Exploring alternative calibration methods for simulating local built-up fraction change in the CRISP model.

van der Wielen, Thijmen; Jacobs-Crisioni, Chris; Koomen, Eric; Andree, Bo Pieter Johannes

2025

Exploring alternative calibration methods for simulating local built-up fraction change in the CRISP model

Thijmen van der Wielen

Chris Jacobs-Crisioni

Bo Pieter Johannes Andrée

Eric Koomen

19 December 2025

Colophon

Title

Exploring alternative calibration methods for simulating local built-up fraction change in the CRISP model
Spinlab Research Memorandum SL-27

Authors

Thijmen van der Wielen[1]
Chris Jacobs-Crisioni[2]
Bo Pieter Johannes Andrée[3]
Eric Koomen[1]

[1] Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam
[2] Bureau Jacobs-Crisioni
[3] World Bank, Development Data Group

Contact

Vrije Universiteit Amsterdam

School of Business Economics
Department of Spatial Economics/ Spatial Information Laboratory (SPINlab)
De Boelelaan 1105
1081 HV Amsterdam
Netherlands
Phone: +31 20 5986095
Email: t.j.a.van.der.wielen@vu.nl
Website: spinlab.vu.nl

Contents

1 Introduction

The new CRISP model ('Cities and Rural Integrated Spatial Projections') developed by the European Commission's Joint Research Centre projects future built-up area development and population patterns at a global level (Jacobs-Crisioni et al., 2025). The model uses calibrated suitability functions to guide the allocation of built-up land and population. The initial calibration of the model uses a balanced logistic regression to explain the occurrence of built-up area in the base year for the model's projection (Van der Wielen and Koomen, 2024). This binary measure to determine the presence of built-up area directly fitted CRISP's predecessor 2-UP that simulated built-up area development as a dichotomous outcome (Andrée and Koomen, 2017; Ferdinand et al., 2021). However, one of the improvements implemented in the CRISP model is the adoption of a continuous measure of built-up area fractions.

While the use of binarised balanced logistic regression to calibrate CRISP yields realistic outcomes, Van der Wielen and Koomen (2024) mention that improvements may be possible through the utilisation of statistical methods more suited for fractional observations (Kieschnick and McCullough, 2003; Papke and Wooldridge, 1996). This may have the additional benefit of making the required binarisation redundant, allowing us to directly ascertain a degree of suitability.

This report explores alternative strategies for estimating suitability functions with the goal of improving the calibration step of the CRISP model. We aim to provide an overview of econometric techniques that are either specifically designed or are particularly well suited for the estimation of models with bounded fractional dependent variables. Specifically, we review two fractional regression models (fractional logit & beta regression) and a method that combines linear regression with machine learning (cubist). In addition to this exploration, we want to provide an explanation of the particularities relating to fractional regression and give some insight into which techniques work best for different analytical scenarios. By doing so, we contribute to the growing literature on the application of machine learning techniques in calibrating land-use models (see, e.g., Black et al., 2023).

2 Methods

The initial calibration of the CRISP model, which is explained in greater detail in Ferdinand (2020), utilises a balanced logistic regression model[1] to explain the occurrence of built-up area in a single year, which corresponds with the base year of the projections. Since logistic regression requires the dependent variable to be binary (or at the very least *discrete*), it will only return the probability that an observation will take the true status, in this case whether a cell is likely to contain more than the threshold amount of 2.5% built-up area. Astute readers may have noticed this logistic probability does not map one-to-one to a prediction of the Built-Up Area Fraction (BUA). In fact, it only predicts whether a grid cell is likely to exceed the threshold value, but it does not matter by how much this value is exceeded. We can, of course, presume that a higher estimated likelihood will correspond to a higher BUA, a not uncommon assumption in the larger land-use modelling literature (see, e.g., Koomen et al., 2015; Loonen and Koomen, 2009).

An important issue with this presumption in the current CRISP model is the nature of the built-up fractions we try to simulate. These show a peculiar distribution: lots of 0s and a cap around 0.5 (discussed in more detail below). The latter is problematic, as directly using suitability values representing the likeliness of being urban would result in a model attempting to simulate fractions approximating a value of 1. The CRISP version applied for the World Urbanisation Prospects (2025) solves this by using a series of transformations (starting with the binary logit predictions that are first rescaled per level of BUA and subsequently multiplied with so-called Expected Top-up (ET) values) to provide a proxy of change in built-up development (Jacobs-Crisioni et al., 2025). These additions make the model more complex and this study aims to see if alternative statistical techniques are better suited to specify suitability. Potentially even integrating some form of ETs directly into the suitability model.

In the remainder of this section, we will discuss the theory and practical implementation of several alternative estimation methods and explain how they may help improve calibration of CRISP. We will first examine regression techniques that were developed for fractional responses on open intervals, before moving on to more specialised techniques for dependent variables on closed intervals. These methods are discussed separately because, as will become clear, fractional regression is substantially simpler when dealing with bounded dependent variables on open intervals.

2.1 Distribution of fractional dependent in the CRISP calibration

Before we go into the theoretical background of the estimation techniques, it is useful to discuss the different types of fractional dependent variables and clarify some of the issues that may occur when attempting to predict fractional values. Fractional response variables can be

---

[1] The reason for using *balanced* logit over regular logit is due to the high prevalence of zero observations. These may cause regular logit to produce biased outcomes but can be compensated for in balanced logit through weighted maximum likelihood.

divided into roughly four distributional categories (Kieschnick and McCullough, 2003). The first of these is comprised of proportions on an open interval (0,1), where the observations of the variable will never be equal to its boundary values. The second distributional category describes proportions on a closed interval [0,1], where the observed variable *can* be equal to its boundary values. The third and fourth categories simply describe multivariate extensions of the first two categories. While a dependent variable on an open interval (0,1) is easily modelled via a continuous probability distribution, allowing one to predict these values using (relatively) standard fractional regression methods. Note, however, that the nature of the open interval prevents predicted values of exactly zero (or one). The closed interval [0,1] will include these boundary values, but as result probability masses at these boundary values (Tu, 2012). Giving it a mixed continuous-discrete nature (Ospina and Ferrari, 2012). In effect, the probability distribution of dependent variable $y \in [0,1]$ consists of two components; the discrete probability that y is exactly zero or one, and the continuous probability values for the remaining values of $y \in (0,1)$.

The built-up fraction simulated in the CRISP model is taken from the Global Human Settlement Layer (Pesaresi and Politis, 2023). More specifically, the built-up fraction of both residences and non-residences is used. This includes all buildings but excludes non-built up surfaces like gardens and roads. Meaning observations theoretically fall into a closed interval [0,1], though cells rarely reach this maximum value. A more practical description may be that BUA is observed on an interval $[0, \bar{r}) \subset [0,1]$. Where $\bar{r}$ is some effective maximum level of built-up development that is functionally unattainable, but lower than the theoretical maximum value of one.

Figure 1 shows this distribution for a representative subsample of the BUA observed for North America. Note that the probability massing at the zero-value is clearly reflected in the high prevalence of values of exactly zero, shown in the highlighted column. We can also see the $\bar{r}$ boundary reflected in the distribution, with BUA rarely above 0.5 and never exceeding 0.6. This has been assumed to be the value of this development limit, and in fact, in 2020 only 571 km^2 of surface (0.006% of all surfaces with at least 1% built-up area) exceeds 0.6 BUA according to the GHSL data[2]. The fact that BUA falls on an open interval $[0, \bar{r}) \subset [0,1]$ complicates our situation, since we cannot capture our dependent variable using standard fractional regression methods (Kieschnick and McCullough, 2003; Ospina and Ferrari, 2012). Ideally, we would adopt an estimation strategy that is able to handle the discrete-continuous nature of this dependent variable.

[2] The exceedingly rare cases where the built-up fraction exceeds 0.6 correspond to extensive production facilities for, for example, agriculture in the greenhouse areas of Almeria (Spain) and Westland (The Netherlands).

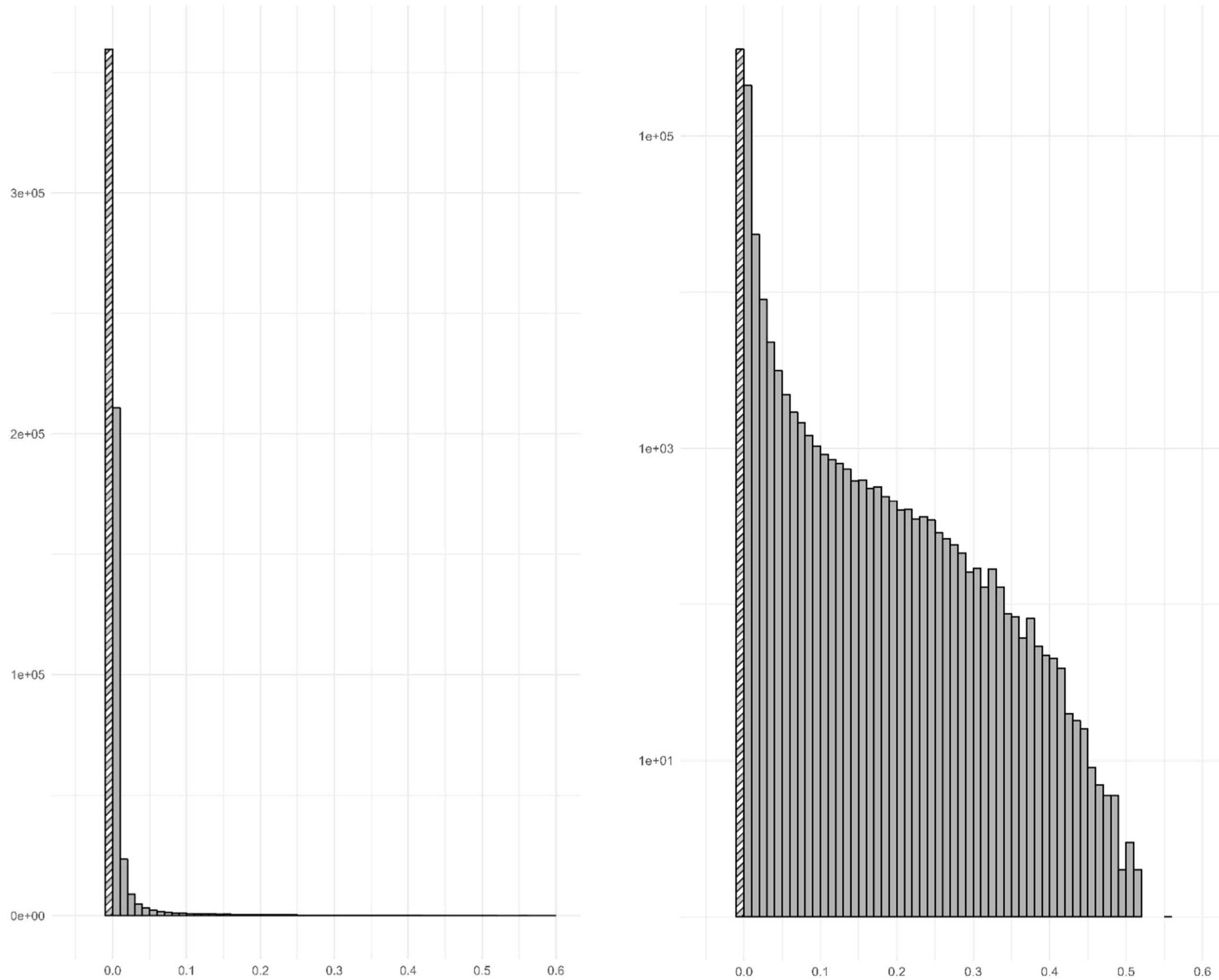

Figure 1 – Histograms of Built-Up Area Fractions for North America using a 5% sample of the full dataset. Zero-values shown isolated in highlighted column (left: linear axis, right: logarithmic axis).

2.2 Fractional logit

Fractional logit, sometimes referred to as *quasi-binomial*, was first proposed by Papke and Wooldridge (1996) as a technique for estimating models involving fractional response variables. They propose that a response variable y, capable of assuming any value within the open interval (0,1), can be effectively estimated through the application of a linear model in combination with a logit link function. This has the advantageous characteristic of having the same log-likelihood function as the Bernoulli log-likelihood used in ordinary binary logit, elegantly minimising the theoretical alterations to the model. It is important to note that, contrary to binary logit, fractional logit is semiparametric and makes no distributional assumption[3].

[3] If GLM assumptions are met, the model is fully parametric under the Bernoulli distribution. Papke and Wooldridge (1996) stress this is not likely to occur.

In practice the implementation of a fractional logit approach is relatively straightforward. Since it shares a functional form with binary logit, we simply exchange a probabilistic binary y in a generalised linear model (GLM) for a fractional equivalent and are then able to directly estimate BUA. No significant technical changes required to the model (or scripts) itself should be necessary. The lack of a viable distributional assumption does require some compensation and requires us to adjust the calculation of standard errors to use a robust estimator. Which tends to produce much larger standard errors. An additional disadvantage of the fractional logit approach is that it cannot predict values of exactly zero (or one), which is undesirable given the distribution of our dependent variable, BUA.

2.3 Beta regression

An alternative technique for handling a fractional response variable is the so-called *beta regression model* (Ferrari and Cribari-Neto, 2004; Kieschnick and McCullough, 2003). This method assumes that the dependent variable y follows a beta distribution. Beta distributions can take on many different shapes, depending on the parameters that index the distribution, making them very versatile (Johnson et al., 1995). This flexibility makes them particularly good candidates for modelling fractional variables on an interval (0,1) (Ferrari and Cribari-Neto, 2004). By assuming a beta distribution, the model remains fully parametric meaning it is more efficient than *fractional logit* (as long as the provided y is beta-distributed).

From a practical point of view, integrating a beta regression approach into the existing CRISP calibration infrastructure requires more substantial adjustments. A beta regression model is more complex by nature, but fortunately most of the tools required are provided in the 'betareg' R-package (Cribari-Neto and Zeileis, 2010). A more significant practical issue is that the base beta regression model was designed for a response variable on an open interval (0,1), which for our purposes means we would need to censor or data to not include any zero (or near-zero) observations of BUA.

The beta regression approach has some shortcomings. Firstly, the strong distributional assumption allows the model to be more efficient but also opens the possibility of distributional misspecification. In essence, we are trading the robustness of the fractional logit model for increased efficiency in the beta regression model. We just need to ensure the response variable, BUA, follows a beta-distribution. Another possible concern one could have is that the increased complexity of the model might cause issues when dealing with small sample sizes. This is worth noting but should not pose a problem for our intended application given the considerable size of the dataset we use. Finally, the most significant limitation of a beta regression approach is its inability to predict values of exactly zero, a characteristic it shares with fractional logit.

2.4 Cubist model

Based on the *M5 model tree* from Quinlan (1992), cubist models are –in a nutshell– a cross between a decision tree and linear regression models. This structure allows for a great deal of versatility in estimation. Approaching the effectiveness of highly advanced machine learning methods while maintaining a reasonable level of legibility, producing outcomes that we can interpret (given enough time and effort). Since the cubist model does not assume a fixed functional form, it has the capacity to approximate and capture non-linear or other complex relationships in the data. This means cubist models have the implicit ability to deal with the zero-inflated nature of our dependent variable, as well as to dynamically integrate the functionality of ETs. It is also worth noting that the structure of cubist models allows the production of regional sub-models within the same model, eliminating the need for manual intervention (unlike the current GLM-based approach).

Implementing a cubist model would be the most impactful on the current CRISP modelling infrastructure. This is a consequence of the cubist methodology; the data is split according to a set of rules, with a distinct linear model being estimated for each of the resulting subsets (Kuhn and Quinlan, 2025). To accommodate this, the resulting output must also follow this rule-based structure. This has knock-on effects outside of the calibration step of CRISP, since these would have to be modified to accommodate this less traditional model structure. The actual calibration step itself also requires some extra attention, as a cubist model uses parameters closer to those used in machine learning. The most important of these for our application, is the maximum number of rules we allow the model to impose on the data. Setting this limit too high will lead to a wholly illegible outcome, which cannot practically be implemented in the larger CRISP model. Setting the limit too low, however, will (partially) negate the advantages of using a cubist model. For this reason, we have first estimated a model with a relatively limited number of rules (10) before moving to a more complex model with a much higher number of rules (100). This approach allows us to find a general solution for integrating a rule-based model into the larger CRISP infrastructure at a manageable level of complexity. We can then integrate more complex cubist models using the same process.

Choosing a cubist model does come with several drawbacks. The most prominent of which is the increased complexity, stemming from the rule-based nature of the method. Which strongly reduces the interpretability of the results in comparison single regression equations. This becomes a particularly strong issue with a larger maximum number of rules allowed. At which point the resulting models are theoretically interpretable, but any actual meaningful interpretation becomes practically unfeasible. Additionally, the structure of the cubist model makes it fully nonparametric, effectively trading statistical inference for increased prediction accuracy. In other words, we would be shifting the focus from the right side of our equation to the left. This does not need to be a problem, as long as the decision to do so is made deliberately. Finally, there are some smaller, additional effects that we should make note of; Cubist models tend to be more computationally intensive than standard regression methods; and the linear nature of the sub-models may occasionally cause predicted values that fall outside of the [0,1] interval (which should be corrected to the respective boundary values).

3 Results

To compare the different regression methods and assess determine the relative effectiveness of each method in the context of the CRISP calibration, we use a common set of regressors to for a set of different regression models. Since our intention is to improve on the calibration methods and suitability functions created for the initial version of the CRISP model, we use the same selection of variables used and described by Van der Wielen and Koomen (2024). In addition to estimating a baseline model based on the binary logit method used in the original calibration. To ensure a valid comparison between the original calibration method and the new methods introduced in this report, as well as between the new methods themselves, we split the dataset into a 50/50 training and validation split and use the same training and validation data for each model. To circumvent any issues stemming from the inclusion of (near) zero values in model estimation, we apply a threshold value to the training data. Meaning that any observations above the threshold value $t = 10^{-4}$ are excluded from the training data. We apply several different censored and uncensored sets of out out-of-sample data during our validation step to help mitigate some of the impacts of this thresholding.

Estimating a model on the scale required for the CRISP model is very computationally intensive. For this reason, we have opted to only estimate models for the continent of North America. Although there is no reason to assume that the potential improvements gained would not be replicable for the calibration of other continents in the CRISP model, such as Europe.

3.1 Performance metrics

We use several performance metrics to assess the relative accuracy, goodness-of-fit, and overall performance of the models. This subsection provides a brief explanation of each of the metrics used. Since some metrics can be calculated using different methods, we also indicate the specific method applied.

Root Mean Square Error (RMSE) measures the standard deviation of the differences between predicted values $\hat{y}$ and observed values y, penalising large errors more strongly. RMSE is expressed in the same unit as y. Its value is always non-negative, with lower values indicating a more accurate model. A value of 0 would signify a perfect fit.

$$RMSE = \sqrt{\frac{1}{n}\sum (y_i - \hat{y}_i)^2}$$

R-squared (R^2) represents the proportion of the variance in y explained by a model. It is not a measure of accuracy, but rather a measure of goodness-of-fit. R^2 usually ranges from 0 to 1, where a value of 1 signifies that the fitted model perfectly explains the total variance. Occasionally R^2 may take on a negative value, which signifies that taking the mean $\bar{y}$ would provide a better fit than the fitted model.

$$R^2 = \rho^2 = \frac{cov(y_i, \hat{y}_i)}{\sigma_{y_i}\sigma_{\hat{y}_i}}$$

Mean Absolute Error (MAE) measures how far, on average, a model's predictions $\hat{y}$ are from the observed y. It is less sensitive to outliers relative to RMSE, since errors contribute linearly instead of quadratically. MAE is non-negative, and a lower MAE indicates a more accurate model. A value of 0 signifies a perfect fit.

$$MAE = \frac{1}{n}\sum|\hat{y}_i - y_i|$$

Weighted Mean Absolute Percentage Error (wMAPE) is an expression of the sum of absolute errors, weighted by the observed values y. This naturally weights the percentage error and helps avoid the problem where (near) zero values result in extremely large and misleading percentage errors. wMAPE is non-negative, with lower values indicating a more accurate model. A value of 0 signifies a perfect fit.

$$wMAPE = \frac{\sum|y_i - \hat{y}_i|}{\sum|y_i|}$$

3.2 Regression results

The task of estimating a model on the scale of the CRISP model requires significant computational resources. For this reason, we have elected to estimate models exclusively for the North American continent. Although there is no reason to assume the potential improvements gained would not be replicable for the calibration of other continents within the CRISP model, such as Europe. The various performance metrics for the models we estimated are table 1. Starting with the binary logit model that was used in the initial calibration, it is immediately obvious that it underperforms relative to any of our proposed alternative methodologies. The model most similar to the original suitability model is the fractional logit. From the metrics it is, again, clear that even this most naive alternative method yields significantly more accurate predictions. The reported wMAPE is relatively high, which is consistent with high prevalence of (very) low values observed for BUA. This is highlighted even stronger in the metrics for the censored validation sample containing only observations below the threshold value. This sub-selection of the data consists of almost exclusively (near) zero observations, leading to the significant drop in performance seen in the performance metrics.

Table 1 – Performance metrics for different model types

Sample	Metric	Binary logit*	Fractional logit	Beta (logit)	Cubist (10)	Cubist (100)
Training	RMSE	0.313	0.021	0.023	0.010	0.007
Training	R^2	0.267	0.791	0.752	0.950	0.973
Training	MAE	0.163	0.009	0.010	0.003	0.002
Training	wMAPE	25.63	0.529	0.619	0.134	0.131
Validation	RMSE	0.313	0.019	0.015	0.006	0.005
Validation	R^2	0.266	0.797	0.759	0.958	0.965
Validation	MAE	0.164	0.004	0.006	0.001	0.001
Validation	wMAPE	25.78	0.687	0.687	0.178	0.178
Validation $>$ t	RMSE	-	0.021	0.023	0.009	0.009
Validation $>$ t	R^2	-	0.789	0.750	0.955	0.963
Validation $>$ t	MAE	-	0.009	0.010	0.003	0.003
Validation $>$ t	wMAPE	-	0.529	0.620	0.156	0.153
Validation $\leq$ t	RMSE	-	0.004	0.005	0.005	0.001
Validation $\leq$ t	R^2	-	0.013	0.020	0.020	0.001
Validation $\leq$ t	MAE	-	0.002	0.004	0.004	0.0003
Validation $\leq$ t	wMAPE	-	531.67	1175.84	74.65	84.30

** Binary logit model validation split is different*

The relevant estimated coefficients reported in columns (1 & 2) of table 2 give very few surprises. The addition of the variables omitted in the initial suitability model does not seem to lead to affect the other coefficients in any unexpected manner. The estimated coefficients for the beta regression models, shown in columns (3 & 4), are stronger for the time-lagged BUA, flood prone area, and landslide prone. The other estimated effects are similar or slightly weaker. Based on the performance metrics the fractional and beta logit models seem to be relatively comparable in accuracy.

Table 2 – Coefficients for the fractional and beta logit models explaining Built-Up Area Fraction in 2020.

		(1)	(2)	(3)	(4)
Explanatory variable	Unit	Fract. Logit	Fract. Logit	Beta(logit)	Beta(logit)
(Intercept)		-5.098***	-5.228***	-4.928***	-4.994***
		(0.012)	(0.013)	(0.005)	(0.005)
Built-Up Area Fraction (BUA) 2000	Fraction [0,1]	7.309	7.290***	10.003***	9.977***
		(0.023)	(0.023)	(0.008)	(0.008)
ln(Pop. Dens. 2000)	Inhabitants/km^2	0.431***	0.430***	0.257***	0.256***
		(0.001)	(0.001)	(0.0004)	(0.0004)
Distance to coast	Kilometres	0.0002***	0.0002***	0.0004***	0.0004***
		(1.26e-05)	(1.25e-05)	(6.52e-06)	(6.52e-06)
Distance to large inland water	Kilometres	-0.001***	-0.0005***	-0.005***	-0.001***
		(6.23e-05)	(6.30e-05)	(2.58e-05)	(2.59e-05)
Distance to major road	Kilometres		0.003***		0.003***
			(0.0001)		(3.75e-05)
Distance to secondary road	Kilometres	-0.003***	-0.006***	-0.001***	-0.004***
		(2.22e-05)	(0.0001)	(8.41e-06)	(3.89e-05)
Grid-cost distance to nearest city	Kilometre Eq.	0.002***	0.002***	0.0001***	0.0002***
		(2.89e-05)	(2.88e-05)	(1.70e-05-05)	(1.70e-05)
Grid-cost dist. to nearest city/town	Kilometre Eq.	-0.004	-0.003***	-0.003	-0.002***
		(6.12e-05)	(6.14e-05)	(2.71e-05)	(2.73e-05)
Grid-cost dist. to nearest	Kilometre Eq.	-0.011***	-0.010***	-0.0002***	0.0001***
city/town/village		(0.0002)	(0.0002)	(4.99e-05)	(5.01e-05)
Elevation	Metres	0.00003***	0.00004***	0.00001***	0.00001***
		(2.75e-06)	(2.85e-06)	(1.07e-06)	(1.10e-06)
Slope	Degrees		-0.026***		-0.016***
			(0.001)		(0.0005)
Terrain roughness index	(1-7)	-0.788***	-0.635***	-0.336***	-0.257***
		(0.010)	(0.011)	(0.004)	(0.005)
Protected area	(dummy)	-0.184***	-0.184***	-0.109***	-0.110***
		(0.008)	(0.008)	(0.003)	(0.003)
Flood prone area	(dummy)	-0.027***	-0.032***	-0.042***	-0.047***
		(0.004)	(0.004)	(0.002)	(0.002)
Earthquake intensity	(0-8)	-0.024***	-0.022***	-0.028***	-0.027***
		(0.001)	(0.001)	(0.0002)	(0.0002)
Landslide prone	(dummy)	-0.032***	0.001	-0.044***	-0.021***
		(0.005)	(0.005)	(0.002)	(0.002)
Precision (ϕ)				76.364	76.689
				(0.083)	(0.084)

*Note: *p<0.1; **p<0.05; ***p<0.01*

The cubist model fitted here is composed of 100 sub-models that apply to different subsets of the data. Those subsets can overlap, and results of sub-models are averaged where this overlap occurs. Unfortunately, the nature of cubist models makes it hard to show results in a report format without compromising legibility. But they do allow for insight into how the various branches and leaves of the estimated model are shaped. Table 3 shows the occurrence of the CRISP variables in the rules and sub-models. Both in the rules and sub-models, prior built-up fraction (BUA 2000) features foremost, with this variable present in 99 out of 100 rulesets, and 98 out of 100 sub-models. Other variables that occur in more than half the rulesets are population density, elevation, and grid-cost distance to the nearest city. Notably the Built-Up Area Fraction in 2000 is the only variable that occurs more in the rules than as an estimator, as all other variables occur more often in the sub-models as an estimator. Many more variables occur as estimators in at least half of the sub-models, including BUA in 2000, population density, elevation, grid-cost distance to cities, towns or cities, secondary roads, coast and large inland water, and slope. The direction of these estimated effects cannot be generalized though without looking at individual sub-models, which is outside of the scope of this report.

Table 3 – Occurrence of variables in cubist (100) rules (i.e. model leaves) and in the fitted sub-models. Occurrence is counted as one when a variable is applied at least one time in a rule or sub-model, else it is counted as zero.

Variable	Occurrence in rules	Occurrence as estimator
Built-Up Area Fraction (BUA) 2000	99	98
ln(Pop. Dens. 2000)	66	81
Elevation	66	80
Grid-cost distance to nearest city	57	65
Grid-cost dist. to nearest city/town	49	54
Distance to secondary road	45	63
Distance to coast	44	58
Terrain roughness index	21	36
Distance to large inland water	18	54
Slope	16	56
Earthquake intensity	13	42
Distance to major road	6	45
Flood prone area	5	43
Grid-cost dist. to nearest city/town/village	2	36
Protected area	0	1

3.3 Validation within CRISP

To assess the accuracy of the estimated suitability values we evaluate how well they predict observed built-up changes according to the GHSL-BU total built-up layers of 2000 and 2020. To this end we have two main approaches:

1. Comparing observed built-up development with the direct built-up development predictions from the models tested here, so that other CRISP characteristics (regional demands and decadal results) do not play a role; and
2. Comparing observed built-up development with built-up development as modelled in CRISP routines in which the simulation uses predictions from the various suitability models tested here, but is constrained with regional demand settings and decadal time steps to allow a more direct comparison with the CRISP setup.

The difference between the two approaches is that the first approach offers a comparison of almost raw prediction values, while the second approach compares observed development with model results when the presented suitability models are integrated in the CRISP model. Thus, in the latter approach, built-up demand (specified as the summed increase in built-up between 2000 and 2010, and 2010 and 2020, in so-called functional areas) is downscaled over the predicted values for 2010 and 2020, respectively, in two decadal modelling steps. Predicted suitability levels are used as a proxy for the downscaling. We execute all analyses for one study area as defined in CRISP, namely the "North America" area. This study area entails the United States including Alaska and Hawaii, Canada, Mexico, other Central American countries, and all Caribbean territories.

Directly comparing suitability models predictions and built-up development

To gain some intuition on how the various suitability approaches differ we start with an analysis of built-up development predictions as they are produced by the various suitability models tested in this report, thus without any effect of other assumptions in the CRISP model. Doing so helps understand the extent to which the different modelling approaches differ, and gives a first insight into their comparative levels of accuracy. CRISP relies on approximations of local built-up development (i.e., the level of increase of built-up fractions), while the fractional logit, beta and cubist models predict absolute future built-up levels. For a fair comparison, we therefore analyse the predicted built-up fractions of these models in 2020 minus the pre-existing built-up fractions in the year 2000. It is also important to note that the cubist model often has sub-models that overlap for a specific location. The average result for all relevant sub-models is taken for such locations.

Figure 2 shows observed development (i.e., BUA in 2020 – BUA in 2000), the built-up development proxy that was used in CRISP for the World Urbanisation Prospects work, and results using the newly introduced approaches described in this report. For this discussion we focus on the East coast of the United States, as it entails both highly metropolitan areas, suburban zones, and remote rural and wilderness areas. Looking at the actual observed built-up development, several aspects stand out. As noted in Jacobs-Crisioni et al. (2025), some built-up development occurs almost everywhere in the built environment. Relatively limited development occurs in the centres of large cities such as New York. However, we can spot clear hotspots of development, relatively small areas in city outskirts where built-up has increased considerably over the simulated 20 years.

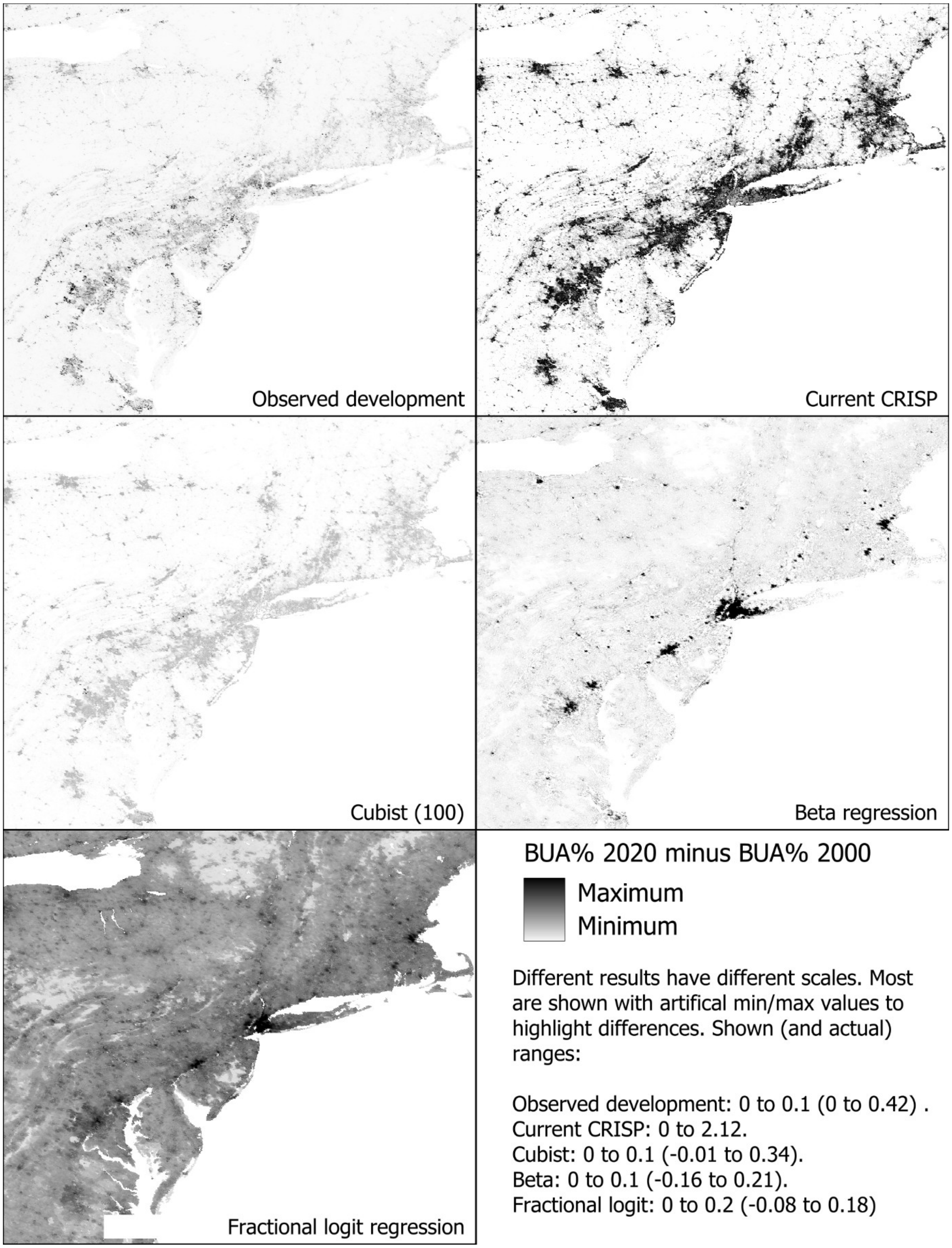

Figure 2 – Observed and predicted built-up development between 2000 and 2020 in the triangle between Washington D.C., Buffalo, and Boston in the United States. Visualization rules have been adapted to highlight spatial variation in built-up development except for the current CRISP results. Full range of the layers is given between parentheses.

The various tested models differ substantially in outcomes even if the set of included explanatory variables is the same.

- The current CRISP proxy values are much higher than the predictions of the other models. This is a consequence of the rescaling of suitability values and their following multiplication with ET values (which can exceed values of 2.5). While these much higher values are not problematic when used as a proxy, they do make direct numerical comparisons with the other results a bit difficult. The results of this proxy highlight large areas within and around existing settlements, with the centres of major metropolitan areas having somewhat subdued values (no doubt because of the limiting effect of ETs in already very developed locations). The hotspots of urban development that are observable in city outskirts, are however not apparent in these predictions.
- By comparison, the cubist model results seem less pronounced but also appear to highlight existing settlements and larger areas around them with subdued values in metropolitan centres similar to the current CRISP values.
- The beta regression results strongly emphasize the centres of metropolitan areas, while smaller settlements appear quite subdued.
- The fractional logit results are possibly spatially the least distinctive, with smooth gradients that reach well beyond settlement limits.

To better quantify the relationship between the various predicted levels of built-up or built-up change and the observed changes in built-up fraction, we have performed Pearson correlation analyses on the predicted and observed levels of built-up change. The correlations are shown in table 4 together with correlations with previous (2000) built-up levels. The latter serve as a crude benchmark here. From these results follows that the approach taken by CRISP for the WUP and the cubist model both perform better than previous built-up levels. The cubist model performs best by far.

Table 4 – Correlations between predicted levels of built-up fraction change and observed changes in BUA for five alternative specifications. Values are reported for subsets based on initial built-up levels. The last row summarises the total score over all observations of the observations in the North America study area.

BUA Subset	CRISP	Cubist (100)	Beta(logit)	Fractional logit	Previous built-up
$l \geq 0.1$	-0.145	0.322	0.200	0.208	-0.157
$0.01 \leq l < 0.1$	0.164	0.297	0.055	0.177	0.111
$l < 0.01$	0.552	0.616	0.167	0.435	0.402
All cells	0.538	0.667	0.168	0.330	0.416

Simulations with constrained built-up fraction changes

To better compare the accuracy of the alternative suitability settings with the current implementation of CRISP, we also perform a validation in which the suitability outcomes are used as a proxy for downscaling observed built-up development between 2000 and 2020. This entails that suitability maps were created using the results of the various presented models, and their values (expressed as a logit in the case of the CRISP, beta and fractional logit models) were used to downscale a fixed budget of urban development. Similar to the WUP application, residential built-up development is summed between 2000 and 2020 for every accounted

Functional Area; and subsequently downscaled to grids in two steps. Non-residential built-up land in grids is added to the residential built-up land after the modelling.

To assess the accuracy of our approaches, we compare the observed built-up area fraction in 2020 (BUA_{2020}) with the estimated fraction ($\widehat{BUA}_{2020}$) in 2020. We added a benchmark model that is driven by previous built-up land. In that model, the levels of BUA on the map are acting as suitability for land-use development, so that built-up development is allocated proportionally to prior built-up levels. We consider this a crude predictor of development locations that likely exaggerates development in highly developed locations but believe it offers a better feasible benchmark than e.g. random development.

To improve intuition on how much the various suitability models are wrong where, figure 3 shows the distribution of model errors for the alternative specifications. All results are shown with the same colour scale. By and large, from this figure follows that all our models make the same mistakes – but the intensity of the error differs drastically between models. None of the tested models predicted the hotspots of built-up development in the urban fringes of main metropolises, visible here as red patches of underestimation in the fringes of cities such as Washington DC (left centre) and New York (middle). All tested models overestimated the built-up development in the centre of the three large cities in the map. However, the CRISP and cubist models get much closer numerically than the beta and fractional logit models.

Table 5 shows weighted Mean Absolute Percentage Errors for the various tested approaches. These are calculated as $\sum abs\,(\widehat{BUA}_{2020} - BUA_{2020})/\sum BUA_{2020}$. Next to the introduced suitability models and benchmark we also estimate the comparative accuracy of omitting built-up development between 2000 and 2020 entirely, so that we assess how accurate built-up levels in 2000 are as a descriptor of built-up levels in 2020. From these follows that only two of the estimated approaches (the CRISP approach used for the WUP, and the tested cubist (100) model) perform better than our crude benchmarks. The beta and fractional logit models perform worse than the theoretical omission variant, so that misallocated built-up development in the beta and fractional logit models exacerbate the fundamental error brought forth when assuming no urban development occurred.

Table 5 – Mean absolute percentage errors (MAPE) from the tested models and benchmark approaches.

Model	CRISP	Cubist	Beta(logit)	Fractional logit	Previous built-up	Omission
MAPE	0.174	0.150	0.317	0.292	0.189	0.227

Note: MAPE from omission is computed as the absolute value of summed BU in 2000 – summed BU in 2020, divided by BU in 2020.

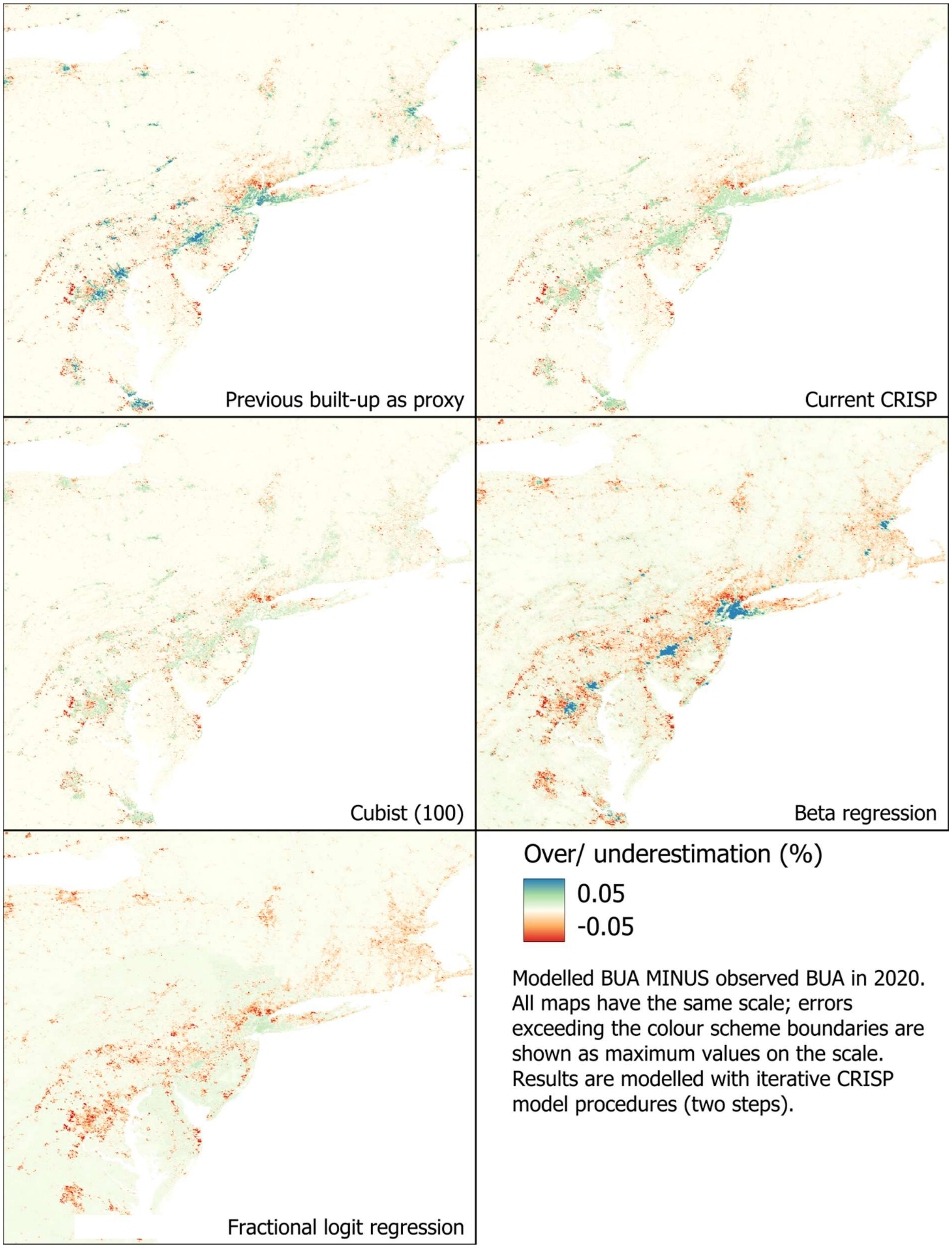

Figure 3 – Errors in predicted built-up levels in 2020 in the triangle between Washington D.C., Buffalo, and Boston in the United States, highlighting locations where models overestimate built-up development (in green and blue) and where models underestimate built-up development (in orange and red).

4 Discussion and conclusion

There are several extensions possible that we have not implemented in this report. To begin, the logit link function we apply in both our fractional and beta regressions could be exchanged for an alternative link function (such as Probit, Cauchit, etc.). This is a relatively minor alteration as far as the estimation of suitability function is concerned but requires major adjustments in the implementation phase of the model, before we can use them to predict future built-up fractions. A further, natural extension to the basic beta regression model would be the *variable dispersion beta regression model,* which allows the precision parameter ϕ to vary between covariates, to better control for heteroskedasticity (Cribari-Neto and Zeileis, 2010). This is likely to further improve performance. Additionally, the beta regression model can be extended to a so-called *zero-and-one-inflated beta regression model* (Ospina and Ferrari, 2012), a mixture modelling approach that allows for response variables on a closed interval [0,1] by separately handling the probability of boundary values (0 and 1) and the continuous values in between.

Regarding the cubist models, it would be possible to employ a larger-scale model to better take advantage of regional sub-modelling. Additionally, the 'cubist' R-package currently uses all variables inputted into a model for both the partitioning and regression components. The ability to provide specific instructions on whether to use certain variates for partitioning and others for the explanatory functioning of the regression components would allow a more informed application of these elements. This could, for example, help distinguish specific spatial regions where drivers are likely to have a different effect, or allow a different behaviour based on the initial built-up fraction and thus replicate the functioning of the Expected Top-ups factors implemented in CRISP.

A common result for the models tested here is that the move from explaining dichotomous built-up in 2UP to fractional built-up in CRISP enables much richer suitability models. Even when accounting for previous built-up development in the explanatory models there is enough variation in built-up developments to also pick up statistically significant effects of the other factors that are known to drive the presence of built-up land. Built-up development appears as a combination of gradual development in existing built-up environments, and the shock appearance of hotspots of urban development in what, from a map, appear to be greenfield urban expansion projects. It may be particularly difficult to capture development hotspots with models that seek averages, as the occurrence of such hotspots is likely explained by strongly local logic. Models such as cubist may be best suited to the prediction of such hotspots as they can likely be captured better with a combination of exceptional rules.

References

Andrée BPJ and Koomen E (2017) *Calibration of the 2UP model*. Spinlab Research Memorandum SL-13, 22 December. Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam.

Black B, Van Strien MJ, Adde A, et al. (2023) Re-considering the status quo: Improving calibration of land use change models through validation of transition potential predictions. *Environmental Modelling & Software* 159: 105574.

Cribari-Neto F and Zeileis A (2010) Beta Regression in R. *Journal of Statistical Software* 34(2).

Ferdinand P (2020) *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, Amsterdam.

Ferdinand P, Andrée BPJ and Koomen E (2021) *Revised calibration of the 2UP model; analysing change and regional variation*. Spinlab Research Memorandum SL-19, 20 December. Spatial Information Laboratory (SPINlab), Vrije Universiteit Amsterdam.

Ferrari SLP and Cribari-Neto F (2004) Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics* 31(7): 799–815.

Jacobs-Crisioni C, Schiavina M, Krasnodebska K, et al. (2025) *Introducing the CRISP Model to Downscale Future Population Projections*. Publications Office of the European Union. Available at: https://data.europa.eu/doi/10.2760/7163875.

Johnson NL, Kotz S and Balakrishnan N (1995) Continuous univariate distributions. 2. 2. ed. New York: Wiley.

Kieschnick R and McCullough BD (2003) Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling* 3(3): 193–213.

Koomen E, Diogo V, Dekkers J, et al. (2015) A utility-based suitability framework for integrated local-scale land-use modelling. *Computers, Environment and Urban Systems* 50: 1–14.

Kuhn M and Quinlan R (2025) Cubist: Rule- And Instance-Based Regression Modeling. Available at: https://topepo.github.io/Cubist/.

Loonen W and Koomen E (2009) *Calibration and Validation of the Land Use Scanner allocation algorithms*. 550026002, 20 August. Netherlands Environmental Assessment Agency (PBL).

Ospina R and Ferrari SLP (2012) A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* 56(6): 1609–1623.

Papke LE and Wooldridge JM (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11(6): 619–632.

Pesaresi M and Politis P (2023) GHS-BUILT-S R2023A - GHS built-up surface grid, derived from Sentinel2 composite and Landsat, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC). Available at: http://data.europa.eu/89h/9f06f36f-4b11-47ec-abb0-4f8b7b1d72ea (accessed 19 July 2024).

Quinlan JR (1992) learning with continuous classes. In: *Proceedings of Australian Joint Conference on Artificial Intelligence*, Hobart, 16 November 1992, pp. 343–348.

Tu W (2012) Zero-Inflated Data. In: El-Shaarawi AH and Piegorsch WW (eds) *Encyclopedia of Environmetrics*. 1st edn. Wiley. Available at: https://onlinelibrary.wiley.com/doi/10.1002/9780470057339.vaz000g.

United Nations Department of Economic and Social Affairs (2025) *World Urbanization Prospects 2025*. United Nations.

Van der Wielen T and Koomen E (2024) *Calibration of the CRISP model; A global assessment of local built-up area presence*. SPINlab research memorandum SL-25. Amsterdam: Vrije Universiteit Amsterdam.