Forecasting patterns of urban expansion

A statistical analysis of forces determining locational urban growth and their regional differences

Master Thesis

MSc Spatial, Transport & Environmental Economics

Vrije Universiteit Amsterdam

Pendula Ferdinand

Student number: 2577664

Supervisor: dr. Eric Koomen

August 2020

ABSTRACT

This master thesis analyzed the driving forces determining locational urban growth and their regional differences. Analysis was performed with an automated stepwise regression. The main purpose was to find a parsimonious model that describes the relationship between the urban pattern in 2010 and several driving forces. The results showed that the number and types of drivers influencing urban growth differ per continent. Furthermore, the same drivers can have both a positive and a negative influence on urban growth in different continents. The secondary research aim was to further investigate possible methodological approaches to improve the predictive power of the models. Therefore, a penalized regression was performed that tries to create a parsimonious model in exchange for an acceptable amount of bias. The stepwise and penalized logit regression models were compared to a random forest model, to see if the complex relationship between driving forces and urban growth would be better described with a non-linear model. The results of this thesis suggest that a regional differentiation of driving forces is urgently needed for global urban growth models. Furthermore, additional economical and political drivers are needed to understand recent urban development and predict future urban growth. Finally, the application of non-linear models has shown great potential and should be investigated further.

ACKNOWLEDGEMENTS

This research took place as part of a larger project led by PBL – Netherlands Environmental Assessment Agency, whose goal it is to further develop and improve the "2UP-model", designed to model future urban growth. More information on the model itself can be found in the background report "Towards an Urban Preview: Modelling future urban growth with 2UP" (Huijstee *et al*., 2018). I am very thankful to everyone involved in this project at PBL for their initial feedback on my ideas and their help along the way. I am especially grateful to Bas van Bemmel for being a great help in preparing and obtaining the large amount of data needed for this research project.

I am indebted to Bo Andrée, who helped me get started with the analyzes in R. Bo Andrée is also the author of the R package 'autoGLM', which was utilized to handle the enormous amount of data for this research project. No analysis would have been possible without the computing power of Jip Classens's computer, which he generously shared with me during this project. Moreover, Jip Classens was a huge help when it came to GeoDMS related problems.

Finally, I want to thank Eric Koomen for being an excellent supervisor and for allowing me the great opportunity to work on this project as a student research assistant.

ACRONYMS

AIC	Akaike's Information Criteria
AUC	Area under the precision-recall curve
caret	Classification And REgression Training
GLM	Generalized linear model
HWSD	Harmonized World Soil Database
JRC	Joint Research Centre
LASSO	Least Absolute Shrinkage and Selection Operator
PBL	Netherlands Environmental Assessment Agency
RF	Random Forest
ROC	Receiver operating characterisitcs
TRI	terrain roughness index
WGI	Worldwide Governance Indicators

TABLE OF CONTENTS

1 INTRODUCTION

The share of the world's population living in urban areas – including cities, towns, and suburbs - has increased from 30 to 55 percent between 1950 and 2018, due to ongoing urbanization (United Nations, 2018). As a result, urban areas have been rapidly expanding in space and will continue to do, as the population is expected to grow by another 700 million people in the next ten years (United Nations, 2019). Over the last decades, the spatial extent of urban areas has grown even faster than their respective populations, indicating an increased demand for urban land per capita (Angel et al., 2011; Seto et al., 2011). While urbanization has driven economic growth and human development (United Nations, 2018), it has also had a severe impact on the global environment. A few of these often irreversible environmental impacts are loss in biodiversity, air pollution, and even climate alterations, such as the urban heat island effect (Grimm et al., 2008; K.C. Seto, Güneralp and Hutyra, 2012). Additionally, urban expansion has increased the exposure of humans to natural hazards, as most urban development occurs in areas of high risk, such as river deltas (Seto, 2011; Neumann et al., 2015).

Due to the double-edged nature of urbanization, much research has been devoted to understanding and modeling the driving forces of urban expansion (Li, Sun and Fang, 2018). A thorough understanding of the processes at play is fundamental to promoting the sustainable development of cities and designing effective policies regarding environmental protection and natural hazards. So far, most studies on urban expansion have focused solely on individual cities or smaller regions with high data availability, often in developed countries (Poelmans and Van Rompaey, 2010; Li, Zhou and Ouyang, 2013). Even though the driving forces identified in those individual studies seem to vary a lot, as of yet, no studies exist that explicitly compare regional differences in driving forces on a global scale. A small number of global studies on driving forces of urban expansion exist. However, they do not account for regional differences and often only use a limited number of drivers, due to computational difficulties and a lack of global high-resolution data (Seto *et al.*, 2011; Güneralp, Güneralp and Liu, 2015). A global assessment of urban growth is not only necessary for data sparse regions, but also for the evaluation of international policies such as the Paris agreement. Regionally specific estimates of driving forces can make a vast contribution to improved global urban expansion models (Huijstee *et al.*, 2018).

To fill this research gap, this study's main objective is to analyze the effect that different drivers have on the spatial growth of urban areas and the difference in relevance and magnitude of those drivers between regions around the globe. Driving forces of urban expansion will be analyzed on a high spatial resolution with grid cells of 1x1 km. Such a fine resolution is in line with recent global land-use models and necessary to model larger as well as smaller urban clusters (Li *et al.*, 2017). There is also a need for urban growth forecasts on a fine resolution for local hazard impact assessments (Nussbaumer *et al.*, 2014; Promper *et al.*, 2014). The second aim of this study is to produce results that can be implemented in global modes that predict futurfoe urban growth, e.g., Huijstee *et al.* (2018). Therefore, in the second part of the analysis, emphasis is placed on the methodological approach that can make the best land-use predictions on new data. Different logistic regression models with an autoregressive specification, as well as a non-linear model will be compared based on goodness of fit and predictive power. This second aim includes differentiating between relevant drivers and model performance for urban growth within a shorter time-period, and the presence of urban sites as a result of millennia of urban development. If the end goal is to model future urban growth, identifying the essential drivers of recent urban growth could be more relevant than identifying those of historical growth.

The master thesis is composed of six chapters, including this introduction. Chapter two focusses on the theory and empirical data relevant to this study. Theoretical concepts to be discussed are urban agglomeration economies, drivers of urban growth, and background on statistical methods for land-use change analyzes. The empirical background will focus mostly on studies of different continents and their respective drivers of urban growth. The third chapter goes into detail on the methodological approach, including a description of the data applied. The methodology section is followed by a presentation of the research findings, primarily focusing on the regression results for the different continents. In the fifth chapter, the results of the analyzes will be discussed and compared to findings from the literature reviewed in chapter two. The focus of the results will first be on the regional differences between driving forces and, in the second part, shift to the varying performances of different regression models. Finally, chapter six provides a summary of the main results as well as suggestions for further research.

2 THEORETICAL AND EMPIRICAL BACKGROUND

From a theoretical viewpoint, most economic concepts relevant to this study belong to the broad field of geographical economics. The central tenet of this discipline is that the distribution of cities across space, or in a broader sense, the concentration of economic activity, is not random. Relevant concepts from the field will be discussed in section 2.1. Not all driving forces of urban expansion and the magnitude and direction of their effect can be identified solely based on theory. Therefore, empirical studies on land-use change, especially studies focusing on driving forces behind urban expansion, will be discussed in section 0 of this chapter.

Before examining the empirical evidence, the concept of driving forces in relation to urban growth will be further defined in section 2.2. For this study, an important distinction is made between forces that merely drive the growth of cities and those that determine the suitability of a spatial unit to become urbanized. In this study, more emphasis will be put on identifying the latter. Section 2.3 includes a theoretical background on statistical methods used for urban growth studies, including a description of binary logistic regression models and penalized regressions.

2.1 Urban agglomeration economies

In economics, the clustering of firms and people in cities, and the associated increased productivity of firms within these cities, is known as urban agglomeration economies or increasing urban returns. Agglomeration is a well-known empirical phenomenon that has been studied for several decades by geographers and regional scientists (Glaeser *et al.*, 1992; Ciccone and Hall, 1996; Melo, Graham and Noland, 2009).

In general, there are two causes of agglomeration; 'first nature' agglomeration causes, which are predetermined advantages due to physical geography, often referred to as 'regional endowments'. Examples of regional endowments are productive soils on relatively flat terrain, which form a good combination for profitable agricultural production. Another example is proximity to navigable waterways, which guarantees efficient transportation and associated trade connections. Apart from these first nature causes, there are second nature causes of agglomeration, which refer to the circular causality between firms and consumers. Producers choose their location based on the existing market and hence choose a location where other producers and consumers are already settled (Brakman *et al.*, 2005). After only a few firms have decided to settle at a specific geographic location, the circular causation between firms and consumers causes the growth of most cities to be self-sustaining (P. R. Krugman, 1991). First nature causes may have influenced the initial location of cities as we know them today, but second nature causes played a considerable role in their persistence and growth. Although both types of agglomeration should be taken into account when selecting drivers of urban expansion, second nature causes might be more predictive of future growth than first nature causes.

The fundamental microeconomic principles of urban agglomeration, which are used by many well-known economists (e.g., Krugman (1991), Porter (1990)) to explain the spatial clustering of economic activity, originate from the work of Alfred Marshall (1842-1924). Marshall defined three leading causes of urban agglomeration and associated increasing returns to scale, namely knowledge spillovers, labor market pooling, and the possibility of sharing and connecting local inputs. Since Marshall published his work in 1920, many empirical studies have proven the existence of these urban agglomeration benefits and their effect on city growth.

Apart from the three sources of agglomeration benefits defined by Marshall (1920), more recent studies have used additional channels to explain the spatial distribution and size of cities. These additional sources of urban agglomeration are natural advantages (first nature agglomeration causes), rent-seeking, the home market effect, and so-called consumption opportunities (Rosenthal and Strange, 2004). Even with recent advances in the theory of urban agglomeration forces and years of empirical research, Rosenthal and Strange (2004) have concluded that the empirical relevance of the individual drivers of urban agglomeration is not always clear, which emphasizes the relevance of this study. Being aware of agglomeration benefits is extremely important in this study's context, to identify relevant drivers and deal with specific methodological issues such as the endogeneity of selected rivers, which will be further discussed in section 2.3.4.

2.2 Driving forces: Quantity versus location

For this study, it is crucial to specify the kinds of driving forces of urban growth we are interested in. A general distinction can be made between driving forces of the quantitative change of urban areas on the one hand, and driving forces of the locational or physical change of urban areas on the other (Veldkamp and Lambin, 2001). With regards to the prospective quantity of urban land needed, population and GDP growth are assumed to be the main determinants (Batisani and Yarnal, 2009; Seto et al., 2011; Seto, Güneralp and Hutyra, 2012), together with an increased need for land-based commodities (Veldkamp and Lambin, 2001). An increasing number of people born in urban areas and increasing wealth lead to an increased demand for urban land. Although both drivers roughly indicate how much new urban land is needed, no spatially-explicit projections can be made.

This study aims to identify drivers that affect the physical land-use change from non-urban to urban land, which is why this study will only focus on driving forces that influence the spatial location of urban growth. These drivers of locational change help us identify locations that are most likely to be urbanized due to the increased quantity of urban land needed. The results of this study can be used in combination with models on the quantitative change of urban areas to model future urban land-use change. Throughout this thesis, the following terminologies are used interchangeably to describe driving forces of locational change: 'driving forces', 'determining factors', 'drivers'.

Driving forces of physical urban growth have been a central theme within the research field of land-use change modeling over the last few decades (Veldkamp and Lambin, 2001). By analyzing state-of-the-art research on driving forces, Bürgi, Hersperger and Schneeberger (2005) have been able to identify five main groups of driving forces: cultural, natural, political, economic and technical driving forces. Figure 1 shows the five groups of drivers and their relationships in a framework developed by Hersperger and Bürgi (2007). This classification of drivers broadly corresponds with other classifications made in similar studies discussed in section 2.4, e.g., Verburg *et al.* (2004), Li, Sun and Fang (2018) or Aguayo, Wiegand, Azócar, Wiegand, & Vega (2007).

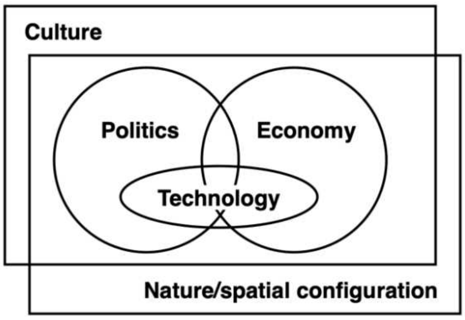


Figure 1 | Conceptual framework for the five groups of driving forces by Hersperger and Bürgi (2007)

2.2.1 Five groups of driving forces

The first group of drivers contains natural drivers and drivers from the current spatial configuration. Natural drivers generally correspond to the first nature agglomeration benefits described in section 2.1. Although natural drivers have a less direct effect on urban growth, they have a significant role to play when it comes to allocation (e.g., soil type or distance to navigable rivers) (Verburg, Schot, et al., 2004). Because of this property, natural/ spatial configuration drivers are depicted in Figure 1 as a physical background on which the other drivers interact. Following the definition of Hersperger and Bürgi (2007), the societal background and its cultural driving forces are similar to natural drivers, a starting position from which the other three groups emerge. The societal background represents the history of a city or country, but also the cultural and demographic composition of a society. The social and cultural background influences, along with personal preferences, the locational choices individuals make (Verburg, Ritsema van Eck, *et al.*, 2004). Inferring from the conceptual framework shown in Figure 1, it becomes clear that even if not explicitly included, the cultural background is always present through its influence on other drivers.

Economic drivers entail market structures, but also consumer demands and incentives from governments, such as subsidies. Models that base their assumptions purely on economic theory often include land-rents as drivers of urban expansion. The underlying assumption, which is based on the insights of Von Thünen (1966) and Ricardo (1817), is that in an equilibrium state, land will always be used in a way that maximizes profits. In economic terms, this is often called the 'bid-rent' approach. Alonso (1964) was one of the first to explain the complex relationship between urban land-use and land rents. Overall, it seems very difficult to include economic drivers in urban expansion models, as we will see during the discussion of empirical studies in section 2.4.

Political drivers include all kinds of policies that influence landscape division, such as infrastructure policies or policies on nature protection (Hersperger and Bürgi, 2007). Recent literature emphasized the vital role of spatial policies for land-use change and urban development (Hersperger *et al.*, 2018). Nevertheless, political drivers are often excluded from land-use change models, especially those that cover a large extent (e.g., global assessments). The main reason for this shortcoming is a lack of large-scale, spatially explicit datasets on spatial policies (Veldkamp and Lambin, 2001). One of the few exceptions is the global study by Seto *et al.* (2011), in which the authors included national policies on car use as a representation of the degree of spatial planning.

The framework in Figure 1 shows that economic and political drivers are interrelated. This means that economic development often affects policymaking. The interplay of economic and political drivers forms the basis for technological drivers. Technological drivers include not only technologies themselves but also the general knowledge on how to implement hese tools efficiently (Hersperger and Bürgi, 2007).

2.3 Statistical methods for analyzing urban-growth

Methods used in land-use change studies range from purely statistical approaches to cellular automata techniques, multicriterial analysis to purely theory-induced approaches. Theory-induced approaches are often used if the focus of the study lies on the causal relationship between urban growth and underlying driving forces. Statistical methods outperform theory-induced methods in modeling existing spatial patterns (Overmars, Verburg and Veldkamp, 2007), and are, therefore, more suited for this study's aim. Most studies that use a statistical method to analyze land-use change patterns use logistic regression (logit) models. Studies that analyze different land-use classes simultaneously use multinomial logit models as a standard. If the goal is to classify only one specific land-use class, such as urban sites, binary logit models are commonly used.

Like standard linear regression, the goal of the logistic regression is to predict an outcome variable[1] based on one or more independent variables[2]. The main difference in logistic regression is that the outcome variable is categorical as opposed to continuous. In the case of a binary outcome variable (for example, 0 or 1), the logistic regression predicts whether the outcome is 0 or 1 based on log odds ratios. Another vital difference between logistic and linear regression is that the optimal model is not fitted with the least squares method, but with the maximum likelihood method. With the maximum likelihood method, the likelihood that the fitted model describes the observed data is maximized.

2.3.1 Binary Logit Models

Binary Logit models, or binary logistic regression models, are commonly used to analyze the determining factors of locational urban growth. Binary logit models are a specific form of generalized linear models (GLMs) that include a wide range of models. **Generalized** linear models are not the same as **General** linear models, which are simply a specific form of linear regression models. GLMs differ from simple linear regression models in that they do not require the dependent variable Y to be normally distributed. The general assumption in GLMs is that Y follows an exponential family distribution, such as a binomial or multinomial distribution. As for this study, the goal is to explain the presence or absence of urban land-cover in a cell. The response variable is binary (categorical with two classes), which means that we are dealing with a binomial distribution.

In cases where the dependent variable is binary ($Y_i = 0$, if the cell is non-urban; $Y_i = 1$, if the cell is urban), one often uses linear logistic models, in which the log-likelihood ratio of the dependent variables is assumed to be in a linear relationship with the explanatory variables ($X = [X_1, X_2, \dots X_p]$):

$$\log(\frac{\Pr(Y_i = 1 \mid X_i = x_i)}{\Pr(Y_i = 0 \mid X_i = x_i)}) = \beta_0 + \beta\, x, \qquad \text{Eq. 1}$$

[1] Other terminologies used are: dependent variable, response variable

[2] Other terminologies used are: predictors, regressors, covariates

In the case of binary logistic regression models, X_i can be discrete, continuous, or a mix of both. The subscript 'i' indicates an individual cell, β_0 is the intercept with the y-axis and β is a vector of regression coefficients for the independent variable x. Eq. 1 can be rewritten in a way that allows us to estimate the probability of Y_i being one, given the set of explanatory variables:

$$\Pr(Y_i = 1 \,|\, X_i = x_i) = \frac{e^{\beta_0 + \beta x}}{1 + e^{\beta_0 + \beta x}} \,. \qquad \text{Eq. 2}$$

A binary logistic regression model, as shown in Eq. 1 and Eq. 2, is just one of the many models that fall under the category of generalized linear models. In this case, the generalization is the logit transformation of the conditional probability that $Y_i = 1$, which then allows us to make estimations with a linear model. Such a transformation is also known as a *link function*, as it links the probability of a binary dependent variable to the linear function of independent variables. GLMs always consist of the probability of the dependent variable (*random component*), a set of independent variables with a linear relationship (*systematic component*), and a *link function* that brings both components together.

The literature review in section 0 has shown that binomial logit models are prevalent for studying the correlation between driving forces and urban growth. Also, for more general land-use change studies, binomial logistic regression is a common approach (Verburg, Ritsema van Eck, *et al.*, 2004; Koomen *et al.*, 2015). Another popular model in urban growth studies is the probit model, which is very similar to the logit model, only it includes a probit (probability + unit) instead of a logit link function (McMillen, 1992; Li, Sun and Fang, 2018).

2.3.2 Penalized Logistic Regression

The theoretical concepts behind urban growth and the review of empirical evidence have highlighted the complicated relationship between a large number of driving forces and locational urban growth. Including all driving forces identified in the literature could make the binomial logit model extremely complex and eventually impair the model's performance, especially on a global scale. One possibility to optimize model complexity and performance is to add a penalty term to the logistic regression. Penalized logistic regression models place constraints on the coefficients, which prevents overfitting and reduces the number of independent variables without the need for the researcher to make a selection a priori. The three most widely used forms of penalized logistic regressions are the ridge regression, lasso regression, and the elastic net regression, which is a mix of a ride and lasso regression.

The LASSO (L1) penalty, first introduced by Tibshirani (1996) with essential adaptations by Zou (2006), penalizes the Maximum Likelihood Estimator by the absolute sum of the coefficients, thereby discouraging high parameter values. The magnitude with which the penalty is applied depends on the tuning parameter λ, which is optimized through cross-validation. A higher λ would mean that the parameters for more noise coefficients are set to zero, which means that fewer variables are kept in the model. The advantage of using the LASSO penalty is that independent variables with little or no predictive power will be automatically removed from the model, increasing efficiency and the risk of overfitting the model in exchange for an acceptable amount of bias. The LASSO method is similar to a stepwise selection method based on information criteria, as introduced by Granger, King and White (1995) and Sin and White (1996). The ridge (L2) penalty following Park and Hastie (2008) is very similar to the L1 penalty with the same goal. However, instead of penalizing the number of coefficients, it penalizes the absolute magnitude of the coefficients. Combining the two primary forms of penalized regression, the elastic net regression model attempts to find a parsimonious model with either a few non-zero coefficients or small absolute magnitude coefficients that best fit the input data.

2.3.3 Random forest models

Besides binary logistic regression models, the random forest model, first introduced by Breiman (2001), is a popular non-parametric classification and regression tool due to its ability to generate good results using standard parameter values and its robustness to noise. It constructs prediction rules without imposing strong prior assumptions on the functional form of the relationship between independent variables and the outcome variable. It is beyond the scope of this thesis to go deeper into random forest models that have been extensively covered in review articles, such as those of Biau (2012); Biau and Scornet (2016). The random forest model can be useful for complex datasets, where linear-based models may not be able to understand the boundary between two binary classes, resulting in poor model accuracy (Couronné, Probst and Boulesteix, 2018; Kirasich, Smith and Sadler, 2018).

2.3.4 Methodological issues related to urban growth studies

When analyzing driving forces behind urban land-use change, there are some relevant properties of the problem at hand that need to be considered. First of all, in spatial cross-sectional approaches as applied in this study, data is generally dependent across space (Anselin, 2003). In particular, the natural (geophysical) driving forces experience strong spatial dependencies, as does the dependent variable itself. Next to the interrelation of observation across space, we see dependencies across time and endogenous interaction between the different driving forces.

Next to spatial heterogeneity, spatial autocorrelation (the dependency of observations close to each other in space) is one of the main issues in spatial econometrics (Paelinck and Klaassen, 1979; Anselin, 1988). As with any statistical issue that threatens the assumption of independent error terms, the main goal of the research is to control for this nuisance to get consistent and efficient parameter estimates. Methodological integration of spatial dependence in binary regression models is a complex matter (McMillen, 1995; Beron and Vijverberg, 2004; Smith and LeSage, 2004). Two widely used general spatial models are the spatial probit and the spatial logit model (McMillen, 1992; Anselin, 2003). Within those models, the spatial dependency can be either included as a spatial lag or a spatial error dependence (Anselin, 2003). In spatial lag models, also known as autoregressive model, a spatially lagged form of the dependent variable is included as an independent variable in the model (Anselin, 1988, p. 35). Autoregressive models for binary responses were introduced by Besag in 1974 under the name of autologistic models. In the cases where the model includes covariates along with the response variable, we are talking about autologistic regression (ALR) models.

In land-use change models, it is common practice to include neighborhood rules to account for autocorrelation (Zeng *et al.*, 2008; van Vliet *et al.*, 2013). This reference to surrounding land-use of the same type is known as an autologistic specification. Dendoncker and colleagues (2006) have demonstrated that introgressive models are especially well suited to reproduce land-use patterns. According to Overmars, De Koning, and Veldkamp (2003), neighborhood effects can account for economic forces that are normally hard to include in regression models, but play an essential role in city growth (see section 2.1), such as the path dependency of urban areas and economies of scale.

Next to endogeneity problems due to spatial autocorrelation, problems can arise because of endogenous interactions between the driving forces. For modeling purposes, it is often assumed that locational urban growth drivers are exogenous, which can potentially lead to methodological issues and biased results, as many factors are actually endogenous. The endogeneity of selected drivers can especially become a problem when studying urban growth over a long period. Path dependency has a great role to play in, for example, infrastructure and urban expansion, as both are interrelated, and the causal relationship is not always clear (Verburg, Schot, *et al.*, 2004). Before selecting driving forces, it should be explicated whether the study aims to explain expansion patterns over a more extended period, or small urban area changes over shorter periods. Especially driving forces connected to second nature agglomeration forces such as the availability of jobs, or other socio-economic factors that become more important as a city expands, should be evaluated carefully before being included in the analyzes to avoid biased results. Instead of including economic drivers that are often endogenous to urban expansion, biases can be avoided by including proxies for market access and trade instead, such as the proximity to rivers or the sea (Verburg, Ritsema van Eck, *et al.*, 2004).

2.4 Empirical evidence

Even though a considerable amount of literature has focused on land-use change and the persistent growth of urban areas over the last few decades, only a few studies have explicitly focused on the drivers determining the spatial location of urban growth. Two general criteria have been applied for the selection of relevant empirical background literature: The first criterion was that the study focus has to be on locational urban growth and not on quantitative change. The second criterion was related to the method applied to determine drivers of locational urban growth. This resulted in the selection of studies that use an inductive (data-driven) approach to determine driving forces of urban expansion and the strength of their effect. Overmars, Verburg and Veldkamp (2007) concluded that inductive approaches are preferred over deductive approaches, when the goal of the study is to reproduce existing land-use patterns. The drawback of inductive approaches is that they do not clarify causal relationships between driving forces and urban growth. The first part of this section will provide a general overview of the different studies and their methodological approaches, followed by a more in-depth analysis of the driving forces of urban expansion identified in these studies.

Generally, the review of different empirical studies has shown that most studies focus on urban patterns that developed over longer periods or on change/growth of urban areas over smaller periods. Only a few global studies on locational urban growth exist, as mentioned in the introduction. One of these few existing global studies by Seto *et al.* (2011), analyzes global urban expansion between 1970 - 2000. A multivariate regression analysis was carried out to test the predictive power of different key drivers of urban areas. The analysis includes driving forces of quantitative change, such as the population growth rate and the driving forces of locational change, such as the coastal zone location. The dependent variable of the analysis was the decadal rate of urban land expansion, and the drivers were selected from "urban theory and models" (Seto *et al.*, 2011, p3). The multivariate regression model is not further specified. The study period and the dependent variable illustrate that the study's focus lies in explaining the resultant change in urban area and not the growth pattern itself.

A second global study by Seto, Güneralp and Hutyra (2012) predicts the extent of the world's urban areas in 2030 with an adopted version of the land-cover change model 'Geomod' (Pontius, Cornell and Hall, 2001). The modified model, called 'Urbanmod', determines the location of the new urban area based on an initial land-cover map and four additional driver maps. The approach is more sophisticated than the one by Seto *et al.* (2011) because the quantitative and locational change of urban areas is calculated in two different steps. The four drivers were used in statistical analysis to generate a suitability map for urban expansion. The description of the methodological approach does not include any additional information on the statistical approach nor any validation of the statistical analyzes. Both global studies use only a small number of drivers, without any regional differentiation.

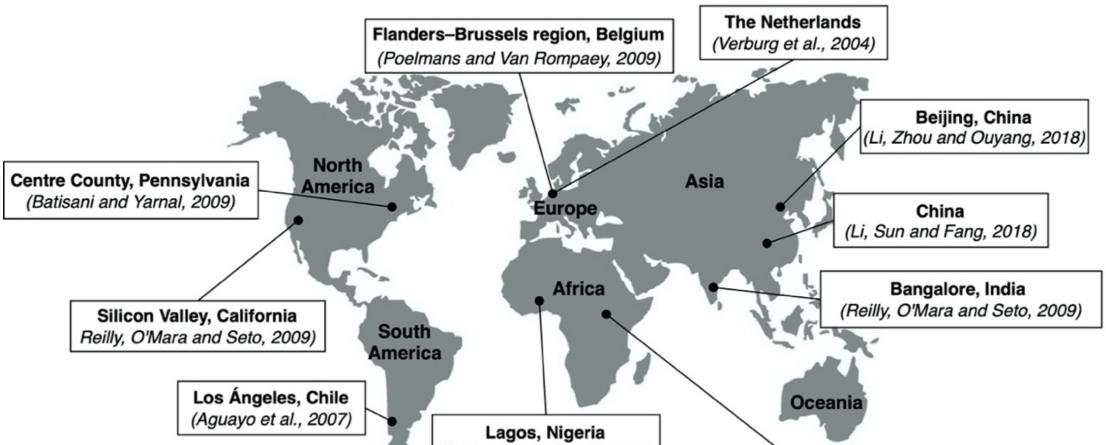

Flanders–Brussels region, Belgium
(Poelmans and Van Rompaey, 2009)
The Netherlands
(Verburg et al., 2004)
Beijing, China
(Li, Zhou and Ouyang, 2018)
Centre County, Pennsylvania
(Batisani and Yarnal, 2009)
North America
Europe
Asia
China
(Li, Sun and Fang, 2018)
Bangalore, India
(Reilly, O'Mara and Seto, 2009)
Africa
Silicon Valley, California
Reilly, O'Mara and Seto, 2009)
South America
Los Ángeles, Chile
(Aguayo et al., 2007)
Oceania
Lagos, Nigeria
(Braimoh and Onishi, 2009)
Kampala, Uganda
(Vermeiren et al., 2012)

Figure 2 | Continents and focus region empirical studies

Besides the few global studies, most of the current literature on locational change of urban areas focuses on a specific region or city. The majority of studies that fit the criteria for this literature review are on Chinese cities; however, only two of those been selected for further discussion. The study by Li, Sun and Fang (2018) includes a thorough review of other studies on Chinese cities. Outside of China, the number of studies fitting the criteria described above is small. To gain a first insight into the possible differences of driving forces between continents, at least one study of each continent (see figure 2) will be discussed. However, for most of the continents besides Asia, not many more studies exist that fit the criteria. For Oceania, no study with an inductive approach could be found, only literature-based studies, e.g., Bohnet and Pert (2010).

2.4.1 Methodological approaches

All studies shown in figure 2 use a binary logistic regression to determine the importance of driving forces behind urban growth. Predominantly the dependent variable is the change in urban area over a relatively short time period, e.g., 1984- 2000 (Braimoh and Onishi, 2007), 1993–2000 (Batisani and Yarnal, 2009). Some studies analyzed several time periods, e.g., 1989–1995, 1995–2003, and 2003–2010 (Vermeiren *et al.*, 2012). The results by Vermeiren *et al.* (2012) show that the effect of driving forces varies over the different time periods, such as the effect of the distance to a road which became weaker over the years. A recent study by Cao *et al.* (2020) looks at Hangzhou's (China) urban area changes within four different time periods. They find that not only the effect of specific driving forces changes over time, but also the number and type of driving forces.

Some deal with spatial autocorrelation by using spatial lag Probit models or spatial error Probit models or both (Li, Sun and Fang, 2018). Many studies account for autocorrelation by including a reference to the surrounding land-use, as will be discussed in the following section. Prevalent among the studies to measure the accuracy of the developed classification models is the receiver operating characteristics (ROC) procedure (Braimoh and Onishi, 2007; Poelmans and Van Rompaey, 2009; Vermeiren *et al.*, 2012; Li, Sun and Fang, 2018). The ROC curve is usually given as a single value, which represents the area under the ROC curve, often abbreviated as AUC. The ROC curve compares the true positive with the false positive rate. If the model randomly assigned all suitable locations for urban development, the ROC value would be 0.5. If there is a perfect fit between the real land-use and the predicted land-use, in our case land-use being urban or non-urban, the ROC value would be 1. Other goodness-of-fit parameters often used to compare models are the Akaike's Information Criteria (AIC) and the log-likelihood (Li, Sun and Fang, 2018).

2.4.2 Driving forces of locational change

The set of driving forces used to explain and model land-use change strongly depends on the simplification of the model and the researcher's theoretical assumptions (Verburg, Schot, *et al.*, 2004).

For this study, a twofold selection method is used based on empirical studies and machine learning. In the first step, drivers of locational change have been identified in the literature, which will be discussed in this section with reference to the framework by Hersperger and Bürgi (2007) (see Figure 1). The second selection step involves stepwise and penalized regression, further discussed in the methodological section 0.

2.4.2.1 Natural/spatial drivers

The natural/spatial drivers, as Hersperger and Bürgi (2007) have defined them, can be divided into two subgroups. The first includes physical characteristics such as elevation and soil type, but also natural hazards such as flooding. Such drivers correspond with the theoretical construct of first nature agglomeration causes. The second group comprises of so-called neighboring factors, which are, for example, the current land use surrounding a cell, as well as existing infrastructure networks.

Natural drivers

The most commonly used natural drivers in the reviewed empirical studies are the slope and elevation (Batisani and Yarnal, 2009; Reilly, O'Mara and Seto, 2009; Bohnet and Pert, 2010; K.C. Seto, Güneralp and Hutyra, 2012; Li, Sun and Fang, 2018). Whereas all studies find a negative correlation between the slope and urban growth, meaning that urban growth becomes less likely on steep terrain, the direction of the effect of elevation on urban growth varies between regions. Studies from the continents Asia, South- and North America find a negative effect of elevation on urban growth (Aguayo *et al.*, 2007; Batisani and Yarnal, 2009; Li, Sun and Fang, 2018). Both studies on regions in Africa, however, find the opposite and conclude that an increasing elevation has a positive effect on urban growth (Braimoh and Onishi, 2007; Vermeiren *et al.*, 2012)

Another natural driver mentioned several times is soil type. Studies for South- and North America have shown that urban land expansion is positively correlated with productive agricultural soils (Aguayo *et al.*, 2007; Batisani and Yarnal, 2009; Reilly, O'Mara and Seto, 2009). Another study by Mundia and Aniya (2005) found that soil is one of the most important factors influencing the spatial patterns of the urban expansion of Nairobi city, Kenia's capital. The authors have found that clay soil reduces the probability of land being transformed into urban area due to their poor water permeability. On the other hand, "well-drained red soils of volcanic origin, covering the western and northern areas have promoted the urban expansion in these directions" (Mundia and Aniya, 2005, p. 2845). Verburg *et al.* (2004) also report soil and drainage conditions as essential drivers. Their results show that in the Netherlands, soils with high loam content are preferred for urban development. Hietel, Waldhardt and Otte (2005) found that biophysical factors such as slope, elevation, and soil mainly function as constrains for urban development. This is in accordance with the definition by Verburg, Schot, et al. (2004) motioned in section 2.2.1, saying that natural drivers steer urban development in an indirect way over allocation decisions.

Other natural occurrences that seem to influence the probability of an area being urbanized are *sea and river floods*. The correlation between flood-prone areas and urban growth can be positive (Bohnet and Pert, 2010; Vermeiren *et al*., 2012) as well as negative (Poelmans and Van Rompaey, 2009). The negative effect can be easily tied to the economic risk imposed by floods. The positive correlation is more difficult to explain. A possible explanation could be that positive effects from being close to the sea or a river outweigh the present risk. If the latter is true, a positive correlation should exist between rivers and urban growth. To better understand the relationship between natural hazards and urban growth, it would be interesting to study other hazards next to floods. So far, no studies could be found that analyze the effects of other natural hazards on urban expansion patterns.

Spatial drivers

The *proximity to freshwater resources* such as rivers and lakes on the probability of urban growth seems to vary around the globe. In China, the presence of a river or lake nearby decreases the probability of urban expansion (Li, Sun and Fang, 2018, p. 70), possibly due to the flood risk. In Africa, on the other hand, the probability of urban development increases near freshwater (Braimoh and Onishi, 2007, p. 511). The *distance to waterworks* is mentioned as a crucial factor for urban development in Africa as well (Braimoh and Onishi, 2007). Both cases show that access to a water source is essential for urban development in Africa, but not in other generally more developed regions. Next to the availability of freshwater, the average distance to the nearest sea-navigable river is also a measure for market access to a city/country. Fujita, Krugman and Mori (1999) highlighted that most old cities in the U.S are "located along the northern part of the Atlantic coast or navigable rivers, reflecting the importance of sea- and river transportation for trade with Europe as well as within the U.S."

Studies from different continents (Asia, Africa, Europe and South and North America) have found the *proximity to infrastructure* to be the most crucial determinant of locational urban expansion (Aguayo *et al*., 2007; Braimoh and Onishi, 2007; Batisani and Yarnal, 2009; Poelmans and Van Rompaey, 2009; Reilly, O'Mara and Seto, 2009; K.C. Seto, Güneralp and Hutyra, 2012; Vermeiren *et al*., 2012; Li, Sun and Fang, 2018). Some of these studies differentiate between types of infrastructures such as highways (main roads), national ways (secondary roads) and railways (Aguayo *et al*., 2007; Poelmans and Van Rompaey, 2009). The *proximity to infrastructure* can be directly linked to travel time and respectively, transportation costs. From extensive work on the core-periphery model, we know that there is a strong relationship between transportation costs and agglomeration (P. Krugman, 1991). If transportation costs are high, we expect dispersion. At intermediate transportation costs dispersion and agglomeration could be the dominant mechanisms. In contrast, low transportation costs, in theory, result in agglomeration. It is, therefore, reasonable to assume from an economic theory perspective that a good infrastructure network stimulates agglomeration. Glaeser and Kahn (2003) suggest that transportation costs have played an important role in the growth of Wall Street, along with agglomeration benefits.

As mentioned in section 2.3.4, some reference to the surrounding urban sites is often included in the analyzes to account for autocorrelation. The *density of the built-up area in the neighborhood* has been included in nearly all empirical studies analyzed for this chapter (Cheng and Masser, 2003; Aguayo *et al*., 2007; Braimoh and Onishi, 2007; Reilly, O'Mara and Seto, 2009; Müller, Steinmeier and Küchler, 2010; Vermeiren *et al*., 2012; Li, Zhou and Ouyang, 2013; Li, Sun and Fang, 2018). The effect of a dense urban area nearby seems to differ regionally. In studies from Chile, Uganda, China, and India, a positive effect was found of dense urban area on urban expansion (Aguayo *et al*., 2007; Reilly, O'Mara and Seto, 2009; Vermeiren *et al*., 2012; Li, Sun and Fang, 2018). In Silicon Valley (U.S.), however, the effect seems to be reversed, and a higher density in the surroundings decreases the chances of urban growth (Reilly, O'Mara and Seto, 2009). This could be an indication of the higher vehicle dependency in North America compared to the other regions. Another or even a complementary explanation could be the higher demand for urban land per capita, often referred to as urban sprawl observed in the United States and other highly developed regions like Western Europe (Wolff, Haase and Haase, 2018).

2.4.2.2 Economic drivers

Economic drivers, such as consumer demands and market structures, are challenging to measure and difficult to include in land-use change models, especially on a global scale. Most often, spatial variables, such as the distance to roads (travel time) *or the distance to the socio-economic center* of a region, are included in the analyzes as a proxy for economic drivers (Veldkamp and Lambin, 2001). Many studies found a negative correlation between urban growth and the distance to the socio-economic center (Braimoh and Onishi, 2007; Reilly, O'Mara and Seto, 2009; Vermeiren et al., 2012; Li, Sun and Fang, 2018). Another commonly used economic driver is the *employment potential*, which is used as a relative measure for job accessibility (Poelmans and Van Rompaey, 2009; Reilly, O'Mara and Seto, 2009). The population density in rural or suburban areas often represents consumer demand and market potential (Liu, Zhan and Deng, 2005). The population density must not be confused with the population growth rate, which is a driver of quantitative urban growth. Li, Sun and Fang (2018) and Seto, Güneralp, & Hutyra (2012b) also found a positive effect of high population densities on urban growth.

2.4.2.3 Political drivers

Only a handful of political drivers have been included in land-use change models. As mentioned in section 2.2.1, especially large-scale studies lack information on political drivers. One policy that has been included in several studies is nature protection. Braimoh and Onishi (2007) found that in Nigeria, the probability of urban development increases with increasing *distance to protected areas*. Nieves et al. (2020) also include the distance to protected areas as a political driver. Verburg and his colleagues (2004) included two different policies on spatial planning in the Netherlands in their analyzes. The policies ensured that urban growth occurred in designated areas and that other cities were not getting too big, and nature would not disappear (Dieleman, Dijst and Spit, 1999).

2.4.2.4 Overall picture

A couple of findings stand out after analyzing the drivers behind urban development. The first one being that key drivers do indeed differ strongly between studies in different regions. Differences have been identified in both the number and types of drivers, as well as the effect they have on urban development. Natural/spatial drivers seem to be overrepresented in most studies analyzed compared to the other four groups of drivers. After analyzing the results of Poelmans and Van Rompaey (2009), Vermeiren et al. (2012) and Li, Sun and Fang (2018), it seems that the drivers behind urban development become more complex, and forecasting models less accurate, for countries or regions with highly developed urban areas.

3 METHODOLOGY

In this section, the methodical approach will be explained in more detail. First of all, an overview of the dependent and independent variables used for the regression analyzes will be provided. The independent variables are a selection of the driving forces identified in the literature in 2.4.2. Not all identified drivers could be included due to missing global datasets on, for example, employment potential or spatial policies. The second part of this chapter includes a description of the different logistic regression models. Models were tested with the urban extent in 2010 as dependent variables, as well as the change in urban area between 1990-2010.

3.1 Data

3.1.1 Dependent variable – Urban Area

Urban area 2010. The dependent variable is a raster with cells that are either urban or non-urban. Cells labeled as non-urban include all other land-use classes, except for water, which has been masked prior to the analysis. The dependent variable is in the context of a binary logistic regression, also known as outcome or response variable. No consistent global definition for "urban land-use" exists, which complicated the initial classification for a global dataset of urban-land use. Eventually, the dataset containing information on the presence of built-up area from the European Commission's Joint Research Centre (JRC) was used to classify cells as urban. A more detailed description of the classification of urban areas can be found in the background report for the urban growth model *'2UP'* (Huijstee *et al*., 2018).

Urban growth 1990 – 2010. The literature study has shown that the influence of different determining factors can vary depending on the time period that is examined. For this reason, a secondary analysis was carried out to explain the change in urban area between 1990 and 2010. Similar to the urban extent in 2010, the data was supplied by PBL and is based on the dataset from the JRC. The variable initially had three categories: cells that were already urban in 1990, cells that became urban between 1990 and 2010, and cells that were non-urban in 1990 and stayed that way. Due to the focus on the change in urban area between 1990-2010, all cells that were already urban before 1990 were excluded from the analysis.

3.1.2 Independent variables

Based on the empirical evidence presented in section 2.4.2, different variables determining the spatial location of urban growth have been selected. PBL Netherlands Environmental Assessment Agency has supplied all datasets for the explanatory variables. A comprehensive list of the source data can be found in appendix 8A. Table 1 provides an overview of the variables, some of which had to be pre-processed, as explained in this section.

Table 1 | Descriptive statistics of the data

Variable	Min	Max	Mean	Type of data
Country grid	0	250	121	Not included in the final analysis
Continent grid	0	7	2.9	Not included in the final analysis
Governance factor	0	1		Dummy
Urban Area Density	0	1	0.004	Continuous
Coastal Urban Area Density	0	36.3	0.01	Continuous
Elevation [meters]	-407	8519	626	Continuous
Slope [degrees]	0	53.2	1.6	Continuous
Terrain ruggedness index (TRI)	0	7	2.6	Continuous
Travel time	0	11.8	0.2	Continuous
Soil type	0	35	13.6	Categorical
Protected Area	0	1	0.12	Dummy
Flood Prone Area	0	1	0.04	Dummy
Landslides (Earthquake)	0	1	0.03	Dummy
Landslides (Precipitation)	0	1	0.06	Dummy
Earthquakes	0	8	1.4	Continuous
Distance to river	0	95.8	0.53	Continuous

Country & Continent grid. The continent codes have been used solely to split the primary dataset in separate datasets for each continent and are not further included in the analysis. Similarly, the country grid is not included in the primary analysis but was used to add a country-specific governance index to the dataset, as explained in more detail hereafter.

Governance factor. The theoretical and empirical background has shown that political drivers are essential in the urban growth process but are often underrepresented in urban growth studies. Therefore, a governance factor in the form of a binary variable (0 or 1) has been included in the analysis. The factor was created with the Worldwide Governance Indicators (WGI) dataset of the World Bank. The governance indicator project was initiated by Kaufmann and Kraay in 1999, with major contributions by Zoido and Mastuzzi. Their definition of governance and a reference to the original datasets can be found in appendix 8A. The dataset includes six different measures for governance, which are available for each country over several years. An average for all categories over time was calculated to reduce nuisance from outliers. To create only one value indicating the country's governance level, the average of the six different categories per country was taken as well. Initially, the values range from -2.5 to 2.5. For simplicity, all negative values have been transformed into zeros and all positive values to ones. This allows us to interpret the index as a dummy variable, and straightforwardly create interaction variables. The governance index has been added to the main dataset based on the country code.

(Coastal) Urban Area Density – Autologistic specification. The urban area density represents the presence of urban sites in the surrounding grid cells and is therefore an autologistic specification, as described in section 2.3.4. Following Anselin (1988), a spatial weight matrix has been applied to obtain a density value ranging from zero to one. A spatial lag is introduced by calculating the surrounding urban area density based on the urban area in 1990, and not on the urban area in 2010. The density of the coastal urban area is calculated in the same way (based on the coastal urban area 1990 and with a spatial weight matrix). The only difference is that the proximity to the coastline is included.

TRI. Following Huijstee and her colleagues (2018), the terrain roughness index (TRI) was added as an independent variable, along with slope and elevation. The TRI is an indicator of the topographic heterogeneity, based on the difference in elevation between a cell and its surrounding cells (Riley, Degloria and Elliot, 1999).

Travel Time. The travel time variable is an index value of the travel time in minutes to the nearest city center. The index ranges from 0 to 12, with 0 representing the exact center of a city and 12 all grid cells with a travel time that exceeds 60 minutes to the city center. The calculations are based on road data and the city centers of urban clusters with more than 50.000 people (see appendix 8A.)

Soil type. Data on the type of soil was obtained from the harmonized world soil database (HWSD). The data contains 33 categories. An overview of all categories can be found in appendix 8F. Category 29-33 are labeled as "non-soil" in the official report and are therefore excluded from the analysis (Nachtergaele *et al.*, 2010). The soil type variable is transformed into 'factor' in a pre-processing step, in order for R to interpret the soil type as a categorical variable. By default, R takes the first category as a reference category and compares all other categories to this reference category. Because the soil type variable was converted from a categorical variable to numerous dummy variables, it is crucial to control for variables with zero or near-zero variance. For example, in the case that one type of soil is not at all present or less abundant within a continent, one of the dummy variables consists mostly out of zeros and has little to no variance. Those variables have little predictive power and can cause problems with the regression models later. Due to these reasons, all soil types with very low variance have been removed prior to the regression analysis. This was done with the `nearZeroVar()` function from the 'caret' package in R (the following section includes a more detailed description of the R package). The function removes variables based on two criteria: the frequency ratio `(freqCut)` and the unique value percentage (uniqueCut). The following values have been used for the criteria: freqCut = 10 and uniqueCut = 15.

Protected areas. Protected areas are one of the few policy-related variables available on a global scale, and the literature review shows that a significant correlation exists between this variable and urban growth. Protected areas have, therefore, been included as a dummy variable, which takes on the value one if more than half of a grid cell is covered by a protected area and zero otherwise. Additionally, an interaction variable has been created between the governance indicator and protected areas, to see if protected areas restrict urban growth less in countries with a negative governance factor.

Natural hazards. As discovered during the review of the empirical studies, flood risk is often included in the analyzes and found to be a determining factor behind urban growth. Other natural hazards have so far not been included in other studies, which is why this study includes next to flood-prone areas, also areas that are affected by landslides and earthquakes. A distinction is made between landslides triggered through precipitation or earthquakes. Both types of landslides, as well as flood-prone areas, are included as dummy variables. Earthquakes are included as a continuous variable, with values ranging from zero to seven, with higher values representing higher intensities.

Distance to rivers. The distance in kilometers to the nearest river from each cell was included as a proxy for freshwater availability, as well as transportation possibility and market access.

Correlation matrix. A common problem in regression models is the occurrence of multicollinearity when independent variables are not, in fact, independent. Therefore, the correlation between the independent variables was tested before the analysis. The results shown in Figure 3 indicate that no strong correlations (correlation > ± 0.8) exist between independent variables. The strongest positive correlation was found between the TRI, elevation, and slope. It is also noticeable that a strong correlation exists between the dependent variable (Urban Area 2010) and the urban area density.

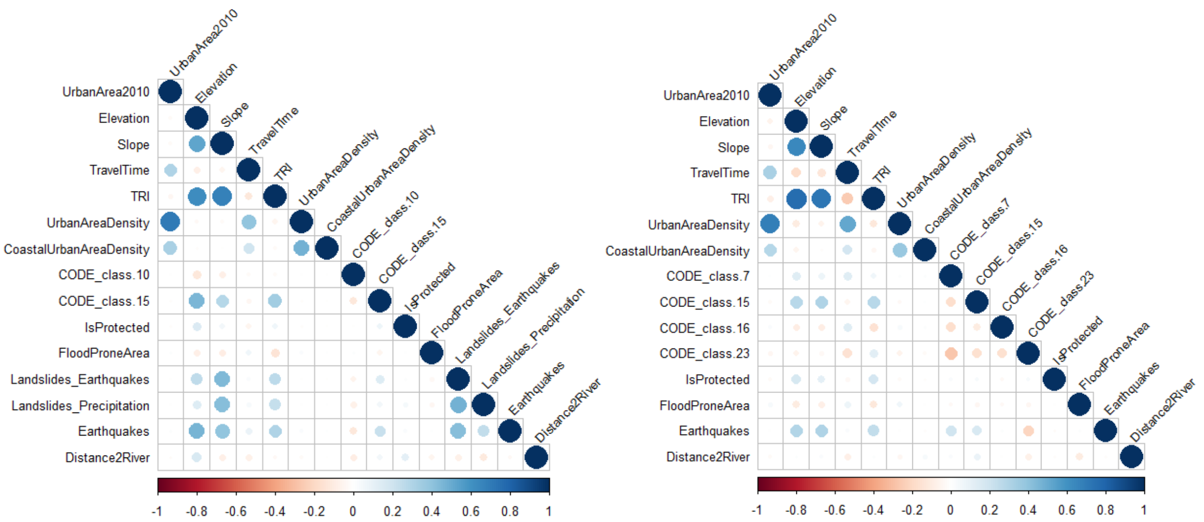


Figure 3 | Correlation matrix Europe (left) and Asia (right) for the urban area 2010

In figure 4 we can see the same correlation matrices as in figure 3, only this time with urban growth between 1990-2010 as the dependent variables. It becomes immediately apparent that the correlation between the dependent variable and the urban area density decreased substantially, viz. for Europe from 0.68 to 0.28. The correlation matrices for the other continents can be found in appendix 8C.

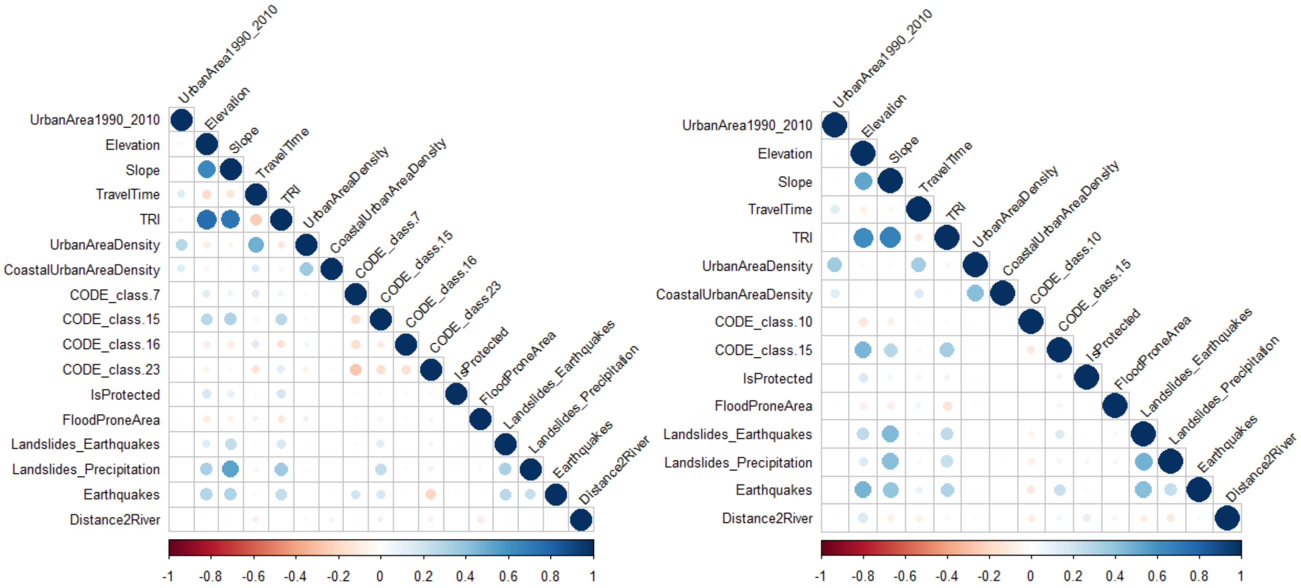


Figure 4 | Correlation matrix Europe (left) and Asia (right) for change 1990 - 2010

3.2 *Analysis*

The analysis consists of two parts, each of which deals with one of the two research questions. This study aims to understand the effect of different drivers on urban growth and their regional differences. A binary logistic regression analysis was chosen to analyze this effect based on the theoretical and empirical background. The analysis was performed with the open-source R package '*AutoGLM*'[3]. This package was developed by Bo Andrée to perform regression analyzes, such as linear, logit, and probit models, with large datasets. The package is very suitable in our case because we do not have to choose a priori which variables to keep in the model. The optimal number of variables for a parsimonious model is found through an automated stepwise regression, further explained in section 3.2.1. The second part of the analysis is focused on the second research aim and was performed with the R package 'caret' (**C**lassification **A**nd **RE**gression **T**raining), developed by Max Kuhn for predictive modeling (Kuhn, 2008). Caret is one of R's most widely used packages for supervised learning, also known as predictive modeling. Supervised learning is a specific case of machine learning with a target variable; in this case, the dependent variable 'UrbanArea2010'. Because the target variable (UrbanArea2010) is a categorical

[3] The package can be installed in R with the following two commands:

```
library(devtools)
install_github("BPJandree/AutoGLM")
```

variable with two classes (either urban or non-urban), a classification will be performed instead of regression with a numeric response variable. ‘Caret’ is particularly well suited to compare different models with one another (Kuhn, 2008). The main expectation of the perfect model we are trying to build is that it predicts the correct class as often as possible, based on the input variables (driving forces) and parameters.

Before carrying out the two parts of the main analysis, a few data processing steps had to be performed. The global assessment on a high resolution, together with a large number of independent variables, has resulted in an extensive dataset of approx. 22 GB, including a little more than 200 million observations. Because a dataset of this size is beyond a 64GM RAM machine's capabilities, the dataset was split into six separate datasets for each continent. The following pre-processing and analysis steps have been executed for each continental dataset separately. Firstly, all rows with missing values have been removed, which reduced the number of observations to approx. 9 – 42 million, depending on the continent. More information on the sizes of the continental datasets can be found in appendix 8E.

Secondly, the size has been reduced further by taking a sample from each continental dataset. The sampling method used for this step is part of the open-source R package ‘*AutoGLM*.’ The smart sampling algorithm draws random samples and compares those to the original data based on t-tests and F-tests. Only samples that are comparable in means and variance to the original dataset are kept. Andrée and Koomen (2017) have shown that this sampling strategy can produce a representative sample dataset that can then be used for further analysis. Because the continent's datasets have not been equal in size, different shares have been taken from the datasets to create the samples (see appendix 8E). Eventually, all continent specific samples are equal in size and in practice just small enough to make further calculations on a 64GB RAM machine computationally feasible.

3.2.1 Logistic classification analysis with autoGLM

For the first part of the analysis, different logit models for each continent are produced with AutoGLM. AutoGLM builds optimal models with a stepwise regression process, by minimizing the information criteria through the elimination of independent variables with low predictive power. AutoGLM’s default optimization strategy was used, which optimizes the regression models based on the AIC. This is done by first fitting all variables and then removing variables with low predictive power until the AIC is optimized. The AIC uses the log-likelihood measure that informs us how good our model fits the data (the less negative the log-likelihood, the better the model fit) and adds to that a penalty for the number of variables. Whereas the log-likelihood only informs us of how good the model fits the data, the AIC also tells us how parsimonious a model is. For example, if two models have the same log-likelihood, but one of them has fewer covariates (independent variables), that model would score better based on the AIC. When comparing different models, the model with the lowest AIC will be the most parsimonious. The AIC is a relative measure, meaning that it becomes only a valuable measure if we compare different AIC values to each other.

Because of the second research aim, logit models for each continent are produced with the urban pattern in 2010 as a dependent variable as well as with the urban change between 1990-2010. This allows us to compare the difference in model performance for both independent variables and the differences in the effects of driving forces. When looking at the effects of the driving forces, it is necessary to remember that due to the logistic classification model, the coefficients represent log-odds ratios that need to be interpreted accordingly.

3.2.2 Predictive modelling with Caret

The second part of the analysis aims to compare different types of models to find the most suitable one for predicting future urban change. The most important property of the model is, therefore, that it performs well on new data. To prevent overfitting and ensure the model performs well on new data, a training and testing sample is created from the original dataset. The models will be trained with the training sample and tested on the test sample. If we train and test the model on the same dataset, there is no way of knowing how well the model generalizes for new data. The Caret package is especially well equipped to perform out-of-sample validations. The train/test split was set at 0.5, splitting the dataset in 50% test and 50% training data. The train/test split was created with the '`createDataPartition`' function, which has the advantage over the more traditional random 'sample' function that the class balance stays intact. Before splitting the data into train and test datasets and training the model, R's random seed is set to two to ensure that all random processes such as the train/test data split are reproducible.

Model training and tuning

The first part of the analysis describes how a parsimonious model can be found by stepwise regression. Another strategy to prevent overfitting is the application of penalized regression, as described in the theoretical background (section 2.3.2). Stepwise and penalized regression are similar in the sense that both are used to create parsimonious models in exchange for an acceptable amount of bias. Since it is not clear in advance which model will make the best predictions, it is common practice in predictive modelling to train and compare different kinds of models. For this analysis we will compare a baseline logit model without any penalty term to a logistic regression with L1 or L2 penalty, or a mix of both (elastic net). These are the three most common forms of a penalized logistic regression as described in section 2.3.2. Furthermore, a random forest model is produced to see if a non-linear model can describe the relationship between driving forces and urban growth better than the linear models. To build the random forest model, R's 'ranger' implementation was used, which is particularly suitable for high dimensional data (Wright and Ziegler, 2017). For all models the data has been centered and scaled to improve the accuracy from the machine learning algorithms.

Customizing trainControl. Before training the different classification models, a custom '`trainControl`' function was used, to specify the method by which the models are created. The code for the '`trainControl`' function can be found in appendix 8D. Within the '`trainControl`' function, it is specified that a repeated cross validation will be executed, each time with a total number of five folds. In this case during the 5-fold 2-repeat cross validation, the test dataset will be split into five sub-samples, two times. Cross-validation is one of the best methods to calculate the out of sample validation. Because the test dataset is split into 5 different subsamples, we are able to obtain multiple estimates for the out-of-sample error.

In the control function, we also specify the summary function ('`summaryFunction`'), that determines which performance summary (evaluation metric) will be computed for the model. Because the data is highly imbalanced with more non-urban than urban cells, it is vital to choose the correct classification evaluation metric. The literature review has shown that the ROC value is often used to create models and compare their predictive power. In our case, however, the ROC can be misleading because of the class imbalance of our dataset. For example, the sample dataset for Africa contains 5.396.133 non-urban cells and only 8601 urban cells. If the model correctly predicts all non-urban cells, the true negative rate (specificity) will be very high. Even if the model falsely classifies all urban cells and the true positive rate (sensitivity) is subsequently meager, the overall area under the ROC curve will still be high because of the relatively small number of urban cells. When selecting the performance metric, it is also essential to consider that we are mainly interested in correctly predicting urban sites. In such a case, the area under the precision-recall curve (AUC) is a more sensitive measure of model performance. The precision tells us what proportion of the positive classified urban sites is actually urban. In contrast, the recall (sensitivity) informs us what proportion of the existing urban sites our model has identified. Other terms for precision and recall are respectively the user's and producer's accuracy, which are often used in similar studies to evaluate model performance (De Vasconcelos *et al.*, 2001; Andree and Koomen, 2017)

A second custom '`trainControl`' function was created that includes one additional line of information, that deals with the general imbalance of the two classes (urban and non-urban). One solution to this problem of class imbalance is to subsample the training data. For this analysis the majority class (non-urban cells) has been randomly down-sampled to match the number of urban cells during the resampling process, by adding a down-sampling option to the second '`trainControl`' function (see appendix 8D).

Custom tuning grid – logit. For the penalized logit models, a custom tuning grid was used in addition to the custom '`trainControl`' function. The custom tuning grid is used because the default tuning grid is very small, and using a custom tune grid allows us to explore more potential models. As discussed in section 2.3.2 there are two main forms of the penalized regression: the lasso and the ridge regression. For the 'glmnet' function we can define two tuning parameters: alpha and lambda. Alpha determines which penalty term is used and lambda regulates the strength of the penalty. To automatically test which penalization results in a better model, the tuning parameter alpha was set to 0, 0.5 and 1, which means the pure ridge regression (α=0), the pure lasso regression (α=1) and an elastic net regression ($\alpha = 0.5$) are tested simultaneously for all defined values of lambda. In total ten values for lambda between 0.0001 and 1 have been tested, where a higher value for lambda yields a simpler model. The full tuning grid can be found in appendix 8D.

Custom tuning grid – random forest. The random forest tuning grid includes several tuning parameters that control the structure of the forest or its trees. Choosing the right parameters is mostly an empirical matter. It is outside of this study's scope to go further into the reasoning behind the chosen parameters. For discussions on the impact of different parameters see, for example, Scornet (2017). For this study, the '*mtry'* was set equal to the square root of the number of variables. The '*min.node.size'* was set to ten, which is standard for probability models (Wright and Ziegler, 2015). As usual, for random forests, the '*gini*' '*splitrule*' was used.

Model evaluation

Once the model training and tuning is completed, it is important to evaluate how well the models work. We generally want to know the goodness-of-fit of the model, and we want quantifiable metrics that give us an objective measure on how good the model performs on the test data. We have already constituted that the log-likelihood and the AIC are suitable goodness-of-fit measures for our application, as those measures account for the number of dependent variables. The model's predictive power will be evaluated based on the area under the precision-recall curve (AUC). The last measure used to evaluate the models produced with the caret framework is the log-loss statistic.

The log-loss is a classification loss function often used to evaluate classification models. The log-loss function is a useful addition to the AUC evaluation metric because it is based on whether the model predicts the correct class, and because it includes the probability for each class. In general, the lower the loss-log statistic is, the more often the model predicts the correct class with high certainty. If the actual cell is urban, and the model predicts with a probability of 1 that this cell is urban, the cost function would be zero—the log-loss increases as the probability with which the correct class is predicted decreases. The log-loss function also puts a high penalty on wrong classifications with a high probability. If the model classifies with a high certainty a non-urban cell as urban, the log-loss value will increase drastically.

4 RESULTS

In this chapter, the results of the analyzes described in the previous section will be presented. First, the results from the logistic classification models produced with AutoGLM will be presented for all continents. From these results, the main conclusion concerning the first research aim will be drawn. In the second part of this chapter, the results of the models produced with 'caret' will be discussed for three continents; Africa, Asia, and Europe. These results will be discussed mainly with the second research aim in mind.

Results logit models with AutoGLM**Table 2** presents the results from the stepwise logit model, with the urban pattern in 2010 as the dependent variable. This means the different coefficients must be interpreted as the effect that driving forces had on an urban pattern that evolved over a long period. The coefficients are given in log-odds ratios. Next to the coefficients, the table includes two goodness-of-fit measures, the log-likelihood, and the AIC. As described in the method section, AutoGLM's optimization strategy is based on the AIC and the AIC in tun is calculated with the log-likelihood and the number of variables. Comparing the AIC's from the different continental models shows that the models for Africa, Australia and South America better fit the data than the ones for Asia, Europe, and North America. The model fit for Europe is by far the worst. The predictive power of the continent specific models will be further discussed in the second part of this section.

Overall, when looking at the regression results for the different continents, we find that the number of significant variables differs per continent. For example, in Europe, thirteen variables were kept in the model, whereas for Africa, only seven variables significantly contributed to the outcome. It is apparent from Table **2** that only significant variables have been kept in the models through stepwise regression, with the soil type 23 in Europe as the only exception. The number of significant variables kept in the models for the other four continents can be found in the second row of Table **2**. Not only the number of relevant variables differs between countries, but also the kind of drivers that influence the presence of urban sites as well as the strength of their effect.

The biggest differences exist between soil types, as not all soil types are present in all continents. Variables that are found to be significant for all continents are urban area density (positive effect), TRI (negative effect), and travel time (positive effect). On all continents except for South America, the presence of a protected area is negatively correlated with urban development. All other variables are only significant in four or fewer continents. Proximate urban land-use seems to be the most crucial determinant of whether a cell becomes urbanized or not for all continents. The positive effect a dense urban surrounding has on an individual grid cell varies per continent. The effect is less strong in Europe and North America compared to the other continents. The probability of a cell becoming urban increases most strongly with the presence of surrounding urban sites in Africa. The effect of the coastal urban area density on the other hand is less commensurate between the continents. First of all, this driver is only significant in Asia, Europe and North America. Secondly, a cell near the coast with a high density of surrounding urban sites is less likely to be urbanized than a cell further away from the coast in Asia, while this effect is reversed in Europe and North America.

The coastal urban area density is not the only driver that has a positive effect in some continents a negative effect in others. For example, soil type 16[4] is significant in Australia and North America but has a negative influence on urban growth relative to the base category[5] in Australia, but a positive influence in North America. Another example is the presence of a flood-prone area, which has a negative influence on urban growth in Australia, North America, and South America and a positive effect in Europe. The coefficient in Europe is 0.13, which means that the presence of flood risk increases the probability of a cell being urban by exp(0.13) = 1.14 times compared to cells without flood risk.

Other natural hazards included were earthquakes and landslides. Earthquakes were found to be significant in Asia, Europe and North- and South America, and the effect is positive which means that if the earthquake intensity of a cell increases by one unit, the probability of that cell to become urban increases. This finding is somewhat counterintuitive, but could be due to the fact that most earthquakes are too weak to form an actual hazard. Landslides triggered by earthquakes are insignificant and the presence of landslides triggered by precipitation has a significant negative effect only in Asia. Another interesting finding is that with an increasing distance to rivers the probability of urban growth decreases in Africa, Asia, Australia and Europe. The effect is by far the strongest in Australia and the weakest in Europe.

[4] Very shallow soils over hard rock or in unconsolidated very gravelly material

[5] Soils with subsurface accumulation of low activity clays and low base saturation

Table 2 | AutoGLM logistic regression results for the urban pattern in 2010

UrbanArea 2010	Afrika	Asia	Australia	Europe	North America	South America
Nr. of variables	7	10	10	13	10	8
Urban Area Density	23.26 (0.39) ***	19.02 (0.23) ***	20.02 (0.41) ***	14.26 (0.10) ***	13.75 (0.18) ***	18.95 (0.31) ***
Coastal UA. Density		-0.13 (0.03) ***		0.28 (0.02) ***	0.08 (0.02) ***	
Distance to river	-0.77 (0.06) ***	-0.99 (0.08) ***	-1.99 (0.20) ***	-0.32 (0.05) ***		
Elevation				0.001(0.0001)***		0.0002(0.0001)**
Slope		0.06 (0.03) **		0.11 (0.01) ***	0.08 (0.02) ***	
TRI	-0.27 (0.03) ***	-0.76 (0.03) ***	-0.49 (0.08) ***	-1.04 (0.03) ***	-0.73 (0.03) ***	-0.22 (0.04) ***
Travel Time	0.30 (0.01) ***	0.21 (0.01) ***	0.09 (0.02) ***	0.07 (0.003) ***	0.22 (0.01) ***	0.31 (0.01) ***
Protected Area	-1.37 (0.21) ***	-1.21 (0.16) ***	-3.53 (0.66) ***	-0.89 (0.06) ***	-2.47 (0.25) ***	
Flood Prone Area			-0.65 (0.23) ***	0.13 (0.04) ***	-0.31 (0.10) ***	-0.56 (0.17) ***
Earthquake		0.057 (0.008) ***		0.08 (0.01) ***	0.07 (0.01) ***	0.05 (0.02) ***
Landslide [PR]		-0.50 (0.13) ***				
Soil Class 3	-0.58 (0.11) ***		-1.36 (0.18) ***			
Soil Class 6	-0.25 (0.14) *		-0.52 (0.15) ***			
Soil Class 7				0.43 (0.03) ***		
Soil Class 9						0.23 (0.09) **
Soil Class 10		-1.03 (0.11) ***				
Soil Class 15				0.20 (0.05) ***		-0.52 (0.19) ***
Soil Class 16			-0.86 (0.16) ***		0.38 (0.05) ***	
Soil Class 23				0.06 (0.04)		
Soil Class 24					-1.36 (0.13) ***	
Soil Class 27			-1.33 (0.19) ***			
Constant	-6.40 (0.07) ***	-5.32 (0.05) ***	-6.04 (0.13) ***	-4.38 (0.03) ***	-5.38 (0.07) ***	-7.81 (0.08) ***
Observations	1,250,635	1,603,720	1,230,718	1,301,614	1,421,123	1,351,126
Log Likelihood	-6,349.625	-12,148.330	-2,981.162	-43,947.050	-13,055.390	-4,886.853
AIC	12,715.250	24,318.650	5,984.325	87,922.110	26,132.780	9,793.707

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 3 shows the results from the logit model with the change in urban area between 1990 and 2010 as dependent variables. One of the first things to notice is that the number of significant drivers has decreased for most continents except for Africa and Asia, where respectively the number has remained the same and one variable has been added. Similar to the results for the urban pattern in 2010, the goodness-of-fit is the highest for Australia and South America and worst for Europe. Only the urban area density and the travel time remained significant for all continents, the TRI is no longer of importance in Africa and South America. The urban area density remains the most important driving force for all continents, however, the strength of the effect did decrease compared to the results for the urban pattern in 2010. For example, in Asia the coefficient decreased from 19.02 to 13.05.

Not only the strength of the effect of urban area density has changed, but also the strength of most other drivers. The coefficient from the flood-prone area in Europe for example increased from 0.13 to 0.51, which means that between 1990-2010 a cell with flood risk is 1.7 time more likely to be urbanized than a cell without flood-risk. It is difficult to see the difference in variable importance between Table **2** and Table 3. Four graphs were made (Figure 5 and Figure 6) to give a clear view of the differences in driving forces for the long time period (urban pattern 2010) and the urban growth between 1990-2010. The differences are presented on the basis of two examples for Europe and Asia.

Table 3 | AutoGLM logistic regression results for the urban growth between 1990-2010

Urban Growth 1990-2010	Africa	Asia	Australia	Europe	North America	South America
Nr. of variables	7	11	7	10	6	5
Urban Area Density	18.74 (0.42) ***	13.05 (0.26) ***	10.68 (0.45) ***	7.73 (0.13) ***	7.23 (0.21) ***	12.82 (0.45) ***
Coastal UA. Density		-0.24 (0.05) ***		0.16 (0.02) ***	0.12 (0.03) ***	-0.17 (0.07) **
Distance to river	-0.95 (0.09) ***	-0.74 (0.1) ***		-0.24 (0.09) ***		
Elevation		-0.0002(0.0001)*	-0.01(0.001) ***			
Slope	-0.12 (0.03) ***	0.13 (0.03) ***		0.05 (0.02) ***		-0.14 (0.04) ***
TRI		-0.83 (0.05) ***	0.31 (0.10) ***	-0.74 (0.04) ***	-0.70 (0.03) ***	
Travel Time	0.353 (0.01) ***	0.27 (0.01) ***	0.42 (0.03) ***	0.19 (0.01) ***	0.39 (0.01) ***	0.40 (0.02) ***
Protected Area	-1.79 (0.27) ***	-1.36 (0.22) ***		-0.79 (0.09) ***		
Flood Prone Area		0.40 (0.09) ***	-0.91 (0.31) ***	0.51 (0.06) ***	-0.36 (0.12) ***	
Earthquake	0.06 (0.02) **		0.09 (0.03) ***	0.09 (0.01) ***	-0.03 (0.01) **	
Landslide [EQ]		-0.55 (0.23) **				
Soil Class 6	-0.36 (0.18) **					
Soil Class 9						0.68 (0.12) ***
Soil Class 10		-1.01 (0.14) ***				
Soil Class 15				0.20 (0.08) **		
Soil Class 16			-0.35 (0.23)			
Constant	-7.04 (0.06) ***	-5.65(0.07) ***	-7.53 (0.15)***	-5.58 (0.05) ***	-6.02 (0.06) ***	-8.99 (0.10) ***
Observations	1,249,502	1,603,720	1,229,504	1,286,483	1,416,191	1,349,509
Log Likelihood	-5,233.067	-8,433.102	-1,909.135	-20,113.130	-8,530.248	-2,299.546
AIC	10,482.130	16,890.200	3,834.269	40,248.260	17,074.500	4,611.092

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

The variable importance graphs shown in Figure 5 and Figure 6 are based on the baseline logistic regressions (without stepwise elimination of variables), which is why more variables are shown than in the two tables above (The regression results of the baseline models can be found in appendix 8G). However, this does not change the relative importance of the variables shown in Table **2** and Table 3. The variable importance is based on the z values and is automatically scaled in such a way that the most important variable has an importance of 100. What immediately strikes us is what was already very clear in the table; namely that the urban area density is the most important variable. Looking at the change in variable importance for Europe between the long and the short time period shown in Figure 5, we see that the variables with the biggest changes are the travel time and the TRI. The travel time became nearly twice as important for the shorter and more recent study period and the TRI became relatively less important. We furthermore notice that the elevation became insignificant for the shorter study period. All other variables stayed relatively constant with an importance between 0 and 20.

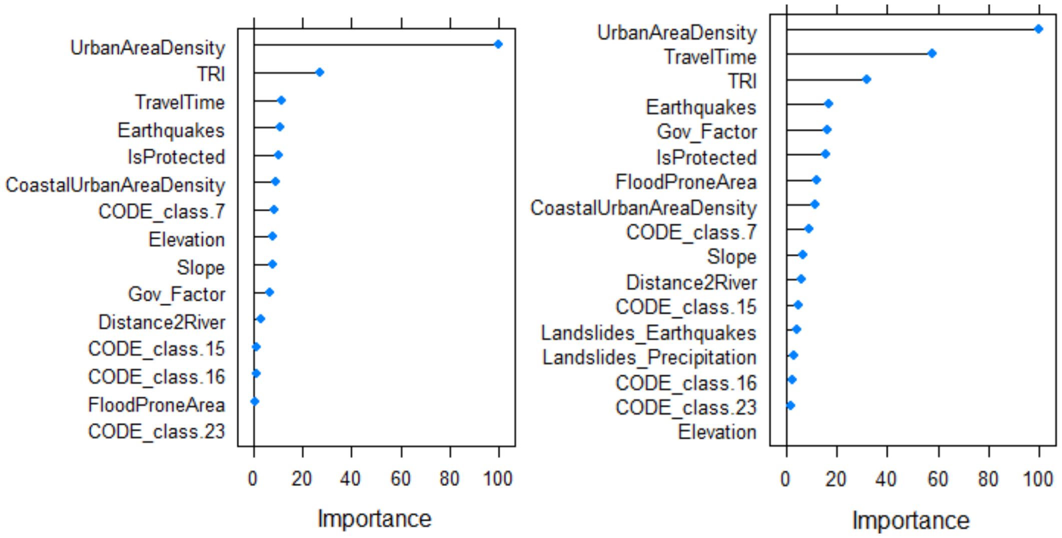


Figure 5 | Variable Importance Europe; Left: Urban Area 2010; Right: Urban Growth 1990-2010

Comparing the variable importance graphs of Europe (Figure 5) and Asia (Figure 6), we notice that the distance to rivers is much more important in Asia. We also see the same increase in importance of travel time for the shorter time period in Asia, as we have seen in Europe. Also, in Asia the importance of the remaining variables seems to be constant between both time periods.

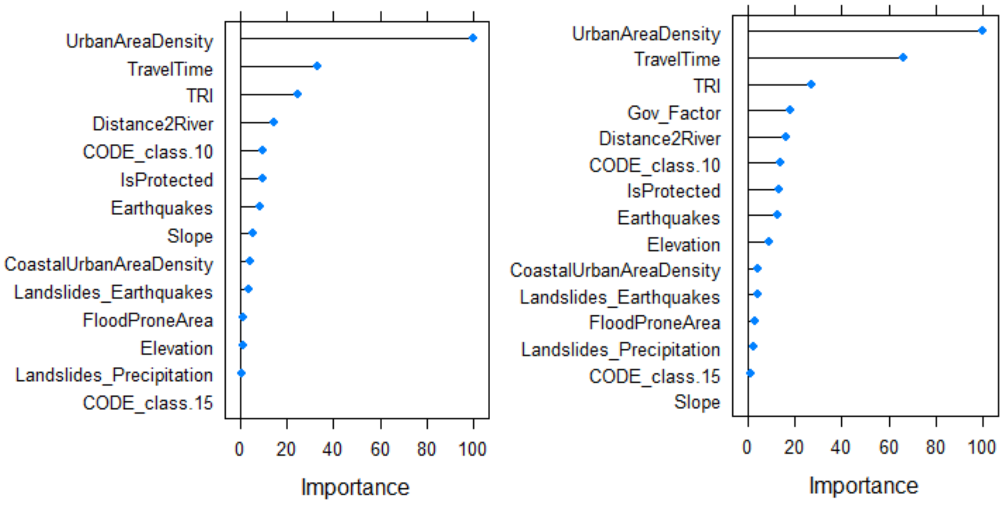


Figure 6 | Variable Importance Asia; Left: Urban Area 2010; Right: Urban Growth 1990-2010

Now that we have analyzed the results of the importance and regional differences of driving forces, we can focus on the predictive power of the models themselves. When analyzing the performance statistics of the models as presented in Table 4 and Table 5, we are mainly interested in the out of sample precision and sensitivity. The overall agreement between the values for the training sample (within sample) and the test sample (out of sample) confirms the validity of our sampling procedure. As mentioned in the methodology section, the high overall accuracy is misleading due to the class imbalance with predominantly non-urban cells. The sensitivity or recall shows us how many of the urban cells have been classified correctly by the model. The lowest recall was achieved for Europe and Africa, where respectively only 48.7% and 49.5% of all urban sites have been correctly classified as urban by our model. The highest recall was attained for Australia and North America, where respectively 68.1% and 61.2% of all urban sites have been correctly classified. The precision (the proportion of the classified urban sites that were correctly identified as such) is high for all continents and ranges between 76.6% for Europe and 85.3% for Australia. This means that both the precision and recall are lowest for Europe and highest for Australia.

Table 4 | Predictive power logit models (Outcome variable: Urban Area 2010)

		Africa	Asia	Australia	Europe	North America	South America
Within sample	Overall accuracy	0.999	0.998	0.999	0.990	0.998	0.999
	Recall (Sensitivity)	0.500	0.508	0.681	0.487	0.612	0.567
	Precision	0.809	0.789	0.854	0.768	0.821	0.790
Out of sample	Overall accuracy	0.999	0.998	0.999	0.990	0.997	0.999
	Recall (Sensitivity)	0.495	0.517	0.688	0.481	0.609	0.580
	Precision	0.809	0.783	0.853	0.766	0.822	0.802

Table 5 shows that with the chosen set of driving forces it is harder to predict urban growth between 1990 and 2010 than the urban pattern in 2010. Both sensitivity and precision are very low for all continents when the outcome variable is the urban change between 1990 and 2010. For Europe the sensitivity dropped to 4.4 %.

Table 5 | Predictive power logit models (Outcome variable: Urban Growth 1990-2010)

		Africa	Asia	Australia	Europe	North America	South America
Within sample	Overall accuracy	0.999	0.999	1.000	0.997	0.999	1.000
	Recall (Sensitivity)	0.132	0.096	0.128	0.044	0.108	0.076
	Precision	0.442	0.366	0.344	0.285	0.343	0.288
Out of sample	Overall accuracy	0.999	0.999	1.000	0.996	0.999	1.000
	Recall (Sensitivity)	0.154	0.086	0.107	0.044	0.120	0.074
	Precision	0.458	0.334	0.283	0.287	0.358	0.255

4.1 *Results caret predictive modelling*

As explained in the methodological section, three different models are produced with the caret package: a logit model (the benchmark model), a penalized logit model, and a random forest model. The model is produced for Europe and Asia; Once with the complete sample dataset and once with an adapted version, in which the number of non-urban cells has been downsampled to match the number of urban cells. Selecting the optimal type and strength of penalization was left to the penalized logit model itself, an example of which is shown in Figure 7. We see that for each alpha (0, 0.5, 1) ten different values of lambda (regularization parameter on the x-axis) are tested. Eventually the combination of alpha and lambda with the highest AUC is selected to fit the model. The optimal combination differs per dataset as can be seen in Table 6. The predictions made with the three models presented in this section will also be compared to the stepwise regression model from the previous section based on their precision and recall values.

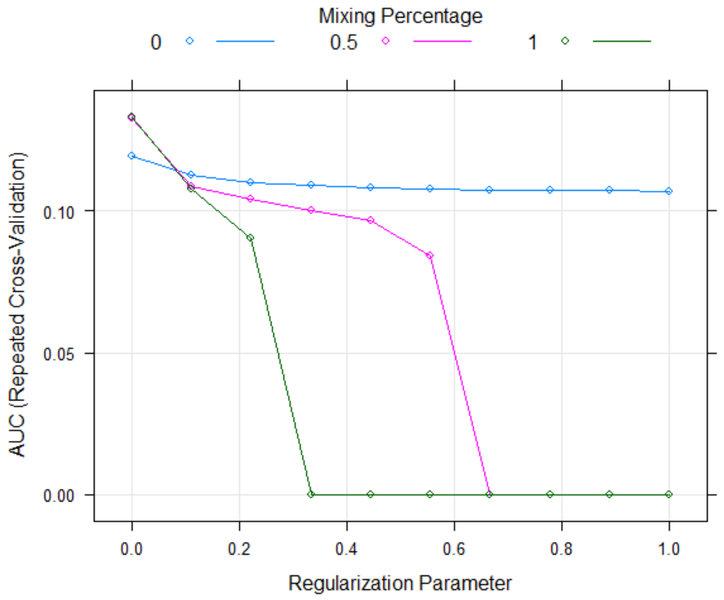


Figure 7 | Penalized logistic regression for Europe 1990-2010 with downsampled majority class

Table 6 | Optimal parameters for the penalized logistic regressions

		2010			1990-2010		
		alpha	lambda	Type of model	alpha	lambda	Type of model
Europe	normal	0.5	0.111	Elastic net	0	0.111	Ridge
	down	1	0.0001	Lasso	1	0.0001	Lasso
Asia	normal	0.5	0.0001	Elastic Net	0	0.111	Ridge
	down	1	0.222	Lasso	0.5	0.0001	Elastic Net

Europe

Figure 8 shows the performance results of the three models for the urban pattern in 2010 in Europe. The benchmark logit model archived a precision of 77% and a recall of 49%, which is very similar to the results of the stepwise regression with autoGLM (precision = 77%, recall = 48%). However, we can see in Figure 8 that the random forest model slightly surpasses the benchmark logit model with a precision of 78% and a recall of 53%. The parameters that yield the best penalized logit model for Europe are alpha = 0.5 and lambda = 0.111 (see Table 6). An alpha of 0.5 means that the penalized regression is a mix of lasso and ridge regression (elastic net), and a lambda of 0.111 means that the penalization of the coefficients was not very strong. The penalized logit model has a precision and recall of zero, which means no urban cells have been correctly or incorrectly classified. The AUC is in this case very misleading. The penalized logit model with a down sampled majority class is a lasso regression and performs better than the normal penalized model, however this is not clear in the figure. We notice that for the logit and random forest model, the AUC is lower when the majority class was down-sampled, due to a very low precision. Meanwhile, the recall is much higher for the models with down-sampled majority class. The F value shown in Figure 8 is the harmonic mean of the precision and recall.

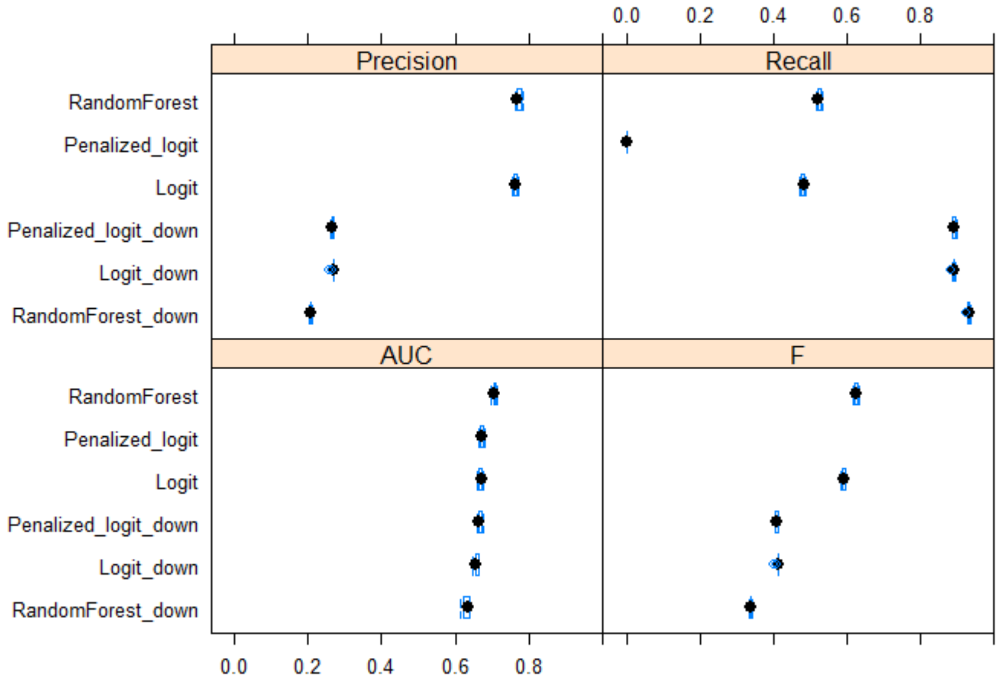


Figure 8 | Model performance Europe (Outcome variable: Urban Area 2010). Penalized logit = Elastic Net. Penalized logit down = Lasso.

In Figure 9 the performance of the model predicting urban growth in Europe between 1990-2010 is shown. As expected, based on the previous results, the predictive power for the shorter time period is much lower than for the urban pattern in 2010. We can see in Figure 9 that the AUC of the random forest model is in turn the highest, although the recall is very low with 2%. The logit model and penalized logit model with downsampled majority class are respectively the second and third best models. Compared to the random forest model the precision of both models is much lower in exchange for a high recall of 83%. The benchmark model with original class proportions performs worst in predicting urban growth between 1990-2010.

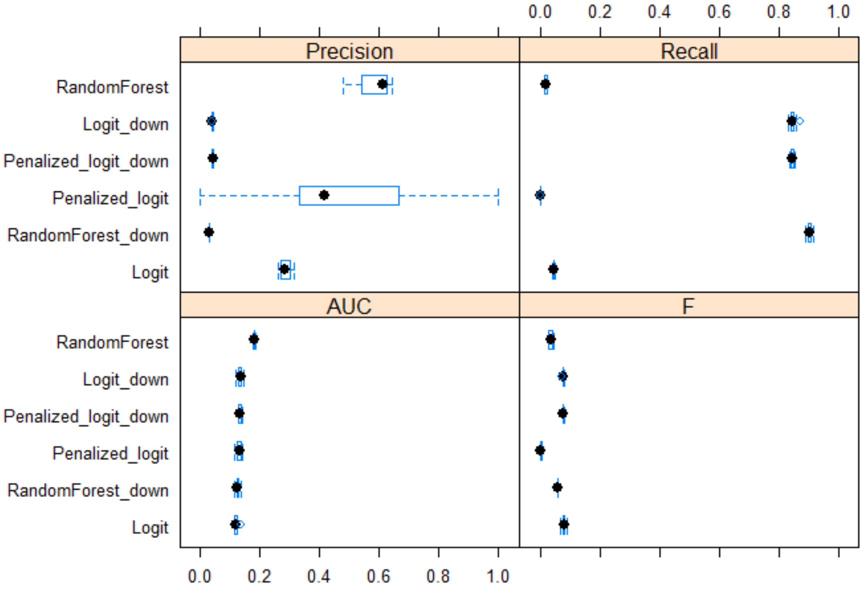


Figure 9 | Model performance Europe (Outcome variable: Urban Growth 1990-2010). Penalized logit = Ridge. Penalized logit down = Lasso.

In this context it is also important to show why a careful decision has to be made with regards to the model evaluation metric. If we for example would have chosen to evaluate the models based on the ROC as shown in Figure 10, we would have come to the conclusion that all models perform exceptionally well. The ROC for all models lies between 0.94 and 0.96. This is due to the fact that all models are especially good in predicting non-urban cells (specificity). The sensitivity is the same as the recall in Figure 9.

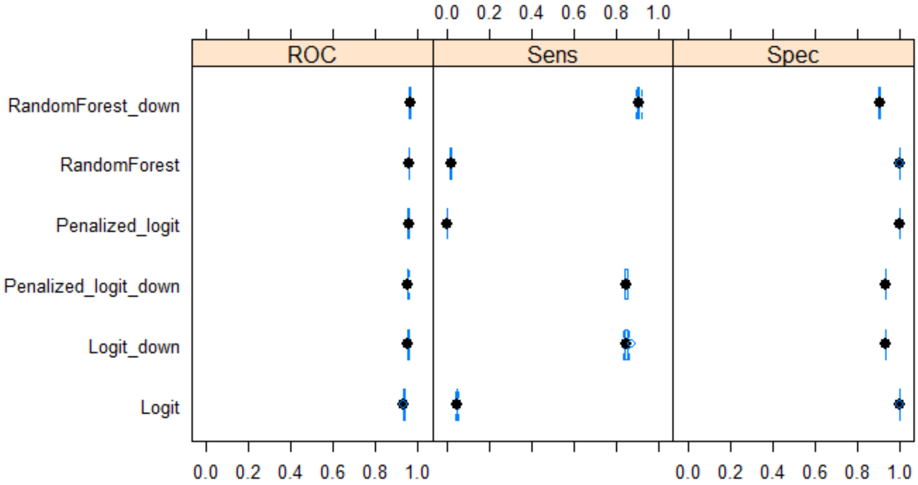


Figure 10 | Model performance based on the ROC for urban growth in Europe between 1990 – 2010. Penalized logit = Ridge. Penalized logit down = Lasso.

Figure 11 shows what happens to the variable importance under down-sampling and penalization. The variable importance from the benchmark model (left figure in the top row) was already shown in the previous section for both time periods (see Figure 5). We saw that the urban area density, the travel time and the TRI were the three most important variables to predict whether a cell will be urbanized between 1990-2010. In Figure 5 we also saw that that the governance factor, that was disregarded by the step wise regression in part one, is within the five most important variables for the short time period between 1990-2010. For the longer time period the governance factor was less important, which is also retraceable in the regression results shown in appendix 8G. The coefficient from the governance index is positive and increased from 0.14 for the long period to 0.22 for the shorter and more recent period. When downsampling the majority class, the influence of the governance factor increases even more. A positive coefficient of 0.22 means that a cell with a positive governance index is 1.3 times as likely to be urbanized than a cell with a negative governance index. The interaction variable between the governance index and the protected areas was insignificant for both time periods, which means that there is no difference in the effect of protected areas between countries that have a positive or negative governance index.

In the bottom row of Figure 11, the variable importance of the penalized regression model with and without downsampled majority class is shown. The variable importance is very similar to the benchmark logit model. The most striking result is the increased importance of the coastal urban area density. With the original class balance, the ridge penalty yields the model with the highest AUC. When the majority class is downsampled, the lasso penalty provides the best model, shown in Figure 7. The strength of both penalizations is low, so the influence of the variables has not changed much compared to the regular logit model. If the lasso penalty had been stronger, more variables would have been disregarded similar to the stepwise regression. That this is not the case shows us that all nearly included drivers, except for one in each model, are important to optimize the AUC.

Asia

Figure 12 and Figure 13 show the model performance for Asia for the long and short period, respectively. The confusion matrix with the exact values can be found in appendix 8G. The results for the long period are very similar to the results for Europe. The most important difference in this case is that the penalized logit models correctly predicted some of the urban cells. Again, the random forest model has the highest AUC with a score of 0.73. The models with downsampled majority class perform worse than the models with imbalanced classes. As we have seen in Europe's results, the precision goes down and the recall up when the majority class is downsampled. The highest precision is achieved by the random forest model with 97.4%, and the random forest model with balanced classes achieved the highest recall with 96.2%. The stepwise linear logit model for Asia form the first part of the analysis only achieved a precision and recall of 78.3% and 51.7%, respectively. The penalized logit model with imbalanced classes performs similarly to the stepwise regression model with a precision and recall of 79.1% and 51%, respectively.

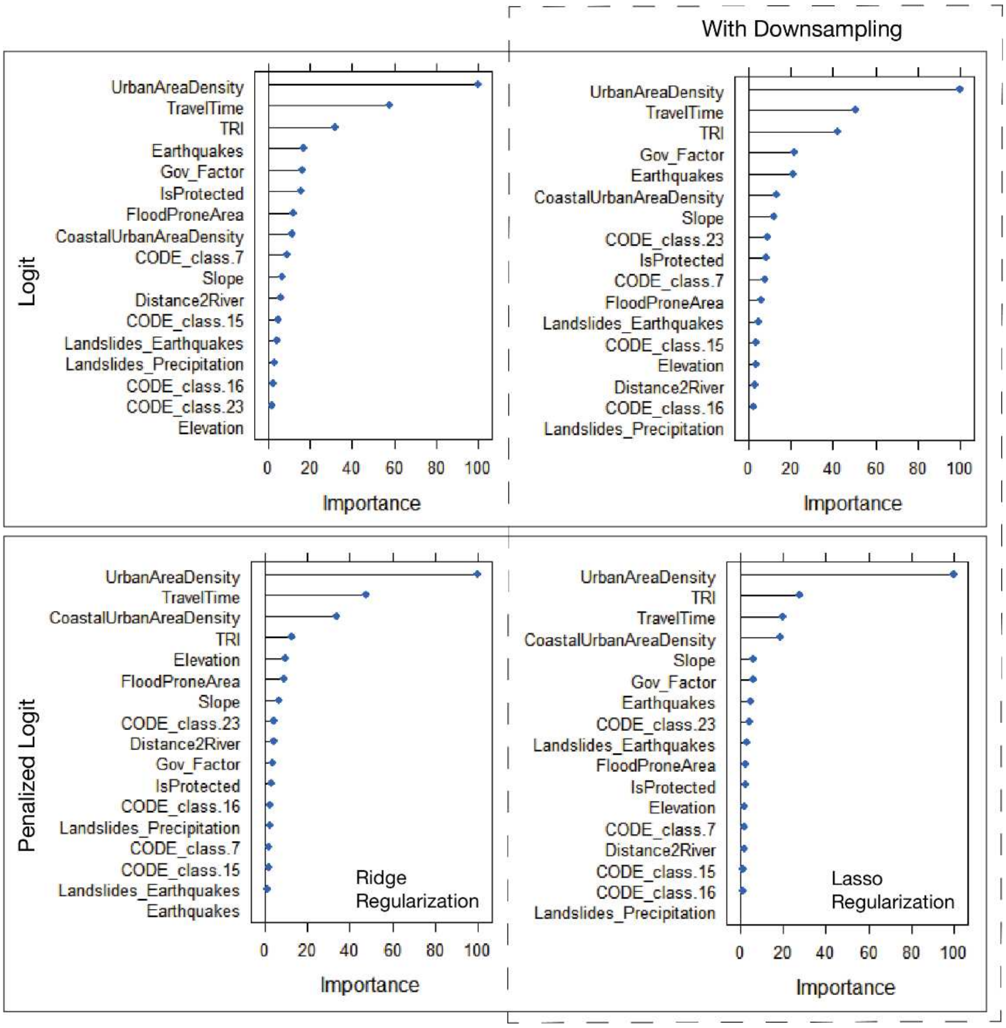


Figure 11 | Variable importance Europe (Urban Growth 1990-2010) for: normal Logit regression in the top row, penalized logit regression on the bottom row, down sampled majority class in the right column.

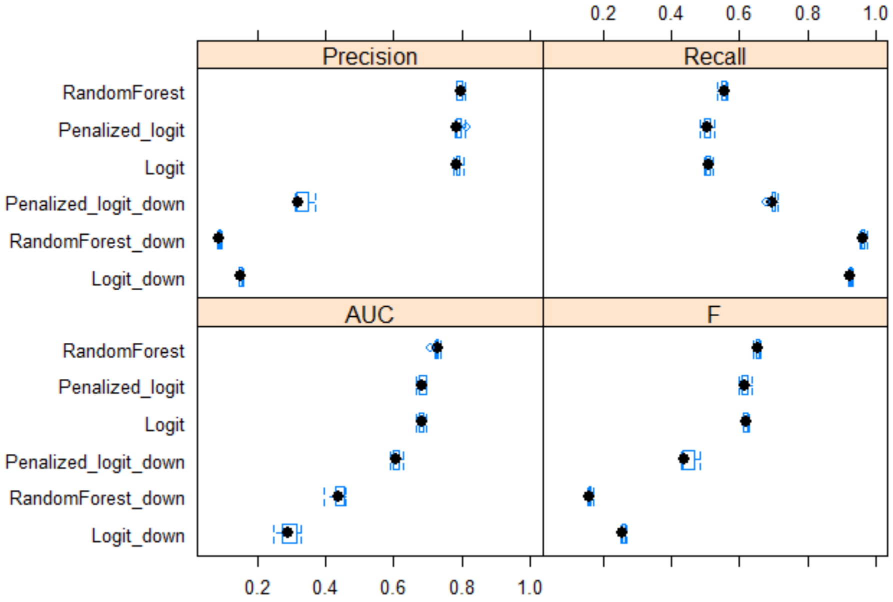


Figure 12 | Model performance Asia (Outcome variable: Urban Area 2010). Penalized logit = Elastic Net. Penalized logit down = Lasso.

Generally, the results of the models that explain urban growth between 1990-2010 in Asia differ from the results for Europe. One similarity is that the random forest model achieves the highest AUC. In general, the AUC of all models is much lower that the AUC of the models for the longer period. Different for this continent is that the penalized logit model and penalized logit model with balanced classes have the second and third best performance.

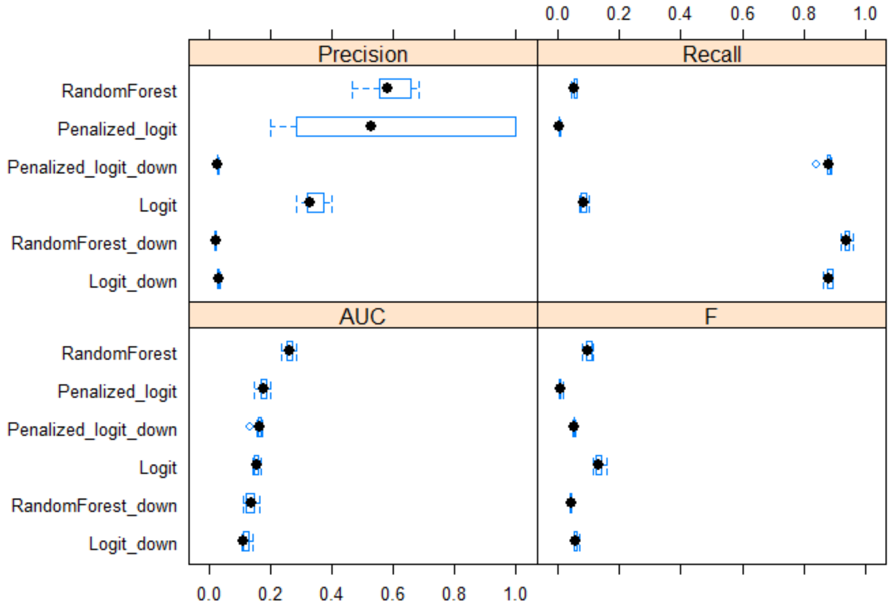


Figure 13 | Model performance Asia (Outcome variable: Urban Growth 1990-2010). Penalized logit = Ridge. Penalized logit down = Elastic Net.

As discussed in the methodological section, the last performance measure of interest is the log-loss that is in compared to the AUC also based on class prediction probabilities. A lower log-loss value is associated with a better model. Table 7 shows that the random forest model performs worst for both continents and both time periods. This is exactly opposite to the evaluation based on the AUC, where the random forest showed to be the best model. The best performing models for both time periods in Europe and for the short time period in Asia are the logit and penalized logit models. The penalized logit model describing the urban pattern in Asia has the best performance with a log-loss of 0.86. This is the only case in which the penalized logit model performs better than the logit model.

Table 7 | Results log-loss statistic for Asia and Europe.

	Asia		Europe	
	1990-2010	2010	1990-2010	2010
Logit	8.87	8.16	7.01	6.03
Logit (down)	3.25	3.81	2.33	2.97
Penalized Logit	6.90	7.69	5.74	4.23
Penalized Logit (down)	3.30	0.86	2.32	2.90
Random Forest	14.44	18.06	11.87	9.86
Random Forest (down)	4.41	6.32	3.53	4.60

5 DISCUSSION

This master thesis set out with the aim to analyze the effect driving forces have on locational urban growth and how these drivers differ regionally. Understanding these drivers can improve global urban growth models and enable better predictions of city expansion patterns. As mentioned in the literature review, so far, no global urban growth models exist that incorporate regionally specific drivers. In reviewing the current literature on locational urban growth, initial insights were gathered on which drivers play an essential role in urban growth and how those drivers might differ per region. For this study, specific binary logistic classification models have been used to analyze continental differences of driving forces. A stepwise regression based on the optimization of the AIC was performed to find a parsimonious model per continent without introducing too much bias.

The stepwise regression analysis results have shown that the number of variables kept in the model to get the most parsimonious fit differs per continent. The stepwise regression was performed with the open-source R-package AutoGLM, which proved to be very useful and reliable in performing the regressions with vast datasets. The model fit based on the AIC has shown that the models produced with AutoGLM for Africa, Australia and South America outperform the ones for Asia, Europe, and North America. The model fit for Europe is by far the worst, which is in line with existing findings that forecasting models become less accurate for countries or regions with highly developed urban areas.

Overall, all variables apart from landslides triggered by earthquakes have, at least in one continent, significantly contributed to the predictive power of the model. The most important driver of urban growth in all continents is the density of surrounding urban sites. This is in line with our expectation of a strong spatial autocorrelation of urban areas. The positive effect surrounding urban sites have on urban growth seems less strong in Europe and North America. Again, this is in line with our expectations, as the literature review has shown that urban sprawl becomes more important in highly developed continents like Europe and North America (Wolff, Haase and Haase, 2018). The literature review shows that the density of the built-up area has a negative effect on the case study for Silicon Valley in North America (Reilly, O'Mara and Seto, 2009). This is not confirmed to be true for the whole continent by our results. This thesis' findings have shown that the coastal urban area density has a positive impact in Europe and North America and a negative impact in Asia. This could be due to the large number of city metropole areas near the coast in Europe and North America.

Other variables that were found to be important in all continents are the TRI, which has a negative effect, and the travel time with a positive effect on urban growth. The negative correlation between an increasing TRI and urban growth and be easily explained, as less homogeneous terrain makes it physically more difficult to build cities. As described in the theoretical background on driving forces, the travel time to the city center is often used as a proxy for economic drivers such as consumer demands, market structure and even employment potential (Veldkamp and Lambin, 2001). The results of a positive correlation between the travel time variable and urban growth in all continents, confirm the high importance of economic drivers and second nature agglomeration forces for urban growth. The travel time therefore forms a very abundant room for future research to make further use of this opportunity. Other variables that could be included next to the travel time to the nearest city center are, for example, travel time to the nearest city edge or the travel time to the nearest employment center, as in many large cities these are relatively far away from the city center (Angel and Blei, 2016).

The second variable referring to surrounding land-use is the coastal urban area density. Inferring from the results, this variable is less important than the urban area density. To disentangle the effect the distance of the coast has on urban growth, we would advise for further research to include the distance to coast as a sole variable. In accordance with the literature review, the results of this thesis have shown that some drivers can have a positive, as well as negative effect on urban growth depending on the continent. Examples are in our case the soil type 16 and the presence of flood risk. These findings emphasize the need for regionally specific drivers in global urban growth models.

The theoretical and empirical background has made it clear that political drivers are very important for urban growth (Hersperger *et al.*, 2018). Nevertheless, they are often underrepresented in urban growth studies, especially on a global scale (Seto *et al.*, 2011). For this reason, the governance index based on the WGT dataset of the World Bank has been included in the analysis. In addition to the governance index, an interaction variable of the protected areas and the governance index was included. The hypothesis was that urban growth in countries with a negative governance index would be less restricted by protected areas. The results show that neither the governance index nor the interaction variable have been kept in the parsimonious models obtained with the stepwise regression in AutoGLM. This shows us that there is no difference in the effect of protected areas between countries with a positive or negative governance index. Protected areas had a strong negative effect on the probability of urban growth, in all continents except South America. Comparing the stepwise regression results with the benchmark model containing all variables, we see that the governance factor itself is significant, although not very important, which is probably why the index was dropped during the stepwise regression.

One of the crucial findings from the literature review was that driving forces vary for different time periods of urban growth (Vermeiren *et al.*, 2012). Furthermore, it became clear that most studies that project future urban growth base their statistical analysis of driving forces on a short and recent time period. Since the goal of this thesis is to contribute to the improvement of urban growth models, the first analysis was repeated for a second, much shorter and more recent time period. As with previous studies, in our case the driving forces vary between the two time periods. Maybe even more interesting than the difference in driving forces was the overall difference in model fit and predictive power of the model between the two time periods. As was shown in the results, the models for each continent were relatively good in predicting the urban pattern in 2010, that is urban growth over an extremely long period. Meanwhile using the same explanatory variables models were performing poorly in predicting urban growth over the in comparison short period of twenty years between 1990 and 2010. The most obvious explanation for this difference in explanatory power is the nature of the included driving forces. As mentioned, policy, economic and technical drivers are very much underrepresented in most urban growth studies, as in this study. It has already been suggested in the explanation of urban agglomeration forces, that second nature causes have become more important over the last couple of decades. Urban growth between 1990 and 2010 is therefore way harder to predict, as unmeasurable forces like agglomeration and path dependency are becoming determining forces.

To improve urban growth models, it is essential to better understand more recent urban growth developments. Therefore, more drivers from categories besides natural drivers need to be included in future regression analyzes. One possibility, namely additional travel time variables, has already been mentioned. The governance index has also shown potential and should therefore be improved on and included in future analyzes. Other variables that could be included in follow-up studies to improve the models' predictive power are autologistic specifications to different kinds of land-use besides urban areas. Examples of such variables used in other studies are the proportion of farmland and the proportion forest/nature in surroundings cells (Braimoh and Onishi, 2007). Ideally the effect of surrounding land-use is tested on different distances, Another variable that could improve the current analysis and detect new urban sites is in density of rural population (Li, Sun and Fang, 2018).

The second part of the study was more focussed on the methodological element of the analysis. The goal of the second part of the analysis was to compare different methodological approaches and determine which one is best applied for urban growth studies. Therefore, the results of the stepwise regression have been compared to a benchmark binary logit model and a penalized binary logit model produced with a more sophisticated approach. Results have shown that, with the stepwise regression, a model generally outperforms the benchmark logit model. This means that the first approach with autoGLM indeed finds a parsimonious model including only significant variables with quite an impressive predictive power. For Asia as well as Europe, the penalized regression results were very similar to the stepwise regression results. The penalized regression therefore does not seem to have an advantage over the stepwise regression for our study with the chosen set of variables.

Most studies analyzed during the literature review used the ROC value to compare model performance (Braimoh and Onishi, 2007; Poelmans and Van Rompaey, 2009; Vermeiren *et al*., 2012; Li, Sun and Fang, 2018). Our study suggests that this is not always the most informative and transparent evaluation method, especially in the case of large class imbalances, as was the case for our data. If the class of interest is in the minority class, the area under the precision-recall curve (AUC) gives a more accurate picture of model performance. This should always be considered when comparing results of different studies. Because of the large class imbalance between urban and non-urban cells, we tested if the model's performance would improve when the majority class of non-urban cell was downsampled. Downsampling did not improve the overall performance of the models. We have seen that downsampling causes the overall precision to decrease in exchange for an increased recall value. This is an interesting finding as it signals that downsampling could be a good tool if the main expectation of a model is to archive a high recall value.

Once again, the literature study showed that in situations where the relationship between variables is very complex, non-linear models such as the random forest model achieve better results than linear models (Couronné, Probst and Boulesteix, 2018; Kirasich, Smith and Sadler, 2018). Results have shown that indeed the random forest model has the highest AUC in both time periods. To calculate the AUC, a cut-off value of 0.5 was used to determine whether a cell would be urban or non-urban. For future studies it is recommended explore different cut-off value as a means to improve the classification models. Random forest models are not often applied in urban growth models. One reason is that it is very difficult to interpret coefficients for the individual drivers. However, if the end goal of a study is to model future growth, and not necessarily to understand historical urban growth, this might be less important or even unnecessary. Another very surprising result is that, based on the log-loss function, the random forest model performs the worst of the three model. Based on the log-loss function the logit and penalized logit models with balanced classes actually perform best.

Discussing all these different results shows that the analysis offers a great starting point for further research into the optimal calibration of urban growth models. We have seen that it is extremely important for the research to determine what the ultimate goal of the model is when choosing a methodological approach and evaluation method.

6 CONCLUSION

This master thesis aimed to analyze the driving forces determining locational urban growth and their regional differences. This analysis was done with an automated stepwise regression. The main purpose was to find a parsimonious model that describes the relationship between the urban pattern in 2010 and several driving forces. The results have shown that the number and kind of drivers influencing urban growth differ per continent. Furthermore, the same drivers can have a positive as well as a negative influence on urban growth, depending on the continent. The predictive power of the models is also different for each continent, with the most considerable predictive power for Australia. The predictive power of all models decreases dramatically when the urban growth between 1990-2010 is modelled instead of the urban pattern in 2010. This shows that the set of chosen driving forces is only suitable to predict urban growth over a long period and not over a short and, importantly, more recent time period. This should be considered when using the regression results to forecast future urban growth.

The secondary research aim was to further investigate possible methodological approaches to improve the predictive power of the models. Therefore, a penalized regression was used, that, like the stepwise regression, tries to create a parsimonious model in exchange for an acceptable amount of bias. The stepwise and penalized logit regression models were compared to a random forest model to see if the complex relationship between driving forces and urban growth would be better described with a non-linear model. The results from this second part of the analysis have brought us to several important conclusions. First, the results from the penalized regression are nearly identical to the results from the stepwise regression, both of which are an improvement over the regular binary logit model without elimination variables. Second, the metric on which the models are evaluated is essential in deciding which model preforms best, and therefore needs sufficient consideration. We argue and show that the area under the recall precision curve is the best evaluation metric in the case of unbalanced classes. The last major finding is that the random forest model surpasses both logit models.

Taken together, the results suggest that a regional differentiation of driving forces is urgently needed for global urban growth models. Furthermore, more economical and political drivers are needed to understand more recent urban development and eventually be able to predict future urban growth. Finally, the application of non-linear models has shown great potential and should, therefore, be further investigated.

7 References

Aguayo, M. I. *et al.* (2007) 'Revealing the driving forces of mid-cities urban growth patterns using spatial modeling: A case study of Los Ángeles, Chile', *Ecology and Society*, 12(1). doi: 10.5751/ES-01970-120113.

Alonso, W. (1964) *Location and Land Use*. Cambridge (MA): Harvard University Press. doi: https://doi.org/10.4159/harvard.9780674730854.

Andree, B. and Koomen, E. (2017) *Calibration of the 2UP model*.

Angel, S. *et al.* (2011) 'The dimensions of global urban expansion: Estimates and projections for all countries, 2000-2050', *Progress in Planning*, 75(2), pp. 53–107. doi: 10.1016/j.progress.2011.04.001.

Angel, S. and Blei, A. M. (2016) 'The spatial structure of American cities: The great majority of workplaces are no longer in CBDs, employment sub-centers, or live-work communities', *Cities*, 51, pp. 21–35. doi: https://doi.org/10.1016/j.cities.2015.11.031.

Anselin, L. (1988) *Spatial econometrics: methods and models*. Dordrecht: Kluwer Academic.

Anselin, L. (2003) 'Spatial Econometrics', pp. 310–330.

Batisani, N. and Yarnal, B. (2009) 'Urban expansion in Centre County, Pennsylvania: Spatial dynamics and landscape transformations', *Applied Geography*. Elsevier Ltd, 29(2), pp. 235–249. doi: 10.1016/j.apgeog.2008.08.007.

Beron, K. J. and Vijverberg, W. P. M. (2004) 'Probit in a Spatial Context: A Monte Carlo Analysis', in Anselin, L., Florax, R. J. G. M., and Rey, S. J. (eds) *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 169–195. doi: 10.1007/978-3-662-05617-2_8.

Besag, J. (1974) 'Spatial Interaction and the Statistical Analysis of Lattice Systems', *Journal of the Royal Statistical Society. Series B (Methodological)*. [Royal Statistical Society, Wiley], 36(2), pp. 192–236. Available at: http://www.jstor.org/stable/2984812.

Biau, G. (2012) 'Analysis of a random forests model', *The Journal of Machine Learning Research*. JMLR. org, 13(1), pp. 1063–1095.

Biau, G. and Scornet, E. (2016) 'A random forest guided tour', *Test*. Springer, 25(2), pp. 197–227.

Bohnet, I. C. and Pert, P. L. (2010) 'Patterns, drivers and impacts of urban growth-A study from Cairns, Queensland, Australia from 1952 to 2031', *Landscape and Urban Planning*. Elsevier B.V., 97(4), pp. 239–248. doi: 10.1016/j.landurbplan.2010.06.007.

Braimoh, A. K. and Onishi, T. (2007) 'Spatial determinants of urban land use change in Lagos, Nigeria', *Land Use Policy*, 24(2), pp. 502–515. doi: 10.1016/j.landusepol.2006.09.001.

Brakman, S. *et al.* (2005) *New Economic Geography, Empirics, and Regional Policy*. CPB Netherlands Bureau for Economic Policy Analysis. doi: ISBN 90-5833-281-7.

Breiman, L. (2001) 'Random forests', *Machine learning*. Springer, 45(1), pp. 5–32.

Bürgi, M., Hersperger, A. M. and Schneeberger, N. (2005) 'Driving forces of landscape change - Current and new directions', *Landscape Ecology*, 19(8), pp. 857–868. doi: 10.1007/s10980-005-0245-3.

Cao, Y. *et al.* (2020) 'Urban spatial growth modeling using logistic regression and cellular automata: A case study of Hangzhou', *Ecological Indicators*. Elsevier, 113(December 2019), p. 106200. doi: 10.1016/j.ecolind.2020.106200.

Cheng, J. and Masser, I. (2003) 'Urban growth pattern modeling: a case study of Wuhan city, PR China', *Landscape and urban planning*. Elsevier, 62(4), pp. 199–217.

Ciccone, A. and Hall, R. E. (1996) 'Productivity and the Density of Economic Activity', *The American Economic Review*. American Economic Association, 86(1), pp. 54–70. Available at: http://www.jstor.org/stable/2118255.

Couronné, R., Probst, P. and Boulesteix, A.-L. (2018) 'Random forest versus logistic regression: a large-scale benchmark experiment', *BMC Bioinformatics*, 19(1), p. 270. doi: 10.1186/s12859-018-2264-5.

Dendoncker, N., Bogaert, P. and Rounsevell, M. (2006) 'A statistical method to downscale aggregated land use data and scenarios', *Journal of Land Use Science*, 1(2–4), pp. 63–82. doi: 10.1080/17474230601058302.

Dieleman, F. M., Dijst, M. J. and Spit, T. (1999) ‘Planning the compact city: The randstad Holland experience’, *European Planning Studies*. Routledge, 7(5), pp. 605–621. doi: 10.1080/09654319908720541.

FAO/IIASA/ISRIC/ISS-CAS/JRC (2009) *Harmonized World Soil Database (version 1.1)*. Rome, Italy and IIASA, Laxenburg, Austria.

Fujita, M., Krugman, P. and Mori, T. (1999) ‘On the evolution of hierarchical urban systems’, *European Economic Review*. Elsevier, 43(2), pp. 209–251.

Glaeser, E. L. *et al.* (1992) ‘Growth in Cities’, *Journal of Political Economy*, 100(6), pp. 1126–1152. doi: 10.1086/261856.

Glaeser, E. L. and Kahn, M. E. (2003) ‘Sprawl and Urban Growth’, *National Bureau of Economic Research Working Paper Series*, No. 9733. doi: 10.3386/w9733.

Granger, C. W. J., King, M. L. and White, H. (1995) ‘Comments on testing economic theories and the use of model selection criteria’, *Journal of Econometrics*, 67(1), pp. 173–187. doi: https://doi.org/10.1016/0304-4076(94)01632-A.

Grimm, N. B. *et al.* (2008) ‘Global change and the ecology of cities’, *Science*, 319(5864), pp. 756–760. doi: 10.1126/science.1150195.

Güneralp, B., Güneralp, İ. and Liu, Y. (2015) ‘Changing global patterns of urban exposure to flood and drought hazards’, *Global environmental change*. Elsevier, 31, pp. 217–225.

Hersperger, A. and Bürgi, M. (2007) ‘Driving Forces of Landscape Change in The Urbanizing Limmat Valley, Switzerland’, in, pp. 45–60. doi: 10.1007/978-1-4020-5648-2_3.

Hersperger, A. M. *et al.* (2018) ‘Urban land-use change: The role of strategic spatial planning’, *Global Environmental Change*, 51(June 2017), pp. 32–42. doi: 10.1016/j.gloenvcha.2018.05.001.

Hietel, E., Waldhardt, R. and Otte, A. (2005) ‘Linking socio-economic factors, environment and land cover in the German Highlands, 1945-1999.’, *Journal of environmental management*, 75(2), pp. 133–143. doi: 10.1016/j.jenvman.2004.11.022.

Huijstee, J. *et al.* (2018) *Towards an Urban Preview: Modelling future urban growth with 2UP*. The Hague.

Kaufmann, D., Kraay, A. and Zoido, P. (1999) ‘Governance matters’, *World Bank policy research working paper*, (2196).

Kirasich, K., Smith, T. and Sadler, B. (2018) ‘Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets’, *SMU Data Science Review*, 1(3), p. 9.

Koomen, E. *et al.* (2015) ‘A utility-based suitability framework for integrated local-scale land-use modelling’, *Computers, Environment and Urban Systems*. Elsevier Ltd, 50, pp. 1–14. doi: 10.1016/j.compenvurbsys.2014.10.002.

Krugman, P. (1991) ‘Increasing returns and economic geography’, *Journal of Political Economy*, 99(3), pp. 483–499. doi: 10.1086/261763.

Krugman, P. R. (1991) *Geography and trade*. MIT press.

Kuhn, M. (2008) ‘Building predictive models in R using the caret package’, *Journal of Statistical Software*, 28(5), pp. 1–26. doi: 10.18637/jss.v028.i05.

Li, G., Sun, S. and Fang, C. (2018) ‘The varying driving forces of urban expansion in China: Insights from a spatial-temporal analysis’, *Landscape and Urban Planning*, 174(February), pp. 63–77. doi: 10.1016/j.landurbplan.2018.03.004.

Li, X. *et al.* (2017) ‘A New Global Land-Use and Land-Cover Change Product at a 1-km Resolution for 2010 to 2100 Based on Human–Environment Interactions’, *Annals of the American Association of Geographers*, pp. 1–20. doi: 10.1080/24694452.2017.1303357.

Li, X., Zhou, W. and Ouyang, Z. (2013) ‘Forty years of urban expansion in Beijing: What is the relative importance of physical, socioeconomic, and neighborhood factors?’, *Applied Geography*. Elsevier Ltd, 38(1), pp. 1–10. doi: 10.1016/j.apgeog.2012.11.004.

Liu, J., Zhan, J. and Deng, X. (2005) ‘Spatio-temporal patterns and driving forces of urban land expansion in China during the economic reform era’, *Ambio*, 34(6), pp. 450–455. doi: 10.1579/0044-7447-34.6.450.

Marshall, A. (1920) *Principles of Economics*. London: Palgrave Macmillan UK.

McMillen, D. P. (1992) ‘Probit with spatial autocorrelation’, *Journal of Regional Science*. Wiley Online Library, 32(3), pp. 335–348.

McMillen, D. P. (1995) ‘Selection bias in spatial econometric models’, *Journal of Regional Science*. Wiley Online Library, 35(3), pp. 417–436.

Melo, P. C., Graham, D. J. and Noland, R. B. (2009) ‘A meta-analysis of estimates of urban agglomeration economies’, *Regional Science and Urban Economics*, 39(3), pp. 332–342. doi: https://doi.org/10.1016/j.regsciurbeco.2008.12.002.

Müller, K., Steinmeier, C. and Küchler, M. (2010) ‘Urban growth along motorways in Switzerland’, *Landscape and urban Planning*. Elsevier, 98(1), pp. 3–12.

Mundia, C. N. and Aniya, M. (2005) ‘Analysis of land use/cover changes and urban expansion of Nairobi city using remote sensing and GIS’, *International Journal of Remote Sensing*, 26(13), pp. 2831–2849. doi: 10.1080/01431160500117865.

Nachtergaele, F. *et al.* (2010) ‘The harmonized world soil database’, in *Proceedings of the 19th World Congress of Soil Science, Soil Solutions for a Changing World, Brisbane, Australia, 1-6 August 2010*, pp. 34–37.

Neumann, B. *et al.* (2015) ‘Future coastal population growth and exposure to sea-level rise and coastal flooding - A global assessment’, *PLoS ONE*, 10(3). doi: 10.1371/journal.pone.0118571.

Nieves, J. J. *et al.* (2020) ‘Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night’, *Computers, Environment and Urban Systems*. Elsevier, 80(101444), pp. 1–14. doi: 10.1016/j.compenvurbsys.2019.101444.

Nussbaumer, S. *et al.* (2014) ‘Risk estimation for future glacier lake outburst floods based on local land-use changes’, *Natural Hazards and Earth System Sciences*. Copernicus Publications, 14(6), pp. 1611–1624.

Overmars, K. P., De Koning, G. H. J. and Veldkamp, A. (2003) ‘Spatial autocorrelation in multi-scale land use models’, *Ecological Modelling*, 164(2–3), pp. 257–270. doi: 10.1016/S0304-3800(03)00070-X.

Overmars, K. P., Verburg, P. H. and Veldkamp, T. (A. . (2007) ‘Comparison of a deductive and an inductive approach to specify land suitability in a spatially explicit land use model’, *Land Use Policy*, 24(3), pp. 584–599. doi: https://doi.org/10.1016/j.landusepol.2005.09.008.

Paelinck, J. H. P. and Klaassen, L. H. (1979) *Spatial Econometrics*. Gower (Studies in spatial analysis). Available at: https://books.google.nl/books?id=xdvtQwAACAAJ.

Park, M. Y. and Hastie, T. (2008) ‘Penalized logistic regression for detecting gene interactions’, *Biostatistics*, 9(1), pp. 30–50. doi: 10.1093/biostatistics/kxm010.

Poelmans, L. and Van Rompaey, A. (2009) ‘Detecting and modelling spatial patterns of urban sprawl in highly fragmented areas: A case study in the Flanders-Brussels region’, *Landscape and Urban Planning*, 93, pp. 10–19. doi: 10.1016/j.landurbplan.2009.05.018.

Poelmans, L. and Van Rompaey, A. (2010) ‘Complexity and performance of urban expansion models’, *Computers, Environment and Urban Systems*. Elsevier Ltd, 34(1), pp. 17–27. doi: 10.1016/j.compenvurbsys.2009.06.001.

Pontius, R. G., Cornell, J. D. and Hall, C. A. S. (2001) ‘Modeling the spatial pattern of land-use change with GEOMOD2: Application and validation for Costa Rica’, *Agriculture, Ecosystems and Environment*, 85(1–3), pp. 191–203. doi: 10.1016/S0167-8809(01)00183-9.

Porter, M. E. (1990) ‘The competitive advantage of nations’, *Harvard business review*. Cambridge, Massachusetts, 68(2), pp. 73–93.

Promper, C. *et al.* (2014) ‘Analysis of land cover changes in the past and the future as contribution to landslide risk scenarios’, *Applied Geography*. Elsevier, 53, pp. 11–19.

Reilly, M. K., O’Mara, M. P. and Seto, K. C. (2009) ‘From Bangalore to the Bay Area: Comparing transportation and activity accessibility as drivers of urban growth’, *Landscape and Urban Planning*, 92(1), pp. 24–33. doi: 10.1016/j.landurbplan.2009.02.001.

Riley, S., Degloria, S. and Elliot, S. D. (1999) ‘A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity’, *Internation Journal of Science*, 5, pp. 23–27.

Rosenthal, S. and Strange, W. (2004) ‘Evidence on the nature and sources of agglomeration economies’, in Henderson, J. V and Thisse, J. F. (eds). Elsevier, pp. 2119-2171 BT-Handbook of Regional and Urban Eco.

Available at: https://econpapers.repec.org/RePEc:eee:regchp:4-49.

Scornet, E. (2017) 'Tuning parameters in random forests', *ESAIM: Proceedings and Surveys*. EDP Sciences, 60, pp. 144–162.

Seto, K. *et al*. (2011) 'A Meta-Analysis of Global Urban Land Expansion', *PLOS ONE*. Public Library of Science, 6(8), p. e23777. Available at: https://doi.org/10.1371/journal.pone.0023777.

Seto, K. C. (2011) 'Exploring the dynamics of migration to mega-delta cities in Asia and Africa: Contemporary drivers and future scenarios', *Global Environmental Change*. Elsevier Ltd, 21(SUPPL. 1), pp. S94–S107. doi: 10.1016/j.gloenvcha.2011.08.005.

Seto, K.C., Güneralp, B. and Hutyra, L. R. (2012) 'Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools', *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16083–16088. doi: 10.1073/pnas.1211658109.

Seto, Karen C, Güneralp, B. and Hutyra, L. R. (2012) 'Supporting Information', *Proceedings of the National Academy of Sciences of the United States of America*, pp. 1–7.

Sin, C.-Y. and White, H. (1996) 'Information criteria for selecting possibly misspecified parametric models', *Journal of Econometrics*, 71(1), pp. 207–225. doi: https://doi.org/10.1016/0304-4076(94)01701-8.

Smith, T. E. and LeSage, J. P. (2004) 'A Bayesian probit model with spatial dependencies', *Advances in econometrics*. Elsevier Oxford, UK, 18(18), pp. 127–160.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*. Wiley Online Library, 58(1), pp. 267–288.

United Nations (2018) *World Urbanization Prospects 2018: Highlights*. Available at: https://population.un.org/wup/.

United Nations (2019) *World Population Prospects 2019: Highlights*. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12283219.

De Vasconcelos, M. J. P. *et al*. (2001) 'Spatial prediction of fire ignition probabilities: Comparing logistic regression and neural networks', *Photogrammetric Engineering and Remote Sensing*, 67(1), pp. 73–81.

Veldkamp, A. and Lambin, E. F. (2001) 'Editorial: Predicting land-use change', *Agriculture, Ecosystems and Environment*, 85(1–3), pp. 1–6. doi: 10.1016/S0167-8809(01)00199-2.

Verburg, P. H., Ritsema van Eck, J. R., *et al*. (2004) 'Determinants of land-use change patterns in the Netherlands', *Environment and Planning B: Planning and Design*, 31(1), pp. 125–150. doi: 10.1068/b307.

Verburg, P. H., Schot, P. P., *et al*. (2004) 'Land use change modelling: Current practice and research priorities', *GeoJournal*, 61(4), pp. 309–324. doi: 10.1007/s10708-004-4946-y.

Vermeiren, K. *et al*. (2012) 'Urban growth of Kampala, Uganda: Pattern analysis and scenario development', *Landscape and Urban Planning*. Elsevier B.V., 106(2), pp. 199–206. doi: 10.1016/j.landurbplan.2012.03.006.

van Vliet, J. *et al*. (2013) 'Measuring the neighbourhood effect to calibrate land use models', *Computers, Environment and Urban Systems*, 41, pp. 55–64. doi: https://doi.org/10.1016/j.compenvurbsys.2013.03.006.

Wolff, M., Haase, D. and Haase, A. (2018) 'Compact or spread? A quantitative spatial model of urban areas in Europe since 1990', *PLoS ONE*, 13(2), pp. 1–22. doi: 10.1371/journal.pone.0192326.

Wright, M. N. and Ziegler, A. (2015) 'ranger: A fast implementation of random forests for high dimensional data in C++ and R', *arXiv preprint arXiv:1508.04409*.

Wright, M. N. and Ziegler, A. (2017) 'Ranger: A fast implementation of random forests for high dimensional data in C++ and R', *Journal of Statistical Software*, 77(1), pp. 1–17. doi: 10.18637/jss.v077.i01.

Zeng, Y. *et al*. (2008) 'Modeling spatial land use pattern using autologistic regression', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 37.

Zou, H. (2006) 'The adaptive lasso and its oracle properties', *Journal of the American statistical association*. Taylor & Francis, 101(476), pp. 1418–1429.

8 APPENDIX

A List of variables

Category	Variable	Description	Source
Nature/ Spatial configuration	Slope	Slope in degrees.	GTOPO30 (USGS, 1996[6])
	Elevation	Elevation in meters.	SRTM V3 (Jarvis et al., 2006[7]) and GTOPO30 (USGS, 1996)
	Terrain roughness index	Index for the overall ruggedness of a grid cell, based on slope and elevation.	Calculated based on method by Riley, Degloria and Elliot (1999) with Slope and Elevation datasets mentioned above.
	Soil type	Type of soil. Overview of the different soil types in appendix 0.	Harmonized World Soil Database (HWSD)
	Urban Area Density	Presence of urban land-use in the surrounding grid cells in a radius from 0 to 10 kilometers.	Global Human Settlement layer (GHSL)[8]
	Coastal Urban Area Density	Density of urban area weighted for its distance to both the central cell and the coastline (weights turn 0 after 20km).	Global Human Settlement layer (GHSL)
	Landslides (Earthquakes)	Presence of landslides triggered by earthquakes.	Global Risk Data Platform (https://preview.grid.unep.ch/index.php?preview=data&events=landslides&evcat=1&lang=eng)
	Landslides (Precipitation)	Presence of landslides triggered by precipitations.	Global Risk Data Platform (https://preview.grid.unep.ch/index.php?preview=data&events=landslides&evcat=2&lang=eng)
	Earthquakes	Earthquakes Modified Mercalli Intensity	Global Risk Data Platform (https://preview.grid.unep.ch/index.php?preview=data&events=earthquakes&evcat=3&lang=eng)
	Flood Prone Areas	Areas with a risk of a flood once in 100 years.	GLOFRIS/inun_oecd_2010_RP100_area
	Distance to rivers	Distance from each grid cell to nearest river.	hydroRivers https://www.hydrosheds.org/page/hydrorivers
Politics	Protected areas	Binary variables. Cell has the value 1 if the cell contains an area under protection and value 0 otherwise.	World database on protected areas (WDPA). https://www.protectedplanet.net
	Governance Indicator	"Governance consists of the traditions and institutions by which authority in a country is exercised. This includes the process by which governments are selected, monitored and replaced; the capacity of the government to effectively formulate and implement sound policies; and the respect of citizens and the state for the institutions that govern economic and social interactions among them." (Kaufman and Kraay, 1999)	Worldwide governance indicators (https://info.worldbank.org/governance/wgi/)
Economy	Travel Time	Index value of the travel time in minutes to the nearest city center. The index ranges from 0 to 12, 12 representing the exact center of a city center and 12 all grid cells with a travel time that exceeds 60 minutes to the city center.	Based on road data (Global Roads Inventory Project (GRIP) dataset v1) and urban clusters with more than 50.000 people.

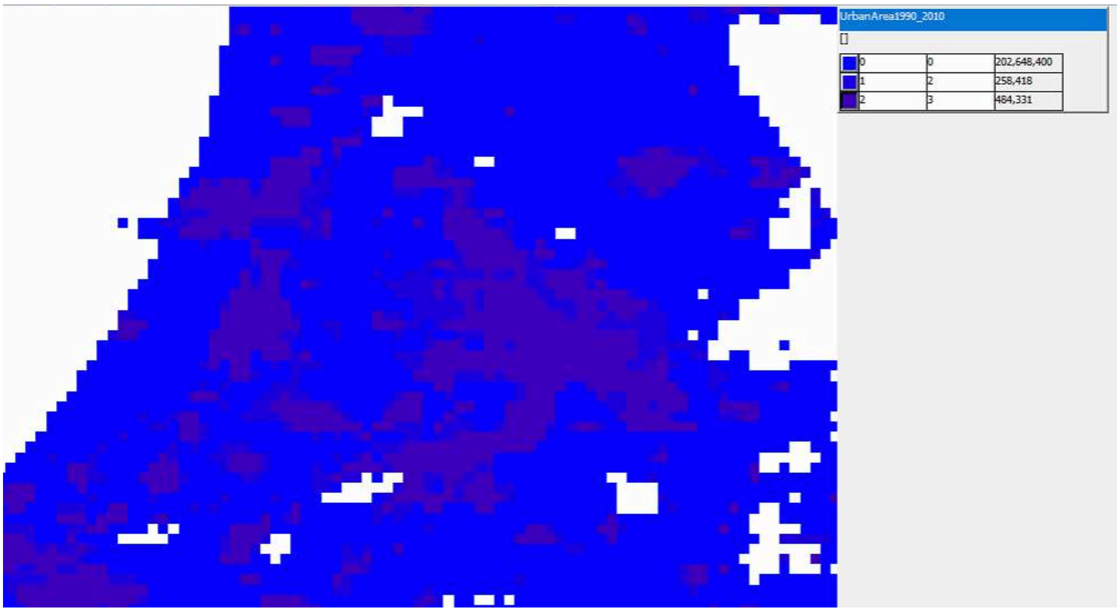
[6] US Geological Survey, 1996. Global Digital Elevation Model (GTOPO30). EROS Data Center Available at https://lta.cr.usgs.gov/GTOPO30.

[7] Jarvis A., Reuter, H.I., Nelson, A., Guevara, E., 2006. Hole-filled SRTM for the globe version 3, from the CGIAR-CSI SRTM 90m database. Available from http://srtm.csi.cgiar.org.

[8] https://ghsl.jrc.ec.europa.eu

B Urban growth 1990 – 2010

Excerpt from GeoDMS showing the urban growth of Amsterdam (The Netherlands) and surroundings between 1990-2010. The darkest blue cells (3) were already urban before 1990. These cells have been excluded from the analysis. The lighter blue cells (2) changed from non-urban to urban between 1990-2010.



UrbanArea1990_2010

□

■	0	0	202,648,400
■	1	2	258,418
■	2	3	484,331

C Correlation Matrices

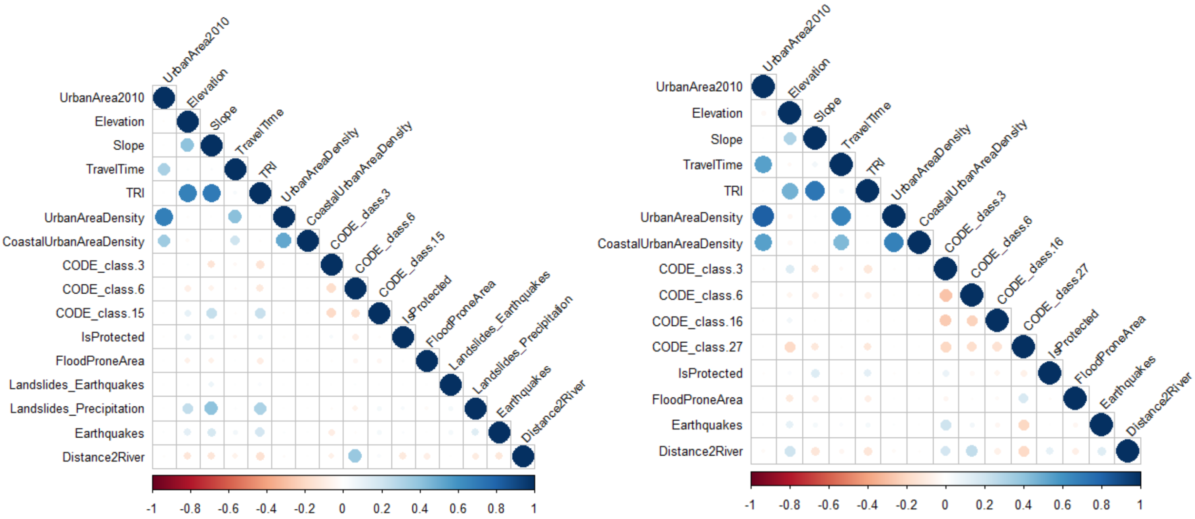


Figure 14 | Correlation matrix Africa (left) and Australia (right) - Urban area 2010

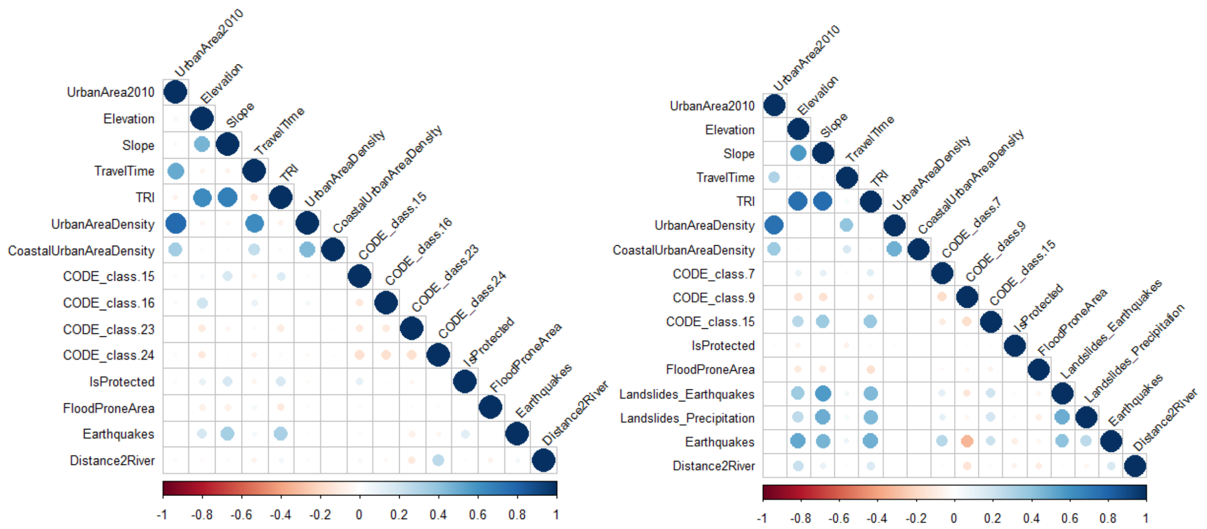


Figure 15 | Correlation matrix North America (left) and South America (right) - Urban area 2010

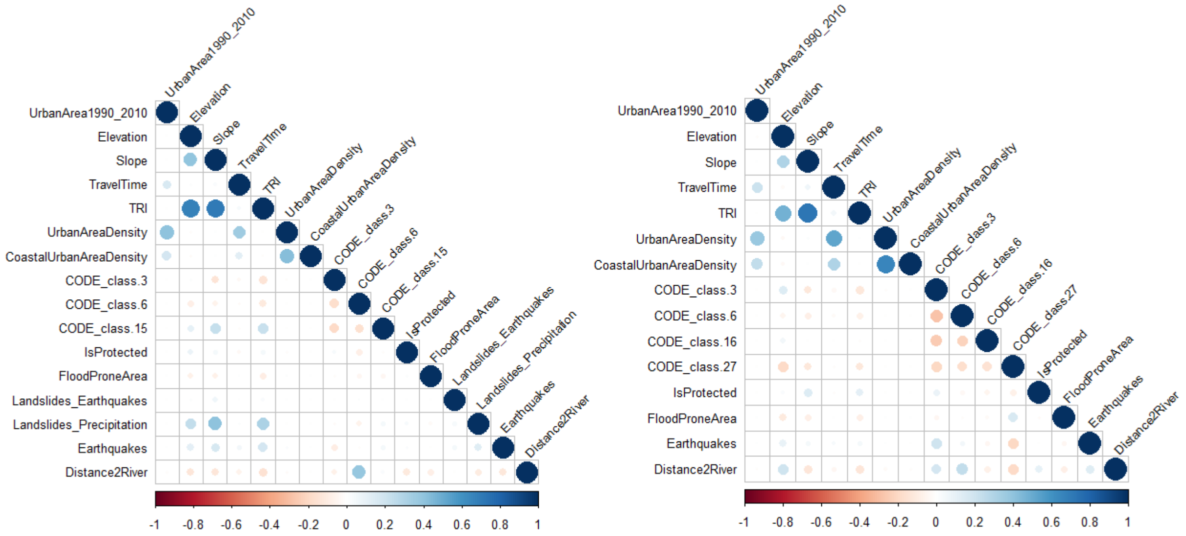


Figure 16 | Correlation matrix Africa (left) and Australia (right) - Urban growth 1990 - 2010

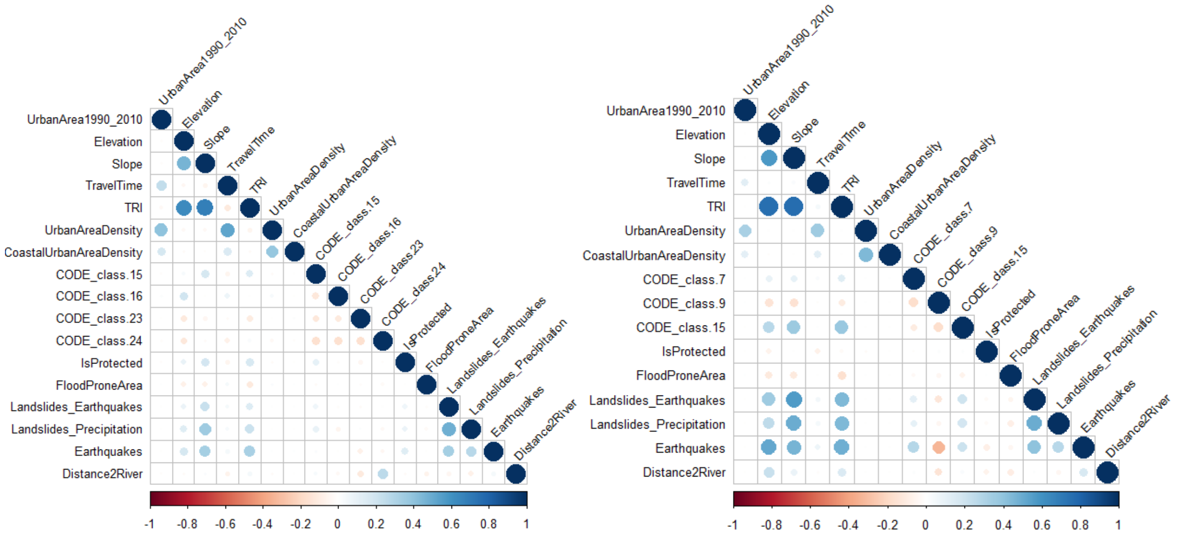
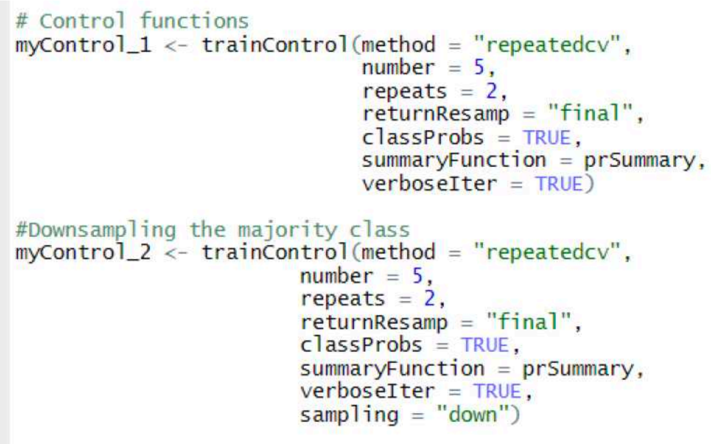


Figure 17 | Correlation matrix North America (left) and South America (right) - Urban growth 1990 - 2010

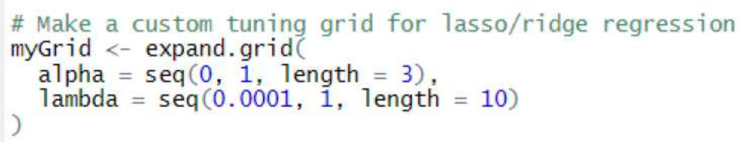
D R code for predictive modelling

Custom control functions



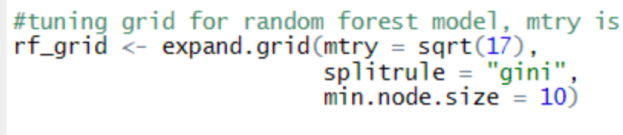
```r
# Control functions
myControl_1 <- trainControl(method = "repeatedcv",
                            number = 5,
                            repeats = 2,
                            returnResamp = "final",
                            classProbs = TRUE,
                            summaryFunction = prSummary,
                            verboseIter = TRUE)

#Downsampling the majority class
myControl_2 <- trainControl(method = "repeatedcv",
                       number = 5,
                       repeats = 2,
                       returnResamp = "final",
                       classProbs = TRUE,
                       summaryFunction = prSummary,
                       verboseIter = TRUE,
                       sampling = "down")
```

Custom tuning grid ElasticNet



```r
# Make a custom tuning grid for lasso/ridge regression
myGrid <- expand.grid(
  alpha = seq(0, 1, length = 3),
  lambda = seq(0.0001, 1, length = 10)
)
```

Custom tuning grid Random Forest



```r
#tuning grid for random forest model, mtry is
rf_grid <- expand.grid(mtry = sqrt(17),
                       splitrule = "gini",
                       min.node.size = 10)
```

E Continent Codes: GeoDMS Legend

Total number of observations before continental split: 203,391,149 observations of 19 variables

Continent	Code	Observations without NA	Sample Size
Africa	0	36,031,563	0.15
Antarctica	1	-	-
Asia	2	42,765,854	0.15
Australia	3	9,845,739	0.5
Europe	4	10,804,034	0.5
North America	5	37,898,362	0.15
Oceania	6	-	-
South America	7	21,618,023	0.25

Table 8 | Proportion of cells that changed from non-urban to urban between 1990-2010

Continent	Urban 1990-2010	Non-urban	Total
Africa	3,722 (0.07 %)	4,994,288	4,998,010
Asia	6,595 (0.10 %)	6,382,477	6,389,072
Europe	16,901 (0.33 %)	5,129,032	5,145,933

F Legend Soil Database

Label	Code	Description from FAO/IIASA/ISRIC/ISS-CAS/JRC (2009)
'Acrisols'	'1'	Soils with subsurface accumulation of low activity clays and low base saturation
'Alisols'	'2'	Soils with sub-surface accumulation of high activity clays, rich in exchangeable aluminum
'Andosols'	'3'	Young soils formed from volcanic deposits
'Arenosols'	'4'	Sandy soils featuring very weak or no soil development
'Anthrosols'	'5'	Soils in which human activities have resulted in profound modification of their properties
'Chernozems'	'6'	Soils with a thick, dark topsoil, rich in organic matter with a calcareous subsoil
'Calcisols'	'7'	Soils with accumulation of secondary calcium carbonates
'Cambisols'	'8'	Weakly to moderately developed soils
'Fluvisols'	'9'	Young soils in alluvial deposits
'Ferralsols'	'10'	Deep, strongly weathered soils with a chemically poor, but physically stable subsoil
'Gleysols'	'11'	Soils with permanent or temporary wetness near the surface
'Greyzems'	'12'	Acid soils with a thick, dark topsoil rich in organic matter
'Gypsisols'	'13'	Soils with accumulation of secondary gypsum
'Histosols'	'14'	Soils which are composed of organic materials
'Kastanozems'	'15'	Soils with a thick, dark brown topsoil, rich in organic matter and a calcareous or gypsum-rich subsoil Very shallow soils over hard rock or in unconsolidated very gravelly material
'Leptosols'	'16'	Very shallow soils over hard rock or in unconsolidated very gravelly material
'Luvisols'	'17'	Soils with subsurface accumulation of high activity clays and high base saturation
'Lixisols'	'18'	Soils with subsurface accumulation of low activity clays and high base saturation
'Nitisols'	'19'	Deep, dark red, brown or yellow clayey soils having a pronounced shiny, nut-shaped structure
'Podzoluvisols'	'20'	Acid soils with a bleached horizon penetrating into a clay-rich subsurface horizon
'Phaeozems'	'21'	Soils with a thick, dark topsoil rich in organic matter and evidence of removal of carbonates
'Planosols'	'22'	Soils with a bleached, temporarily water-saturated topsoil on a slowly permeable subsoil
'Plinthosols'	'23'	Wet soils with an irreversibly hardening mixture of iron, clay and quartz in the subsoil
'Podzols'	'24'	Acid soils with a subsurface accumulation of iron-aluminum-organic compounds
'Regosols'	'25'	Soils with very limited soil development
'Solonchaks'	'26'	Strongly saline soils
'Solonetz'	'27'	Soils with subsurface clay accumulation, rich in sodium
'Vertisols'	'28'	Dark-coloured cracking and swelling clays
'Rock Outcrop'	'29'	Removed from the analysis
'Sand Dunes'	'30'	
'Water Bodies'	'31'	
'Urban, mining, etc.'	'32'	
'Salt Flats'	'33'	

G Additional Results

Table 9 | Regression result GLM (logit) with caret for Europe (Urban Area 2010)

Generalized Linear Model (Logit)					
Europe; Urban Area 2010					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.99	0.02	-357.3	< 2e-16	***
Elevation	0.37	0.02	18.0	< 2e-16	***
Slope	0.35	0.02	17.7	< 2e-16	***
TravelTime	0.13	0.01	25.7	< 2e-16	***
TRI	-1.79	0.03	-56.6	< 2e-16	***
UrbanAreaDensity	0.89	0.004	202.5	< 2e-16	***
CoastalUrbanAreaDensity	0.06	0.003	20.1	< 2e-16	***
CODE_class.7	0.16	0.01	18.7	< 2e-16	***
CODE_class.15	0.05	0.01	4.4	1.13E-05	***
CODE_class.16	0.03	0.01	4.1	3.58E-05	***
CODE_class.23	-0.03	0.01	-2.5	0.014259	*
IsProtected	-0.27	0.01	-22.3	< 2e-16	***
FloodProneArea	0.02	0.01	3.4	0.000655	***
Earthquakes	0.18	0.01	23.7	< 2e-16	***
Distance2River	-0.07	0.01	-8.0	1.61E-15	***
Gov_Factor	0.14	0.01	15.5	< 2e-16	***
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1					
AIC: 177596 Null deviance: 402405 on 2603227 degrees of freedom Residual deviance: 177564 on 2603212 degrees of freedom					

Table 10 | Regression result GLM (logit) with caret for Europe (Urban Growth 1990-2010)

Generalized Linear Model (Logit)					
Europe; Urban growth 1990-2010					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.02	0.03	-253.8	< 2e-16	***
Elevation	0.01	0.04	0.2	0.810015	
Slope	0.21	0.04	5.6	2.42E-08	***
TravelTime	0.36	0.01	49.5	< 2e-16	***
TRI	-1.40	0.05	-27.4	< 2e-16	***
UrbanAreaDensity	0.27	0.003	84.9	< 2e-16	***
CoastalUrbanAreaDensity	0.02	0.002	9.7	< 2e-16	***
CODE_class.7	0.10	0.01	7.6	3.75E-14	***
CODE_class.15	0.07	0.02	4.0	5.87E-05	***
CODE_class.16	-0.02	0.01	-1.9	0.051531	.
CODE_class.23	-0.03	0.02	-1.5	0.122548	
IsProtected	-0.25	0.02	-13.4	< 2e-16	***
FloodProneArea	0.07	0.01	10.5	< 2e-16	***
Landslides_Earthquakes	-0.21	0.06	-3.7	0.000209	***
Landslides_Precipitation	0.05	0.02	2.7	0.00786	**
Earthquakes	0.17	0.01	14.2	< 2e-16	***
Distance2River	-0.08	0.02	-5.1	3.37E-07	***
Gov_Factor	0.22	0.02	13.7	< 2e-16	***
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1					
AIC: 81071 Null deviance: 113529 on 2572966 degrees of freedom Residual deviance: 81035 on 2572949 degrees of freedom					

Table 11 | Regression result GLM (logit) with caret for Asia (Urban Growth 1990-2010)

Generalized Linear Model (Logit)					
Asia; Urban growth 1990-2010					
	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-8.87	0.07	-134.63	< 2e-16	***
Elevation	-0.71	0.11	-6.51	7.53E-11	***
Slope	0.03	0.09	0.36	0.71889	
TravelTime	0.22	0.00	46.08	< 2e-16	***
TRI	-1.55	0.08	-19.23	< 2e-16	***
UrbanAreaDensity	0.18	0.00	69.35	< 2e-16	***
CoastalUrbanAreaDensity	-0.01	0.00	-3.30	0.00096	***
CODE_class.10	-0.28	0.03	-9.93	< 2e-16	***
CODE_class.15	0.04	0.04	1.07	0.28601	
IsProtected	-0.47	0.05	-9.38	< 2e-16	***
FloodProneArea	0.03	0.01	2.16	0.03114	*
Landslides_Earthquakes	-0.13	0.04	-3.23	0.00126	**
Landslides_Precipitation	0.07	0.03	2.08	0.03769	*
Earthquakes	0.19	0.02	9.07	< 2e-16	***
Distance2River	-0.56	0.05	-11.42	< 2e-16	***
Gov_Factor	-0.20	0.02	-12.70	< 2e-16	***
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1					
AIC: 34427					
Null deviance: 51946 on 3194536 degrees of freedom					
Residual deviance: 34395 on 3194521 degrees of freedom					

Table 12 | Confusion matrix for the models explaining urban growth in Asia between 1990-2010

term	**glm**	**glm_down**	**Pen. Glm. (alpha = 0, lambda = 0.1112)**	**Pen. Glm. (a=0.5, l=0.0001**	**rf**	**rf down**
accuracy	0.999	0.972	0.999	0.966	0.999	0.956
kappa	0.140	0.058	0.007	0.048	0.104	0.040
sensitivity	0.088	0.874	0.004	0.883	0.057	0.942
specificity	1	0.972	1	0.9657	1	0.9560
pos_pred_value	0.358	0.031	0.500	0.026	0.551	0.022
neg_pred_value	0.999	1	0.999	1	0.999	1
precision	0.358	0.031	0.5	0.026	0.551	0.022
recall	0.088	0.874	0.004	0.883	0.057	0.942
f1	0.141	0.060	0.007	0.050	0.104	0.042
prevalence	0.001	0.001	0.001	0.001	0.001	0.001
detection_rate	0.000	0.0009	0.000	0.001	0.0001	0.001
detection_prevalence	0.000	0.029	0.000	0.035	0.0001	0.045
balanced_accuracy	0.544	0.923	0.502	0.924	0.529	0.949

Table 13 | Confusion matrix for the models explaining urban growth in Europe between 1990-2010.

term	glm	glm_down	Pen. Glm.(alpha = 0, lambda = 0.111)	Pen. Glm. (alpha = 1, lambda = 0.0001)	rf	rf down
accuracy	0.996	0.937	0.997	0.936	0.997	0.904
kappa	0.076	0.074	0.002	0.073	0.040	0.052
sensitivity	0.044	0.831	0.001	0.835	0.021	0.902
specificity	0.9996	0.937	1	0.936	1	0.904
pos_pred_value	0.280	0.042	0.643	0.041	0.648	0.030
neg_pred_value	0.9969	1	0.997	1	0.997	1
precision	0.280	0.042	0.643	0.041	0.648	0.030
recall	0.044	0.831	0.001	0.835	0.021	0.902
f1	0.077	0.080	0.002	0.079	0.040	0.058
prevalence	0.003	0.003	0.003	0.003	0.003	0.003
detection_rate	0.0001	0.0027	0.000	0.003	0.0001	0.003
detection_prevalence	0.0005	0.065	0.000	0.066	0.0001	0.099
balanced_accuracy	0.522	0.884	0.501	0.885	0.510	0.903