal states, the Becker regions and the undivided study area.

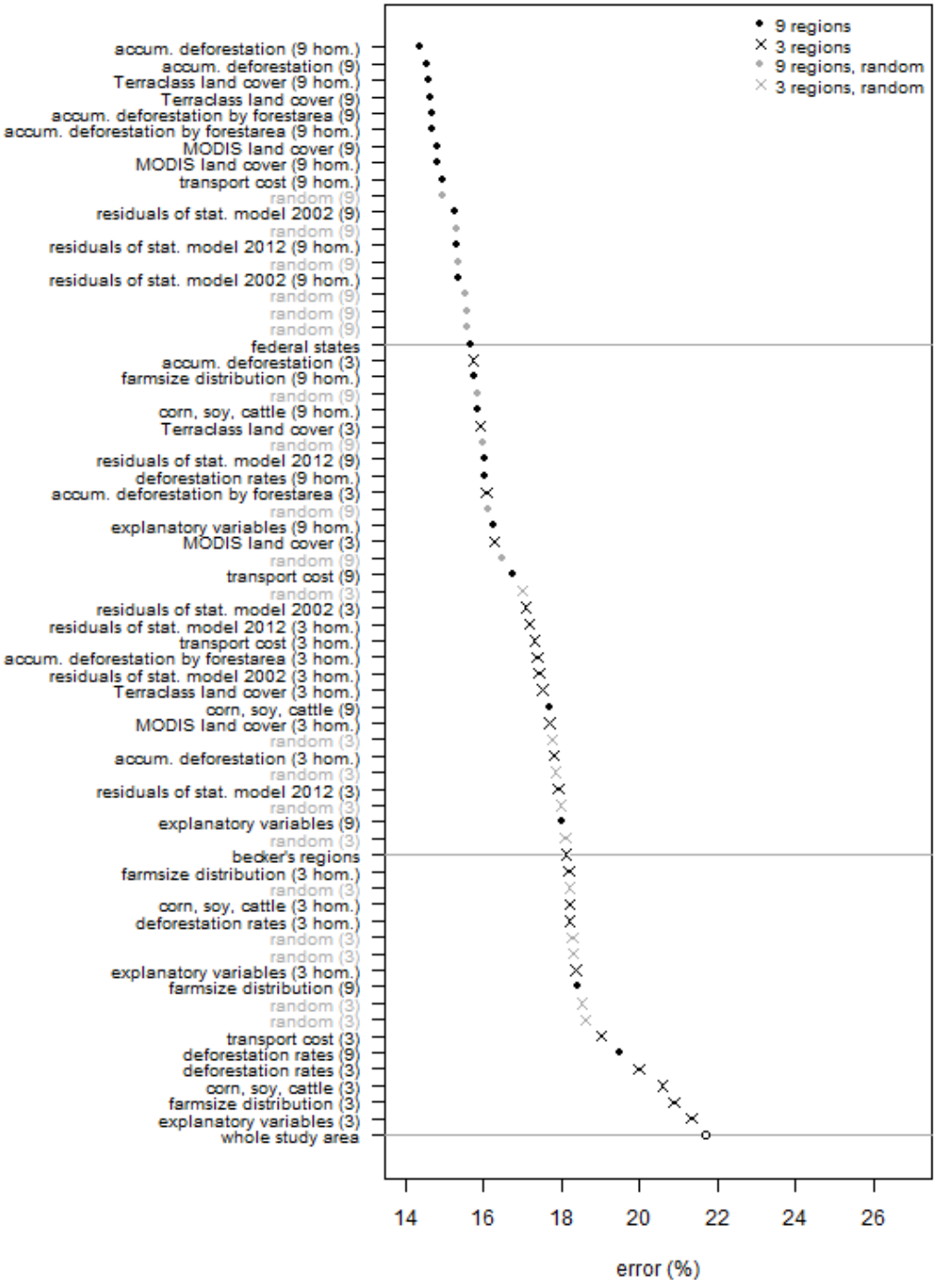

Figure 27: The errors of all partitions in 2012 (scenario B). The partitions on top are performing best. The horizontal grey lines show the performance of the federal states, the Becker regions and the undivided study area.

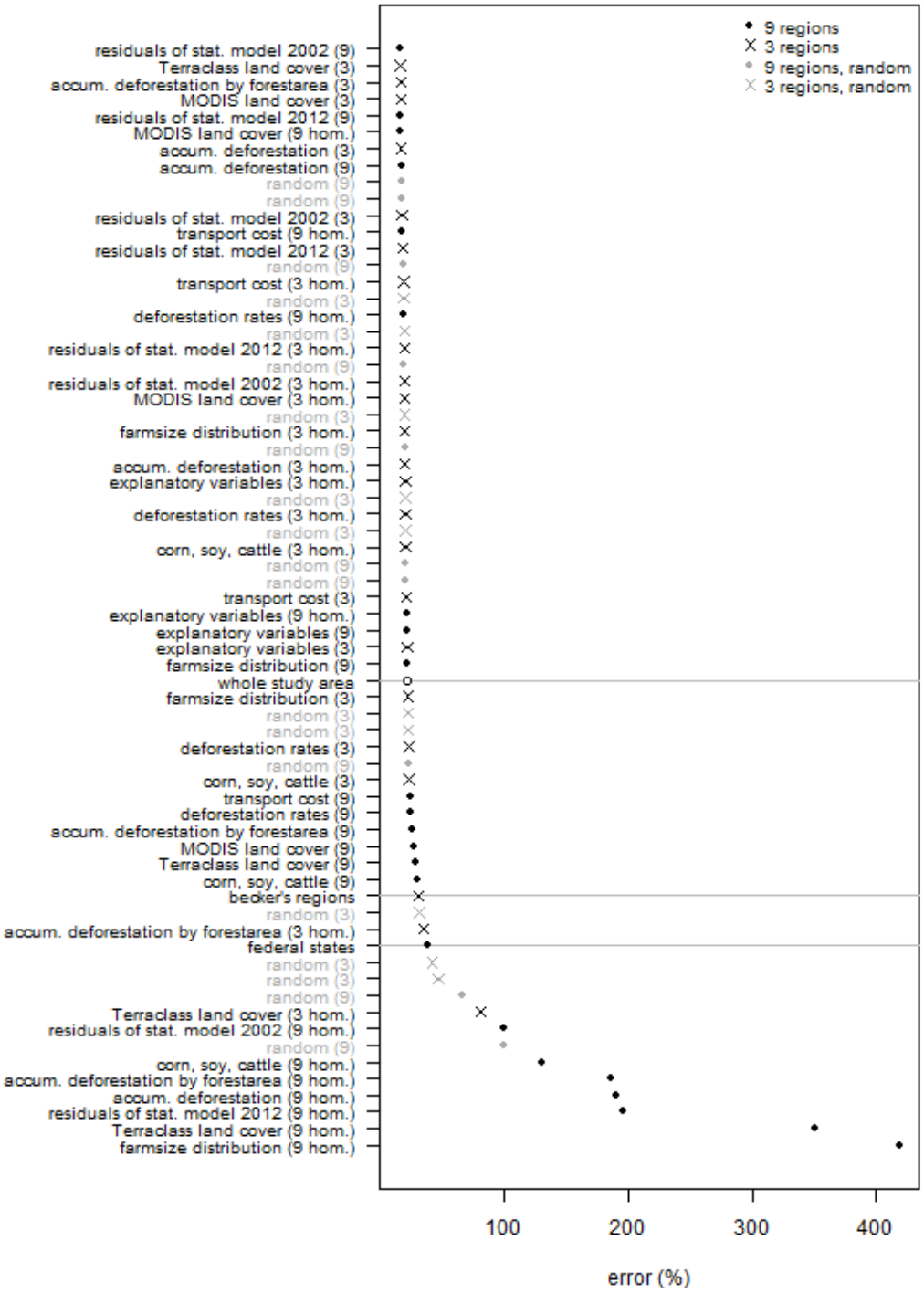

Figure 28: The errors of all partitions for predicting 2012 based on models fitted to 2002 (scenario C1). For each partition, a study area-wide correction factor is applied, making sure that the overall amount of deforestation is correct. Thus, the error measures only the quality of the spatial allocation of the models. It provides no evidence about the models' capacity of predicting deforestation amounts.

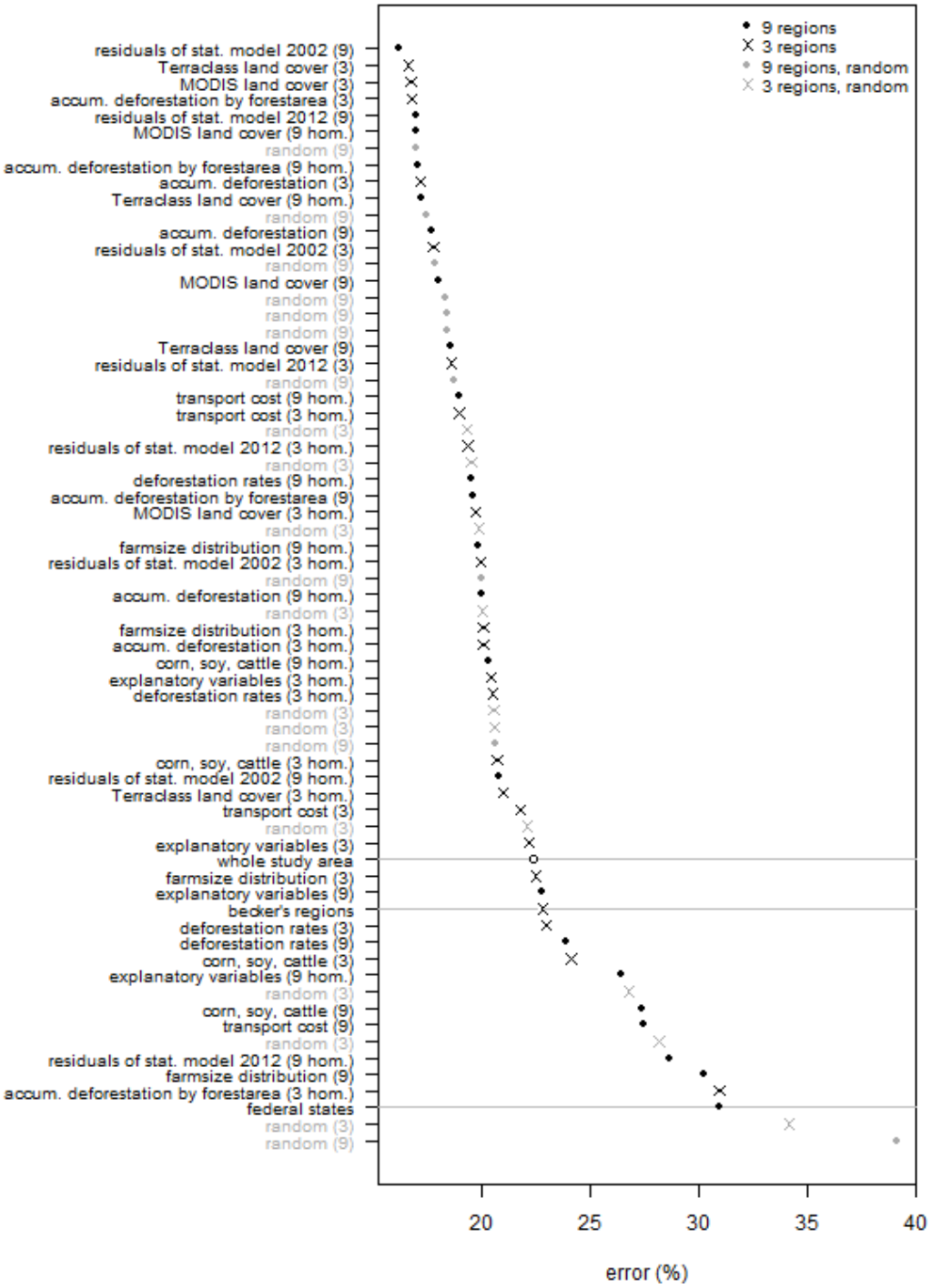

Figure 29: The errors of all partitions for predicting 2012 based on models fitted to 2002 (scenario C2). A correction factor is applied by region, making sure that the amount of deforestation is correct in each region. Thus, the error measures only the quality of the spatial allocation of the models. It provides no evidence about the models' capacity of predicting deforestation amounts.

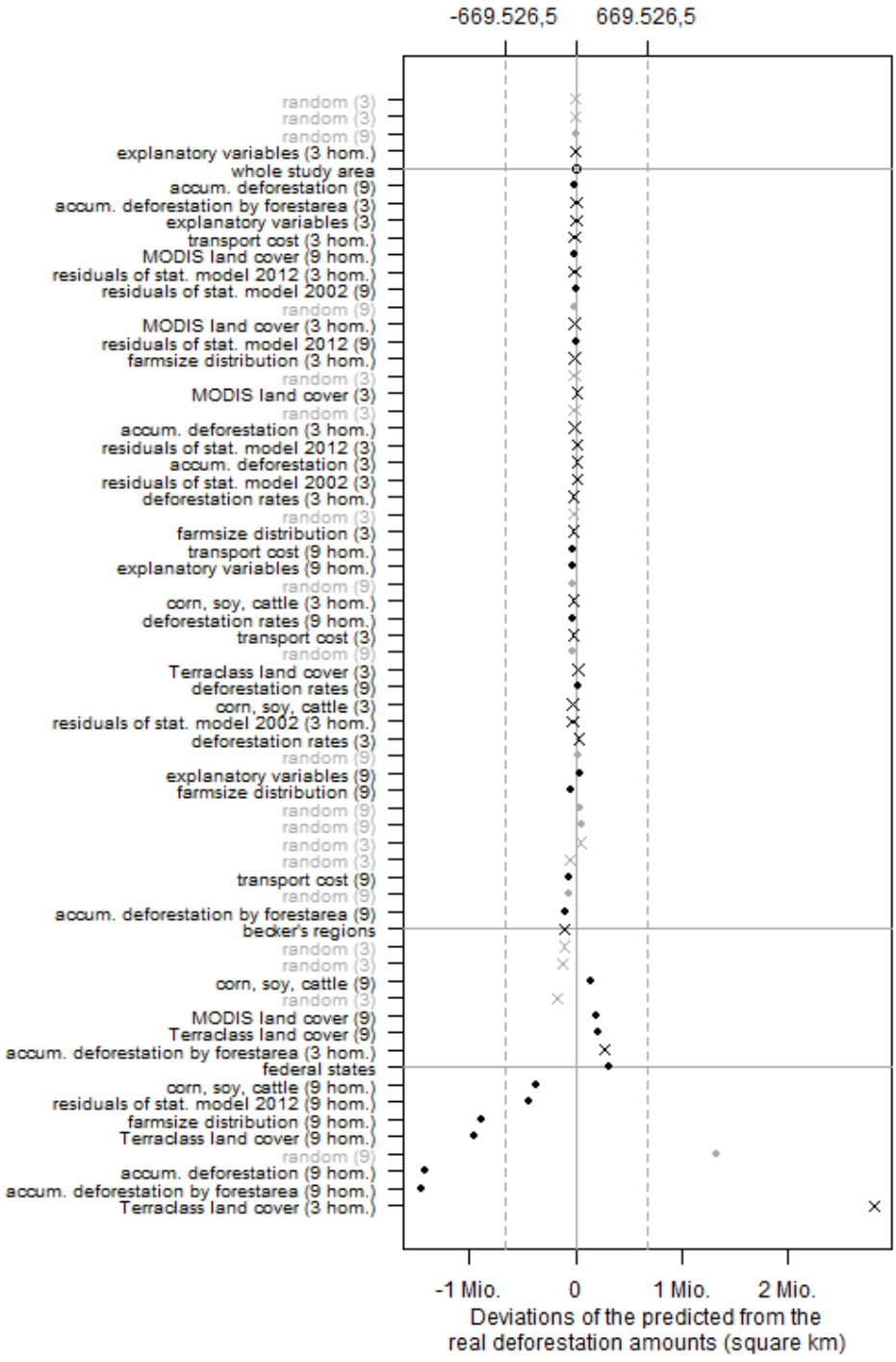

Figure 30: The deviations in km^2 of the predicted deforestation amounts for 2012 for all partitions, obtained by subtracting the real amounts from the predicted amounts. The centered vertical line means no deviation. The dashed grey lines represent deviations of $\pm$ the real amount. We see four of the partitions that have negative total amount (left of the left line), and two that overestimate more than by double. Two extreme negative outliers are omitted to keep the graphic readable.

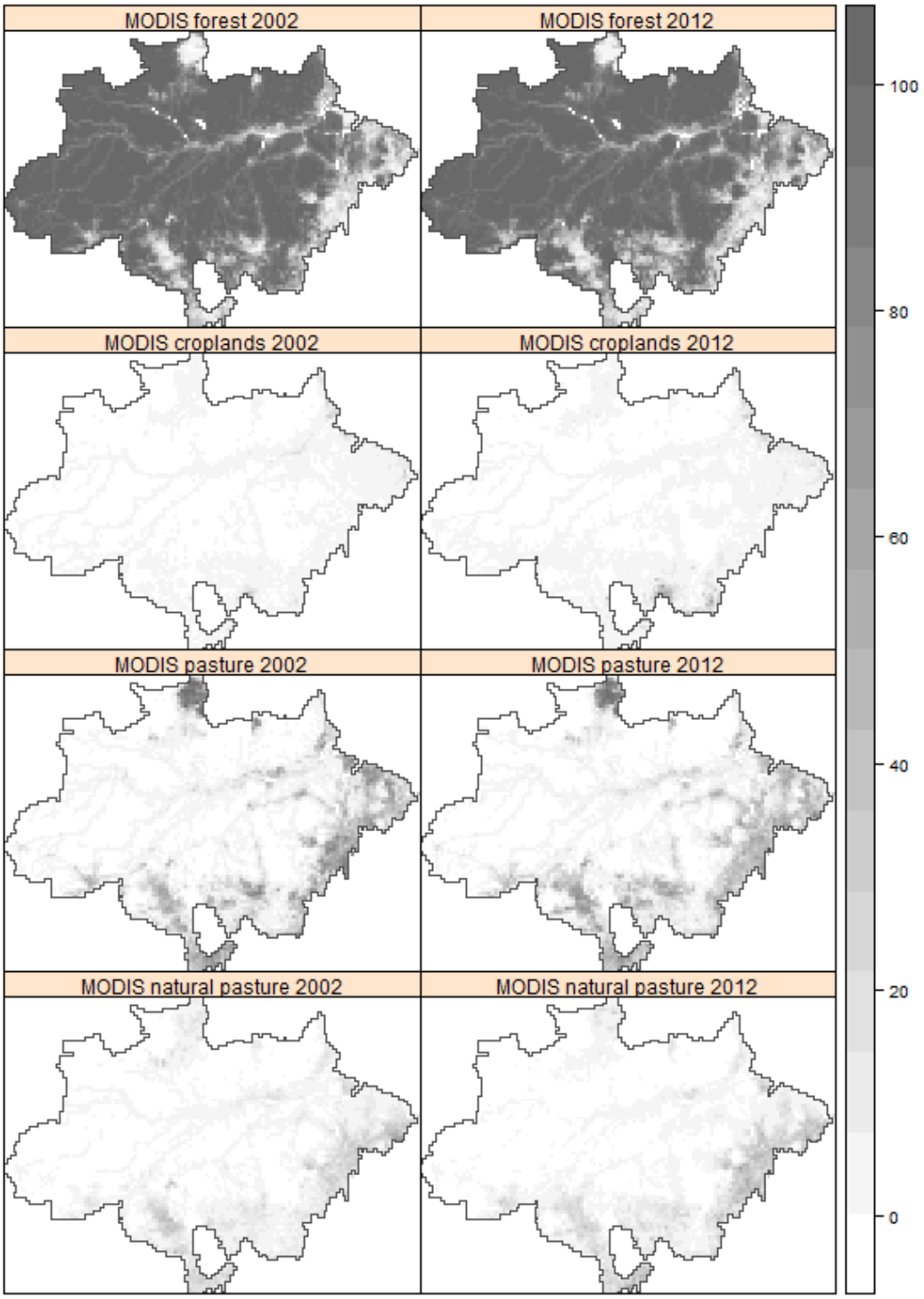

Figure 31: Examples for the *MODIS* land cover data used for partitioning the study area. This figure shows the four classes for 2002 and 2012. For regionalisation, all the data from the years in between these two is also used. The patterns are very similar, but an increase in croplands and pasture in the 10 years is visible (source: NASA/MODIS, adapted).

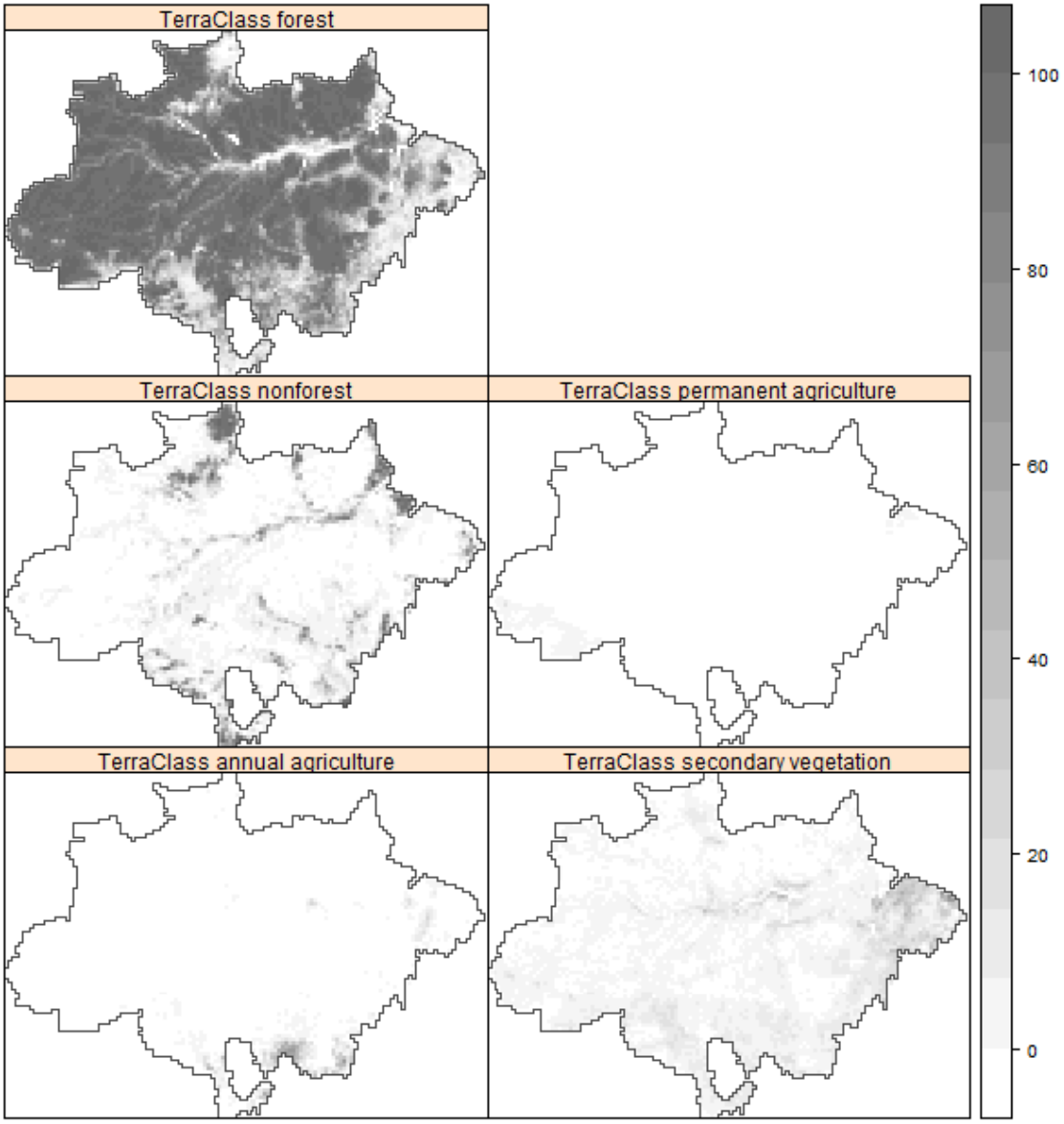

Figure 32: The *TerraClass* land cover data used for partitioning the study area (source: INPE, adapted).

I hereby assert that this master thesis with the title "*Regionalisation of the Brazilian Amazon basin for improved land change modelling*" is written by myself and that I did not use any other than the declared resources. All parts which are literally and logically taken from external sources within this work are marked as being external.

Münster, NRW, Germany 30. April 2014

Merret Buurman