Revised calibration of the 2UP model; analysing change and regional variation

Ferdinand, Pendula; Andree, Bo Pieter Johannes; Koomen, Eric

2021

document version
Publisher's PDF, also known as Version of record

Revised calibration of the 2UP model; analysing change and regional variation

Pendula Ferdinand
Bo Pieter Johannes Andrée
Eric Koomen

December 20, 2021

SPIN LAB SPATIAL INFORMATION LABORATORY

VU VRIJE UNIVERSITEIT AMSTERDAM

COLOPHON

TITLE

Revised calibration of the 2UP model; analysing change and regional variation
Spinlab Research Memorandum SL-19

AUTHORS

Pendula Anne Ferdinand, Bo Pieter Johannes Andrée, Eric Koomen
Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam.

CONTACT

Vrije Universiteit Amsterdam

Faculty of Economics and Business Administration
Department of Spatial Economics/ Spatial Information Laboratory (SPINlab)
De Boelelaan 1105
1081 HV Amsterdam
Netherlands
Phone: +31 20 5986095
Email: e.koomen@vu.nl
Website: https://spinlab.vu.nl/

This study was funded by the JPI-Urban Europe project SIMETRI and prepared for PBL Netherlands Environmental Assessment Agency, Den Haag.

Contents

1 Introduction

The 2UP model was developed by PBL Netherlands Environmental Assessment Agency for spatially explicit simulation of the future growth of cities and population at a global scale (van Huijstee et al., 2018). The model describes urban land use and population at a fine 30″ spatial resolution equivalent to approximately 1x1km near the equator. The model combines scenario-based projections of urban area and population development with local suitability constraints to generate future urban area and population grids. The local suitability for urban development is dynamically generated according to a parameterized suitability model. An initial version of the suitability model was calibrated using logistic regression analysis explaining the global 2014 urban land-use patterns with a limited set of spatially explicit data sets (Andrée and Koomen, 2017). As a follow-up, the current report documents several efforts to improve that initial calibration.

First, the statistical analysis is now performed separately for different continents to allow for a more heterogeneous response of the spatially explicit drivers of urban development. Second, the set of variables capturing the most important drivers of urban spatial expansion has been supplemented to include distance to freshwater (rivers and lakes, potentially important in dry climates), distance to different types of roads, presence of protected natural areas, and exposure to various natural hazards. Third, we have explored more sophisticated regression techniques, including down-sampling, to better deal with the unbalanced data, including a very small amount of urban relative to non-urban observations. Finally, we now also explain where urban land use developed between 1990 and 2014. The latter analysis can be used to explicitly specify suitability for new urban development based on recent growth patterns only. In this report, we reflect on the applicability of this approach as compared to specifying suitability based on analysing urban patterns in a single year, dubbed as the static approach in this report.

An extensive discussion of the rationale and performance of these improvements is provided in the Master thesis in which they were first tested (Ferdinand, 2020). This short report documents the results of the revised calibration analysis that are used to specify the suitability for urban development in the adapted version of the 2UP model. The performance of the revised settings (validation) will be discussed in the scientific paper that VU is preparing in cooperation with PBL.

2 Methods

2.1 Data pre-processing

2.1.1 Continent-specific calibration and dependent variables

The analysis was carried out for each continent separately to improve the calibration and account for continent-specific urban growth patterns. The primary calibration dataset, containing the dependent variables (urban area in 2014 and newly developed urban between 1990 and 2014) and independent variables capturing the essential spatial driving forces, was split into seven continent-specific datasets. Data for Australia and Oceania have been combined because Oceania alone had too few urban cells to fit and validate a proper model.

Table 1 shows the total number of observations (non-urban and urban cells) for each continent for the two analyses after NoData values have been dropped. Note that the cells that became urban between 1990 and 2014 overlap with all cells that were urban in 2014. This overlap ranges between 20% (for Australia and Oceania) and 45% (for Africa). This implies that these separate regressions are also likely to show similar results. For Asia, the number of observations was much higher than for the other continents. To make the statistical calibration for Asia computationally feasible, the smart sampling strategy from the autoGLM package described in the previous calibration report (Andrée and Koomen, 2017)[1] was used to reduce the number of observations for Asia by 50 percent.[2]

Table 1. Number of observations and sample sizes per continent for the analyses explaining urban area in 2014 and urban expansion between 1990-2014 (with the percent overlap with all urban cells in 2014).

Continent	Sample share	Static analysis		Dynamic analysis	
		Valid observations	Urban cells in sample	Valid observations	Urban cells in sample (% of cells from static analysis)
Africa	1	36,325,023	56,545	36,293,574	25,096 (45%)
Asia	0.5	42,306,983	115,002	42,231,171	40,633 (37%)
Australia & Oceania	1	10,869,627	13,369	10,858,720	2,462 (20%)
Europe	1	11,026,506	156,030	10,902,396	31,920 (22%)
North America	1	40,285,937	177,685	40,154,202	45,950 (27%)
South America	1	21,634,881	32,277	21,608,942	6,338 (21%)

The main dataset has 206,825,451 observations (including NoData cells)
The dynamic analysis includes those urban cells that were newly developed between 1999 and 2014. Incidental occurrences of loss of urban area are assumed to result from classification issues and discarded.

[1] The package is available from GitHub (https://github.com/BPJandree/AutoGLM)
[2] This sampling strategy is designed to reduce sample size while keeping the statistical properties of the sample data similar to the full data set, which results in almost identical model parameter estimates as one would obtain when processing the entire data set.

2.1.2 Explanatory variables

Compared to the earlier version of the statistical calibration, several explanatory variables have been added. The newly added variables relate to distances to relevant spatial features (coastline, freshwater bodies, main roads and secondary roads), the presence of specific natural hazards (river floods, earthquakes and landslides) and spatial planning (protected areas). The general characteristics of the explanatory variables are listed in Table 2 and a more extensive description with sources is included in Appendix 1. Two predictive variables included in the previous calibration, urban area density and coastal urban area density, have been omitted due to their endogenous nature. Travel time is now included as individual minutes (up to 1 hour) and not any longer grouped in a limited number of classes. Elevation, slope and terrain roughness are described as in the previous calibration. Soil type was also considered as an explanatory variable and included as a series of dummy variables linking to specific soil types, but these were excluded from the final analysis because their contribution to explaining urban development was limited (see Appendix 1). In total, 13 drivers were used to describe the presence and growth of urban area.

Table 2. Descriptive statistics of the dependent and explanatory variables.

Variable name (unit)	**Africa**		**Asia**		**Australia & Oceania**		**Europe**		**North America**		**South America**	
	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Urban area 2014*	0	1	0	1	0	1	0	1	0	1	0	1
Urban expansion 1990-2014	0	1	0	1	0	1	0	1	0	1	0	1
Ln neighbourhood population density 1990**	-2.3	12.7	-2.3	13.1	-2.3	11.1	-2	11.6	-2.3	11.5	-2.3	11.1
Distance to coast (km, cut off at 250)	0	250	0	250	0	250	0	250	0	250	0	250
Distance to fresh water body (km, cut off at 250)	0	250	0	250	0	250	0	250	0	250	0	250
Distance to main roads (km, cut off at 250)	0	250	0	250	0	250	0	250	0	250	0	250
Distance to secondary roads (km, cut off at 250)	0	250	0	250	0	250	0	250	0	250	0	250
Travel time to city centre (minutes, cut off at 60)	0	60	0	60	0	60	0	60	0	60	0	60
Elevation (meters)	-170	5,825	-406	8,682	0	4,437	-30	4,570	-84	5800	-116	6798
Slope (in degrees)	0	55	0	54	0	59	0	40	0	48	0	58
Terrain roughness index (1-7)	0	7	0	7	0	7	0	7	0	7	1	7
Protected area (0 or 1 =true)	0	1	0	1	0	1	0	1	0	1	0	1
Flood prone area (0 or 1 =true)	0	1	0	1	0	1	0	1	0	1	0	1
Earthquake intensity (0-8 range; 8 = high)	0	6	0	8	0	8	0	8	0	8	0	8
Landslide prone (0 or 1 = true)	0	1	0	1	0	1	0	1	0	1	0	1

* The previous calibration report referred to this urban area layer as 2010 in line with other base data in the 2UP model, but the underlying GHSL data is based on 2014 imagery.
** This represents mean population density in the eight surrounding cells without considering the central cell. A natural logarithm is applied to limit the impact of occasionally very high values, replacing zero values with 0.1 to prevent errors.

Several pre-processing steps were executed for each continent to only retain relevant explanatory variables (see Figure 1). First, the data was split into continent-specific datasets, and subsequently, all rows containing no data values were dropped. In a third step, the bivariate correlation between variables was tested, as exemplified in Figure 2 for Europe. Per variable pair with a correlation coefficient larger than 0.7, we excluded one variable to reduce dimensionality of the data without losing linearly important information. To decide which of the two variables to retain, caret's 'findCorrelation' function was applied that removes the variable with the largest mean absolute correlation with all other variables. After this initial screening, the balanced logistic regression model, described in Section 2.2, was run. The results of this first logistic regression indicated that not all variables contributed to the overall prediction performance of the model. These insignificant predictors were removed from the dataset in a fourth pre-processing step, resulting in the final dataset per continent used for the logistic regression discussed in Section 3.

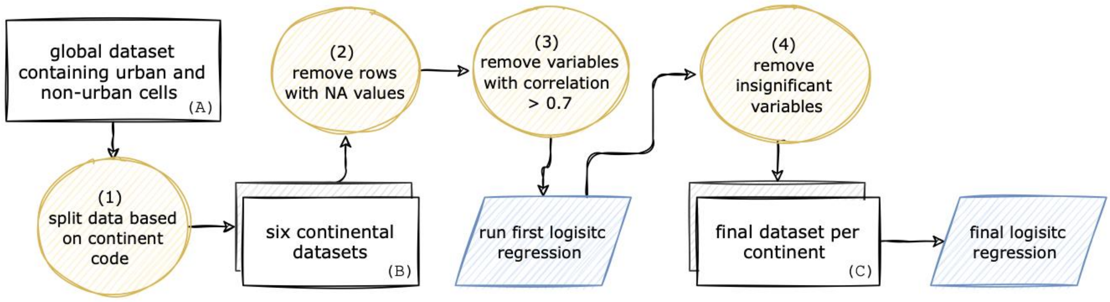

Figure 1.The main data pre-processing and variable selection steps (1-4) in the regression analysis.

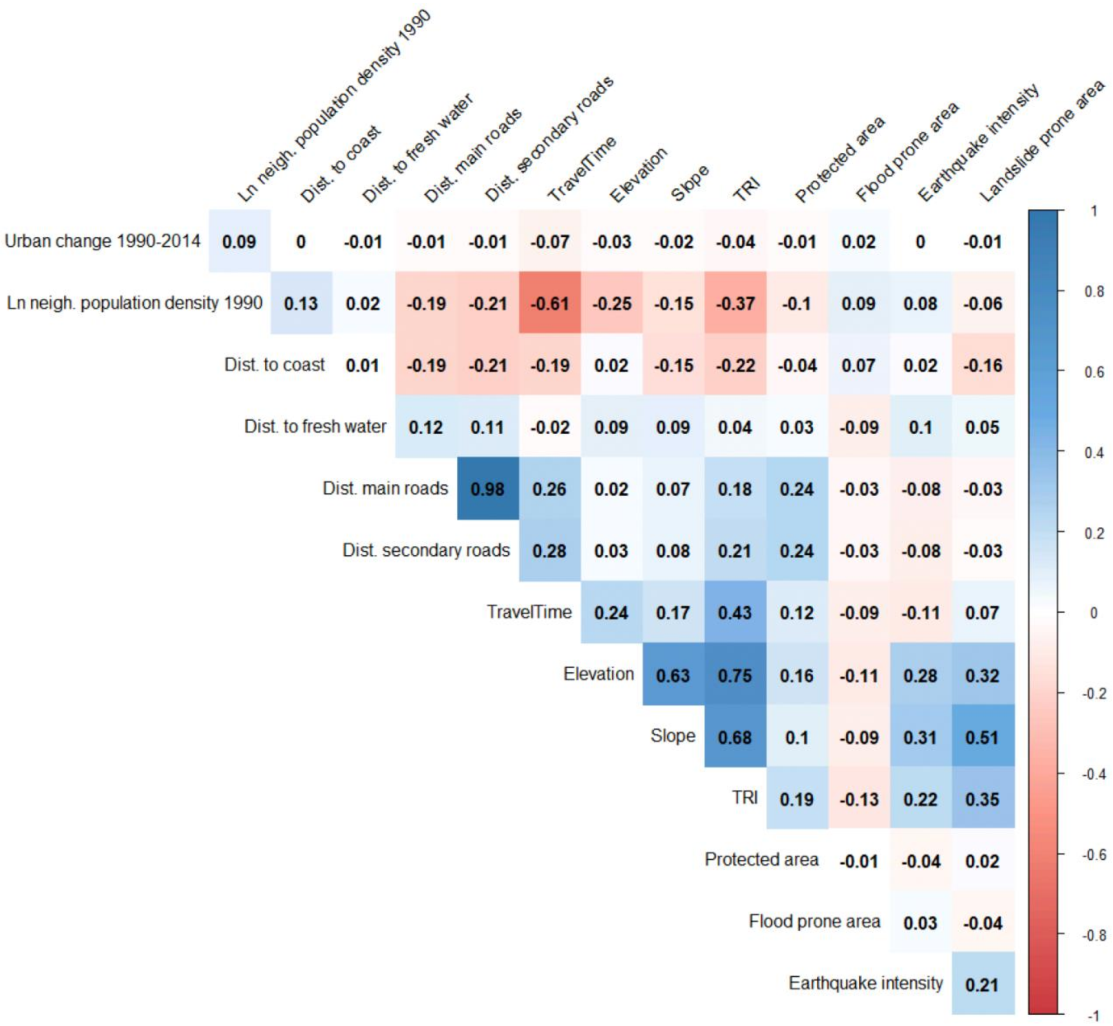

Figure 2. Bivariate correlation coefficients for urban change 1990-2014 and all available explanatory variables in Europe. All correlation coefficients are significant in this example. One variable was excluded per variable pair with a correlation coefficient larger than 0.7.

2.2 Balanced logistic regression model

Several different regression models have been explored to find the most suitable one for the problem at hand (see Ferdinand, 2020). The biggest hurdle in fitting a good regression model is the size of the dataset and the imbalance of the two classes (urban/non-urban). Giving an example, Table 1 shows that less than only 0.27% of all observations for Asia are urban. This means that a model can achieve 99.73% classification accuracy by assigning zero suitability for urban land use to all cells, which would be highly problematic when that suitability is subsequently used to simulate plausible future urban development. Therefore, the final model for the analysis was selected based on balanced accuracy metrics that give equal weight to accuracy on both urban and non-urban validation samples. The performance metrics thus ignore the imbalance in classes in a validation sample and only reward model predictions that differentiate between both urban and non-urban samples within both classes rather than simply reflecting the near-zero rate of occurrence of urban land use itself[3]. By comparison, the uninformative model that would reach 99.73% accuracy by never predicting an urban pixel would only reach Balanced Classification Accuracy of 50%. More discussion on balanced validation metrics, their policy interpretation in the context of model-based decision making, and the connection to a Weighted Maximum Likelihood criterion can be found in Andrée et al. (2020).

Given the large datasets and the training data's imbalancedness, a logit model with a downsampling approach was used for further analysis. This has the benefit that the training data becomes balanced by discarding samples from the majority (non-urban) class and thus, apart from improving model predictions on a balanced performance metric, also smaller and easier to process. The predictive modelling was done in R using the *caret* package, short for Classification And REgression Training. A short overview of the operational steps in the statistical analysis is included in Appendix 2.

A training and testing sample is created from the original dataset to prevent overfitting and ensure the model performs well on new data. The train/test split was set at 0.5 and was generated using a stratified sampling approach that leaves the class frequency intact in both the training and holdout

[3] In particular, a linear model with down-sampling is comparable to a model with case weights that explicitly increases the weight of individual minority class observations (rather than duplicating them in the estimation sample). Maximum Likelihood Estimation is a probabilistic framework for estimating the parameters of a model, and for a logit model, the likelihood function is the same as the log-loss function. Estimating the logit model with down-sampling is thus equal to calibrating it by minimizing a balanced log-loss function. Balanced log-loss, for a completely uninformative model, is minimized at a probability prediction of 0.50. This uninformed guess, in turn, corresponds to an expected balanced accuracy of 50% if residual errors are uniform. The balanced accuracy will be used to interpret and compare the different models for each continent.

data.[4] *Caret's* train function was used to tune the logit model using data from the train split and test its performance using cross-validation techniques on the training data. A standard k-fold cross-validation is used with the number of groups in which the data is split (k) is set at = 10. Cross-validation is often applied in machine learning to estimate the prediction performance of a model; in this case, we estimate the model 10 times on 90% of the training data and validate the predictions on the remaining 10%, finally averaging out the 10 out-of-sample performance estimates to obtain a final estimate of true holdout performance.

Data manipulations, such as the down-sampling scheme used to obtain a more balanced model fit, occur "within" fold; that is, they are repeated separately on each of the 90% samples such that the cross-validation estimate reflects the impact of such procedures on out-of-sample predictions. This procedure can be used to estimate the impact of and decide between various modelling decisions. These decisions are represented by different "hyper-parameters," which are often referred to as "tuning-parameters" when cross-validation is used to optimize over them explicitly. A key tuning parameter used in this analysis is a regularization, or "penalty," parameter that discourages overly complex models by shrinking parameter values toward 0 when predictors are weak. In the context of neural networks, this parameter is often called "decay." This language will be adopted here. To ensure that the penalization similarly impacts parameters of different predictors and thus disregards the variance of that predictor, each variable is normalized to the [0,1] range within each fold. Finally, cross-validation is only performed on the 50% train split, so the remaining 50% test split can be used for final validation.

In the subsequent sections, the results from the regression analyses are described, starting with the analysis of all urban areas present in 2014 in Section 2 and the analysis of the urban area developed between 1990 and 2014 in Section 3.

[4] The data split was created with the createDataPartition function from the caret package.

3 Results

3.1 Results static analysis (urban area 2014)

In this section, the regression results for the existing urban area in 2014 will be presented. By way of example, the cross-validation results for Europe are shown in Figure 3. The figure demonstrates how the cross-validation estimates of a balanced log-loss metric vary across different values of the decay parameter. In particular, it shows that out-of-sample prediction performance did not improve when higher values of decay were used. This highlights that the standard logistic regression model is not overfitting on the estimation sample. Therefore, there is no need to discourage large parameter values for one or several weak predictors. In the case of Europe, the best log-loss value is 0.179 at a weight decay of 0.001. Consequently, this regularization parameter is used for the final model of this continent. The weight decays for the other continents are shown in Table 3.

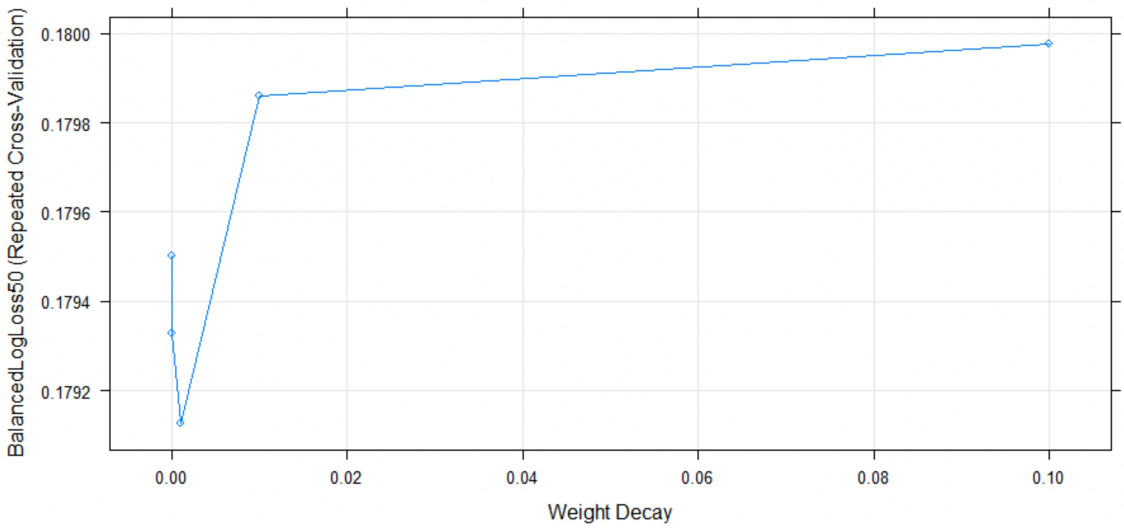

Figure 3. Cross Validation results for Europe showing that in this case increased weight decay does not results in lower log-loss values. The noninformation value for log-loss is approximately 0.693.

The log-loss is a classification loss function often used to evaluate classification models. In general, the lower the log-loss statistic is, the more often the model predicts the correct class with high certainty. For example, if the actual cell is urban, and the model predicts with a probability of 1 that this cell is urban, the cost function (log-loss) 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 a non-urban cell as urban with high certainty, the log-loss value will increase drastically.

In addition to the log-loss values, Table 3 also contains two other important measures for model accuracy, the false negative rate (FNR) and the false positive rate (FPR) for each continent. The FNR represents the probability of failure to predict urban given that the actual class label is urban. The

FNR is also known as the miss-rate and is calculated as follows: FN/(FN+TP), where FN is the number of false negatives, and TP is the number of true positives (FN+TP being the total number of positives). The FPR is the rate of incorrectly classifying urban land use, given that the actual class label is non-urban. The FPR is calculated as follows: FP/(FP+TN), where FP is the number of false positives, and TN is the number of true negatives (FP+TN being the total number of negatives). The different FPRs and FNRs shown in Table 3 are very low, indicating a good performance of the models. The balanced accuracy takes one minus an equal weight of FPR and FNR and is very high for all continents, as shown in Table 3. Recall that an unpredictive model would score only 50%. This means that we can predict whether a cell is urban or not for all continents with high accuracy.

When interpreting these performance metrics, it is important to consider that an arbitrary 0.5 probability value is applied as threshold value for classification. This deviates from the envisioned application in a land-use model that is constrained by an expected amount of urban development. The model generates urban pixels based on the interaction between expected demand for urban development and the probabilities generated by the suitability assessment discussed in this report. In effect, the threshold for urban classification would be controlled by the demand equation in the simulation model. The balanced log-loss metric is thus the appropriate measure for the intended application as it measures how close the predicted probabilities are to observed outcomes. The classification rates here are merely discussed because of their more straightforward interpretation. False positives can be expected to occur less frequently in the simulation output as the amount of simulated demand for urban development would imply a classification threshold higher than 0.5, resulting in a smaller number of urban classifications. Regardless of these considerations, all validation metrics point to very high predictive power of the regional suitability models.

Table 3. Confusion matrix for logit model with down sampling, explaining presence of urban area in 2014.

Statistics	Africa	Asia	Australia & Oceania	Europe	North America	South America
Weight decay	0	0.0001	0.0001	0.001	0	0
True positive (TP)	26,837	56,537	6,531	74,787	87,283	15,771
False positive (FP)	1,009,406	1,269,644	109,447	489,814	692,640	344,484
False negative (FN)	1,435	964	153	3,228	1,559	367
True negative (TN)	17,124,833	19,826,346	5,318,682	4,945,424	19,361,486	10,456,818
False negative rate (FNR): FN/(FN+TP)	5.58%	6.00%	2.02%	8.15%	3.46%	3.24%
False positive rate (FPR): FP/(FP+TN)	4.87%	1.73%	1.93%	4.67%	1.66%	2.40%
Balanced log-loss	0.150	0.113	0.061	0.179	0.078	0.091
Accuracy: (TN + TP)/(TN+TP+FN+FP)	0.944	0.940	0.980	0.911	0.966	0.968
Balanced accuracy	0.947	0.962	0.978	0.934	0.974	0.973

Besides the performance results, we are very interested in the regression coefficients for each continent that show which explanatory variables are relevant in explaining urban land use and how these differ between regions. What stands out first is that the total number of relevant explanatory variables and the variables themselves differ between the continents (Table 4). For example, for North America, twelve statistically significant variables remain after the pre-processing steps and the two regressions, whereas for Europe, only eight variables remain. Most of the included variables are significant at the 1% level. For Africa, Asia, and Europe, one of the significant variables in step number four (Figure 1) was no longer significant after removing the variables that did not yield statistically significant coefficients during the first logistic regression.

For most variables, their influence on the probability of a cell being urban is either positive or negative across all continents, however there are some differences depending on the regional setting. For example, in all continents but Australia and North America, the presence of protected areas negatively influences the probability of finding an urban area. The z-statistics included in Appendix 3 show that neighbourhood population density is by far the most important variable for explaining the presence of urban land use. Travel time is the second most important variable across all continents. A full discussion of the continent-specific variation in the importance of these drivers is beyond the scope of this short report. But being able to specify this variation will allow a more specific definition of suitable locations for urban development.

Table 4. Regression results for logit model with down sampling, explaining presence of urban area in 2014.

Variables	Africa	Asia	Australia & Oceania	Europe	North America	South America
Intercept	2.534***	-1.133***	1.763***	-5.683***	-0.631***	-0.633***
Ln neighbourhood population density	0.531***	0.840***	0.870***	1.275***	0.975***	0.724***
Distance to coast	-0.008***	-0.005***	-0.005***	-0.001***	-0.003***	-0.004***
Distance to fresh water				-0.007***	-0.007***	0.011***
Distance main roads	-0.014***	-0.011***	-0.032***	0.006	-0.076***	-0.034***
Distance secondary roads	-0.050***		-0.012***			
Travel time to city centre	-0.038***	-0.035***	-0.030***	-0.033***	-0.053***	-0.037***
Elevation	0.002***	0.001***	0.001***	-0.002***	0.001***	-0.0001**
Slope	-0.066***	0.033***	-0.299***		-0.040**	-0.064***
Terrain roughness index	-0.653***	-0.620***	-0.720***		-0.652***	
Protected area	-0.076	-0.384***	1.005***	-0.575***	0.578***	-0.706***
Flood prone area	0.788***			0.199***	0.179**	1.684***
Earthquake intensity		0.007	-0.198***	-0.008*	-0.029***	0.095***
Landslide prone		-0.440***	1.554***		-0.857***	-0.876***
Number of explanatory variables	10	10	11	9	12	11

Significance coding: *** $p<0.01$, ** $p<0.05$, * $p<0.1$

3.2 Results dynamic analysis (urban expansion 1990–2014)

This final section documents the results for the dynamic analysis explaining where new urban development occurred between 1990 and 2014. As expected, the performance of the models is not as good as for the static analysis that predicts the location of all urban cells, which means we are less well able to explain urban growth between 1990 and 2014 than we are in enplaning the static urban pattern in 2014. The dynamic analysis results show a decreased balanced accuracy and increased FNRs and FPR for all continents (Table 5). Based on these three performance measures, the model for Europe performs worst, and the model for North and South America perform the best. This may very well relate to the fact that recent developments in Europe are more likely to be steered by spatial planning. Since we cannot include local planning regulations because of a lack of data, our simulations, to some extent, produce a counterfactual representation of what could happen without regulations. This may help planners decide at which locations planning attention is most desirable.

Table 5. Confusion matrix for logit model with down sampling, explaining urban expansion 1990-2014.

Statistics	**Africa**	**Asia**	**Australia & Oceania**	**Europe**	**North America**	**South America**
Decay	0.01	0.001	0.01	0.0001	0.1	0
True positive (TP)	11,709	19,875	1,173	14,864	22,457	3,076
False positive (FP)	1,173,224	1,509,489	156,864	669,543	895,092	427,882
False negative (FN)	839	441	58	1,096	518	93
True negative (TN)	16,961,015	19,585,780	5,271,265	4,765,695	19,159,034	10,373,420
False negative rate (FNR)	6.45%	7.11%	2.96%	12.32%	4.46%	3.90%
False negative rate (FPR)	6.25%	2.06%	3.82%	6.95%	2.13%	3.03%
Balanced log-loss	0.178	0.135	0.105	0.256	0.097	0.115
Accuracy	0.935	0.929	0.971	0.877	0.955	0.960
Balanced accuracy	0.934	0.953	0.962	0.904	0.966	0.966

The regression coefficients from the dynamic analysis, as shown in Table 6, are mostly in line with the coefficients from the static analysis (Table 4). Mean population density remains the most important variable for all continents. However, the effect decreased compared to the static analysis.

Table 6. Regression results for logit model with down sampling, explaining urban expansion 1990-2014.

	Africa	**Asia**	**Australia & Oceania**	**Europe**	**North America**	**South America**
Intercept	2.725***	0.495***	2.766***	-4.122***	1.124***	0.217
Ln neighbourhood population density	0.495***	0.658***	0.758***	1.102***	0.777***	0.692***
Distance to coast	-0.007***	-0.006***	-0.003**	-0.002***	-0.004***	-0.005***
Distance to fresh water		0.003***			-0.005**	0.012***
Distance main roads	-0.022***	-0.037***	-0.057***	-0.112***	-0.101***	-0.016***
Distance secondary roads	-0.052***	-0.003***				
Travel time to city centre	-0.036***	-0.035***	-0.037***	-0.039***	-0.061***	-0.042***
Elevation	0.001***	0.001***		-0.001***	0.001***	
Slope		0.045**	-0.727***	-0.139***	0.095***	
Terrain roughness index	-0.647***	-0.567***	-0.216		-0.714***	-0.055
Protected area	0.470***	-0.411**	1.624***	-0.414***	0.940***	
Flood prone area	0.822***			0.494***		1.405***
Earthquake intensity	0.039**		-0.235***	0.020**	-0.058***	
Landslide prone		-0.571***	1.454***		-1.078***	-0.506**
Number of explanatory variables	10	11	9	9	11	8

Significance coding: *** p<0.01, ** p<0.05, * p<0.1

4 Recommendations for further research

The updated calibration results presented in this report were used to help improve the simulation of urban growth patterns with the 2UP model. However, further improvement is possible as is discussed in this concluding section.

One aspect of the current research design that could be improved is the incorporation of land-use data. Several global datasets are available that distinguish different types of use, such as forest, agricultural land, barren land, and grassland (e.g., MODIS or CCI-LC, see Diogo and Koomen, 2016). This data can be used to characterise local neighbourhoods and may indicate types of use that are more likely to transition into urban land. The data can also be used to help overcome the current imbalance between urban and non-urban land by creating the opportunity for a sampling strategy that samples proportionally from different land-use types.

To further improve the calibration, regression analysis could be performed for smaller and more homogeneous regions than the continents applied in this study. Especially on large and diverse continents such as Europe and Asia, urban expansion patterns are expected to differ between countries. Analysis of individual countries could also provide the opportunity to include country-specific drivers such as zoning policies. Alternatively, countries could be grouped according to their urban development trajectory. For example, distinguishing between maturity stages based on urban share of population and urban land growth rate (as proposed by Gao and O'Neill, 2020) or urban change trajectories defined as a function of changes in population and urban land intensity (as described by Li et al., 2022).

Finally, by utilising cloud computing, more advanced regression techniques can be considered, including interaction variables and spatial lags.

References

Andrée, B.P.J., Chamorro, A., Kraay, A., Spencer, P., Wang, D., (2020) Predicting Food Crises, Policy Research Working Paper. World Bank, Washington DC.

Andrée, B.P.J., Koomen, E. (2017) Calibration of the 2UP model: Spinlab Research Memorandum SL-13. Vrije Universiteit Amsterdam.

Diogo, V., Koomen, E. (2016) Land Cover and Land Use Indicators; review of available data. OECD Publishing, Paris.

Ferdinand, P.A., (2020) Forecasting patterns of urban expansion; A statistical analysis of forces determining locational urban growth and their regional differences. MSc thesis., STREEM. Vrije Universiteit Amsterdam.

Gao, J., O'Neill, B.C. (2020) Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways. Nature Communications 11, 2302.

Li, M., Verburg, P.H., van Vliet, J. (2022) Global trends and local variations in land take per person. Landscape and Urban Planning 218, 104308.

van Huijstee, J., van Bemmel, B., Bouwman, A., van Rijn, F., (2018) Towards an urban preview; Modelling future urban growth with 2UP. PBL Netherlands Environmental Assessment Agency, The Hague.

Appendix 1 Data sources and processing

An overview of all included datasets including a short description and their sources is listed below.

Table 7 Overview of included data sets and their sources.

Variable name	Description	Source
Urban area	Presence of urban land use (1 =true), classified at 30 arc seconds based on aggregation of original 38m raster, applying a 50% threshold	Global Human Settlement layer (GHSL, Pesaresi et al., 2016)
Ln neighbourhood population density	Mean population density in eight surrounding cells, not considering central cell. Ln transformation is applied to limit impact of occasionally very high values, replacing 0 values with 0.1 to prevent errors.	Global Human Settlement layer (GHSL, Pesaresi et al., 2016)
Distance to coast	Euclidean distance to coast line in kilometres; cut off at 250km (higher values set to 250km)	Coastlines are inferred from national boundaries in the Database of Global Administrative Areas (available from: https://gadm.org/)
Distance to fresh water	Euclidean distance to nearest river or lake (whichever is closest) in km, cut off at 250km	HydroRIVERS database (Lehner and Grill, 2013) and Global Lakes and Wetlands Database (Lehner and Döll 2004)
Distance main roads	Euclidean distance to nearest highway, primary or secondary roads (types 1-3) in km, cut off at 250km	Global Roads Inventory Project (GRIP) dataset; (Meijer et al., 2018)
Distance secondary roads	Euclidean distance to nearest tertiary or local road (types 4 or 5) in km, cut off at 250km	Global Roads Inventory Project (GRIP) dataset; (Meijer et al., 2018)
Travel time to city centre	Travel time in minutes to the nearest city centre, with maximum value of 60 minutes	Travel time to settlements with over 50,000 inhabitants (from CIESIN et al., 2017) over a cost surface grid based on the GRIP road data set.
Elevation	Elevation in meters	SRTM V3 (Jarvis et al., 2006) and GTOPO30 for high latitudes (USGS, 1996)
Slope	Slope in degrees	GTOPO30 (USGS, 1996)
Terrain roughness index	Index describing overall ruggedness within a cell, based on slope and elevation	Calculated based on method by Riley et al. (1999) using Slope and Elevation datasets
Protected area	Presence of protected area	World database on protected areas (WDPA; UNEP-WCMC, 2019)
Flood prone area	Location exposed to river floods with a 100-year return period	Global Risk Data Platform (UNEP/GRID-Geneva, 2018)
Earthquake intensity	Estimate of the Modified Mercalli Intensity of earthquakes with a 475-year return period	Global Risk Data Platform (UNEP/GRID-Geneva, 2018)
Landslide prone	Location exposed to a potentially destructive landslide with a 200,000-year return period. Probabilities are corrected for the share of the cell that is potentially affected.	Global Risk Data Platform (UNEP/GRID-Geneva, 2018). Precipitation-induced landslide areas are selected here that overlap the areas facing landslides triggered by earthquakes, but also cover additional area.

To characterise accessibility, we use the Global Roads Inventory Project (GRIP) dataset, distinguishing between main roads and secondary roads. As we assume the impact of roads negligible at very long distances, we apply a maximum distance of 250 kilometres for both main and secondary roads. Higher values are kept at that maximum value. Figure 4 shows the spatial distribution of the cut-off values in red. This procedure has the additional advantage of preventing unrealistic values at, for example, isolated islands that would otherwise range in the 1000's. The same cut-off value is applied to other distance-based variables.

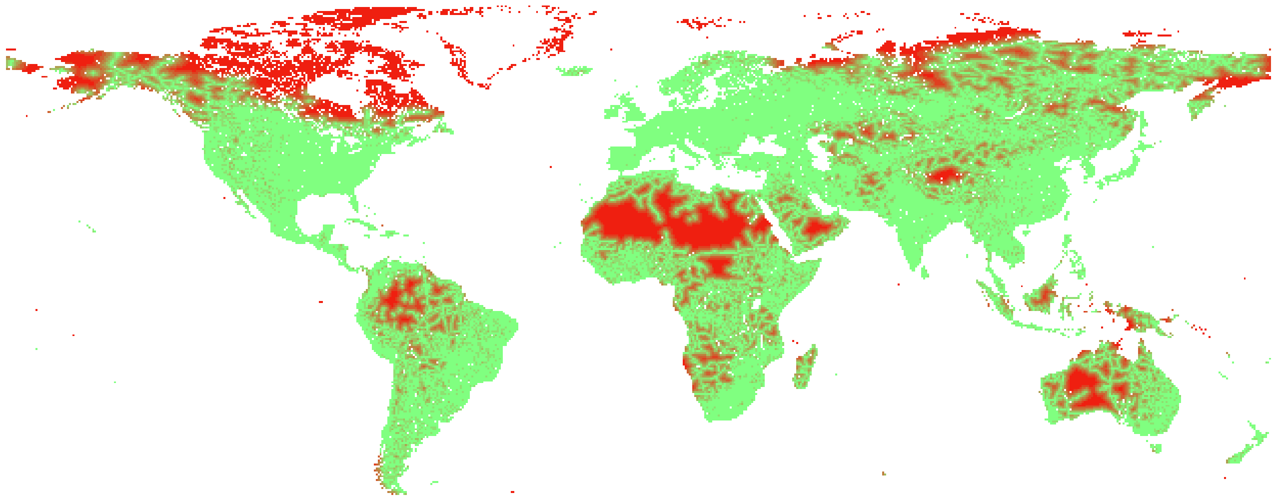

Figure 4. Distance to the nearest main road cut off at 250 kilometres (in red).

Soil type is generally considered to be an important driver for urban development. Studies for South- and North America, for example, showed 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). In initial analysis we, therefore considered soil type as independent variable. Data on the type of soil was obtained from the harmonized world soil database (HWSD; FAO et al., 2009). The data contains 33 categories, but several types showed near-zero variation (meaning they are relatively rare) while others occurred only in a number of continents. After removing the types with very low variance, we included reference to over 20 soil types in our logistic regression as an equal number of dummy variables indicating the presence (or absence) of individual soil types. Figure 5 below shows the cross-validation results of this initial test.

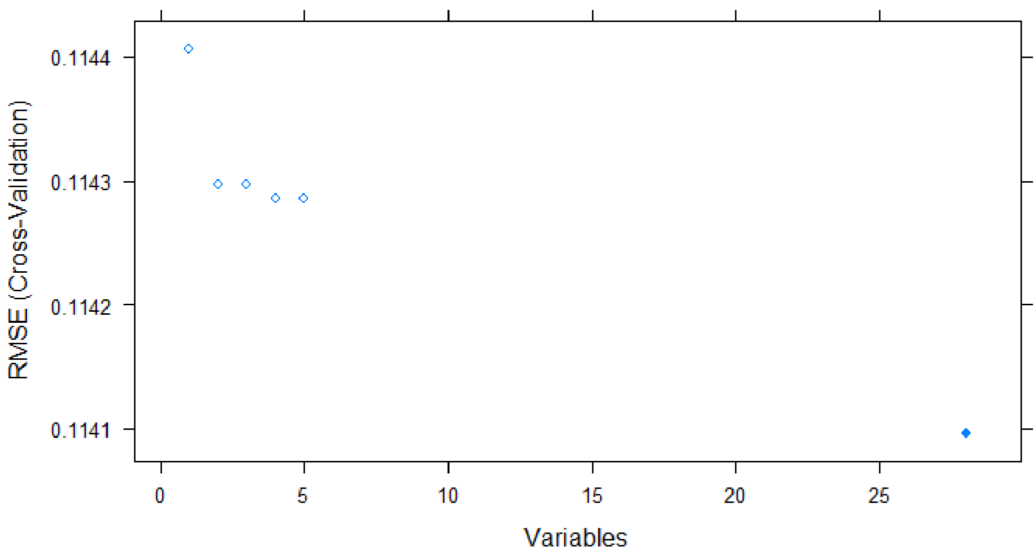

Figure 5. Cross-validated root mean square error (RMSE, also referred to as Brier Score in this context) for models including an increasing number of soil class dummies. Lower error is obviously preferred. The graph highlights that the improvement in prediction performance is extremely low when using soil dummies, compared to the model that does not include any soil dummies.

The figure indicates that the impact of including the set of over 20 variables that refer to soil classes on top of a basic model with 5 variables is minimal in terms of on limiting the root mean square error. Adding so many individual variables does limit the possibility of including other more relevant variables, however, so it was decided to exclude soil type from further analysis.

References for data sources

Aguayo, M., Wiegand, T., Azócar, G.D., Wiegand, K. and Vega, C.E. (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): 13.

Batisani, N. and Yarnal, B. (2009) Urban expansion in Centre County, Pennsylvania: Spatial dynamics and landscape transformations. Applied Geography 29(2): 235–249.

Center for International Earth Science Information Network (CIESIN) Columbia University, CUNY Institute for Demographic Research (CIDR) City University of New York, International Food Policy Research Institute (IFPRI), The World Bank and Centro Internacional de Agricultura Tropical (CIAT) (2017) Global Rural-Urban Mapping Project, Version 1 (GRUMPv1); Settlement Points, Revision 01. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, NY. Data available from: https://doi.org/10.7927/H4BC3WG1

FAO, IIASA, ISRIC, ISS-CAS and JRC (2009) Harmonized World Soil Database (version 1.1). Rome, Italy and IIASA, Laxenburg, Austria. Data available from: https://www.fao.org/land-water/land/land-governance/land-resources-planning-toolbox/category/details/en/c/1028012/

Jarvis A., Reuter, H.I., Nelson, A. and Guevara, E. (2006) Hole-filled SRTM for the globe version 3, from the CGIAR-CSI SRTM 90m database. Data available at: http://srtm.csi.cgiar.org

Meijer, J.R., Huijbegts, M.A.J., Schotten, C.G.J. and Schipper, A.M. (2018) Global patterns of current and future road infrastructure. Environmental Research Letters, 13-064006. Data available at: www.globio.info

Lehner, B. and Döll, P. (2004) Development and validation of a global database of lakes, reservoirs and wetlands. Journal of Hydrology 296 (1-4): 1-22. Data available at: https://www.worldwildlife.org/publications/global-lakes-and-wetlands-database-lakes-and-wetlands-grid-level-3

Lehner, B. and Grill G. (2013) Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15): 2171–2186. Data available at www.hydrosheds.org

Pesaresi, M., Ehrilch, D., Florczyk, A.J., Freire, S., Julea, A., Kemper, T., Soille, P. and Syrris, V. (2016) GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014). European Commission, Joint Research Centre (JRC). Data available from: https://ghsl.jrc.ec.europa.eu/

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): 24–33.

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

UNEP/GRID-Geneva (2018) Global Risk Data Platform. Created and hosted by UNEP/GRID-Geneva. Data available from: https://preview.grid.unep.ch

UNEP-WCMC (2019) User Manual for the World Database on Protected Areas and world database on other effective area-based conservation measures: 1.6. UNEP-WCMC: Cambridge, UK. Data available at: https://www.protectedplanet.net/

US Geological Survey (1996) Global Digital Elevation Model (GTOPO30). EROS Data Center. Data available at https://lta.cr.usgs.gov/GTOPO30

Appendix 2 General steps in the calibration process

Below is an overview of the general workflow to calibrate the 2UP model. A more detailed description of the individual steps can be found in the R script files mentioned below and available from the first author.

1. Export dataset from GeoDMS (export csv)
2. Open R and use file: *Continent_calibration_2010_rfe.R* or *Continent_Calibration_1990_2010_New.R* for following steps:
    a. Import csv file
    b. Make sample for each continent (Asia has to be sampled with AutoGLM)
3. Use R file: Caret_predicitve_models_2UP_FINAL.R for each continent separately:
    a. Import clean data continent
    b. Create training and testing data
    c. Run regression model

Appendix 3 Full regression results

Table 8. Results for logit model with down sampling, explaining presence of urban area in 2014.

Variables	Africa				Asia				Australia & Oceania			
	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.
Intercept	2.534	0.082	30.922	0.000	-1.133	0.085	-13.262	0.000	1.763	0.268	6.580	0.000
Ln neighbourhood pop. dens.	0.531	0.006	87.515	0.000	0.840	0.010	83.385	0.000	0.870	0.027	31.715	0.000
Distance to coast	-0.008	0.000	-30.003	0.000	-0.005	0.000	-27.878	0.000	-0.005	0.001	-5.595	0.000
Distance to fresh water												
Distance main roads	-0.014	0.001	-15.350	0.000	-0.011	0.002	-4.765	0.000	-0.032	0.006	-5.411	0.000
Distance secondary roads	-0.050	0.003	-14.408	0.000					-0.012	0.003	-3.836	0.000
Travel time to city centre	-0.038	0.001	-32.010	0.000	-0.035	0.001	-37.473	0.000	-0.030	0.004	-7.051	0.000
Elevation	0.002	0.000	22.639	0.000	0.001	0.000	13.727	0.000	0.001	0.000	4.077	0.000
Slope	-0.066	0.019	-3.475	0.001	0.033	0.013	2.603	0.009	-0.299	0.091	-3.275	0.001
Terrain roughness index	-0.653	0.031	-21.076	0.000	-0.620	0.025	-25.035	0.000	-0.720	0.105	-6.878	0.000
Protected area	-0.076	0.093	-0.810	0.418	-0.384	0.108	-3.567	0.000	1.005	0.277	3.631	0.000
Flood prone area	0.788	0.087	9.025	0.000								
Earthquake intensity					0.007	0.006	1.178	0.239	-0.198	0.032	-6.204	0.000
Landslide prone					-0.440	0.074	-5.914	0.000	1.554	0.337	4.613	0.000

Table 8. continued

Variables	Europe				North America				South-America			
	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.
Intercept	-5.683	0.054	-105.270	0.00	-0.631	0.075	-8.471	0.000	-0.633	0.124	-5.123	0.00
Ln neighbourhood pop. dens.	1.275	0.008	152.796	0.00	0.975	0.010	102.519	0.000	0.724	0.013	55.019	0.00
Distance to coast	-0.001	0.000	-5.999	0.00	-0.003	0.000	-12.822	0.000	-0.004	0.000	-10.520	0.00
Distance to fresh water	-0.007	0.001	-9.113	0.00	-0.007	0.001	-5.759	0.000	0.011	0.002	6.649	0.00
Distance main roads	0.006	0.005	1.068	0.29	-0.076	0.015	-5.116	0.000	-0.034	0.006	-5.571	0.00
Distance secondary roads												
Travel time to city centre	-0.033	0.001	-38.096	0.00	-0.053	0.001	-58.809	0.000	-0.037	0.002	-19.632	0.00
Elevation	-0.002	0.000	-33.142	0.00	0.001	0.000	9.684	0.000	-0.0001	0.000	-2.012	0.04
Slope					-0.040	0.018	-2.274	0.023	-0.064	0.019	-3.272	0.00
Terrain roughness index					-0.652	0.031	-21.144	0.000				
Protected area	-0.575	0.043	-13.288	0.00	0.578	0.160	3.621	0.000	-0.706	0.168	-4.210	0.00
Flood prone area	0.199	0.041	4.807	0.00	0.179	0.071	2.511	0.012	1.684	0.148	11.345	0.00
Earthquake intensity	-0.008	0.005	-1.757	0.08	-0.029	0.008	-3.600	0.000	0.095	0.017	5.605	0.00
Landslide prone					-0.857	0.094	-9.111	0.000	-0.876	0.116	-7.527	0.00

Table 9. Results for logit model with down sampling, explaining urban expansion (1990-2014).

Variables	Africa				Asia				Australia & Oceania			
	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.
Intercept	2.725	0.112	24.405	0.000	0.495	0.112	4.409	0.000	2.766	0.511	5.412	0.000
Ln neighbourhood pop. dens.	0.495	0.008	61.751	0.000	0.658	0.012	53.085	0.000	0.758	0.052	14.643	0.000
Distance to coast	-0.007	0.000	-19.108	0.000	-0.006	0.000	-19.895	0.000	-0.003	0.002	-2.083	0.037
Distance to fresh water					0.003	0.001	3.826	0.000				
Distance main roads	-0.022	0.001	-14.669	0.000	-0.037	0.005	-7.021	0.000	-0.057	0.011	-5.014	0.000
Distance secondary roads	-0.052	0.004	-11.618	0.000	-0.003	0.000	-6.188	0.000				
Travel time to city centre	-0.036	0.002	-21.809	0.000	-0.035	0.001	-23.755	0.000	-0.037	0.009	-4.050	0.000
Elevation	0.001	0.000	15.863	0.000	0.001	0.000	7.381	0.000				
Slope					0.045	0.019	2.406	0.016	-0.727	0.206	-3.533	0.000
Terrain roughness index	-0.647	0.036	-17.900	0.000	-0.567	0.038	-15.105	0.000	-0.216	0.157	-1.373	0.170
Protected area	0.470	0.110	4.293	0.000	-0.411	0.171	-2.404	0.016	1.624	0.441	3.684	0.000
Flood prone area	0.822	0.122	6.731	0.000								
Earthquake intensity	0.039	0.019	2.104	0.035					-0.235	0.058	-4.028	0.000
Landslide prone					-0.571	0.109	-5.227	0.000	1.454	0.480	3.032	0.002

Table 9. continued

Variables	Europe				North America				South-America			
	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.	beta	std.err.	z-stat.	sign.
Intercept	-4.122	0.101	-40.754	0.000	1.124	0.124	9.102	0.000	0.217	0.256	0.847	0.397
Ln neighbourhood pop. dens.	1.102	0.017	65.825	0.000	0.777	0.015	52.299	0.000	0.692	0.024	28.244	0.000
Distance to coast	-0.002	0.000	-8.116	0.000	-0.004	0.000	-12.079	0.000	-0.005	0.001	-6.001	0.000
Distance to fresh water					-0.005	0.002	-2.306	0.021	0.012	0.003	3.519	0.000
Distance main roads	-0.112	0.014	-7.912	0.000	-0.101	0.024	-4.290	0.000	-0.016	0.006	-2.824	0.005
Distance secondary roads												
Travel time to city centre	-0.039	0.002	-22.232	0.000	-0.061	0.002	-36.787	0.000	-0.042	0.004	-10.972	0.000
Elevation	-0.001	0.000	-6.780	0.000	0.001	0.000	7.475	0.000				
Slope	-0.139	0.015	-9.588	0.000	0.095	0.030	3.151	0.002				
Terrain roughness index					-0.714	0.053	-13.383	0.000	-0.055	0.051	-1.068	0.286
Protected area	-0.414	0.084	-4.923	0.000	0.940	0.227	4.135	0.000				
Flood prone area	0.494	0.081	6.103	0.000					1.405	0.302	4.654	0.000
Earthquake intensity	0.020	0.009	2.282	0.022	-0.058	0.015	-3.982	0.000				
Landslide prone					-1.078	0.162	-6.663	0.000	-0.506	0.233	-2.174	0.030